UNCERTAINTY QUANTIFICATION FOR INTERPRETABLE AND RELIABLE MACHINE LEARNING

CARINA NEWEN

Dissertation
zur
Erlangung des Doktorgrades (Dr. Ing.)
der
Fakultät für Informatik
der
Technischen Universität Dortmund

JULI 2025, DORTMUND

Angefertigt mit Genehmigung der Fakultät für Informatik
der Technischen Universität Dortmund

Carina Newen : *Uncertainty Quantification for Interpretable and Reliable Machine Learning*

GUTACHTER:
Prof. Dr. Emmanuel Müller
Prof. Dr. Barbara Hammer

DORTMUND, JULI 2025

ABSTRACT

In the booming research field of machine learning and artificial intelligence, uncertainty quantification is often overlooked as an essential quality guarantee. When interacting with and applying artificial intelligence, it is common to evaluate the performance of such learners by metrics such as efficiency, accuracy and other task-specific performance-based metrics without emphasizing the importance of quantifying potential hazards.

Uncertainty characterizes the proximity between observations and predictions, providing a measure of how well a model reflects the true underlying data distribution. This thesis places uncertainty quantification at the center of its investigation, and we investigate three interconnecting subareas:. Explanations and visualizations of uncertainties, robustness with regard to uncertainties, and trustworthiness and human interpretability of uncertainties.

This interdisciplinary setup is crucial to establishing new connections between the domains. While each of the subjects has been extensively studied in isolation, it is often at the interconnections that new research paradigms emerge: Neural networks as we know them today are a result of a fusion of neuroscience-inspired models of cognition, mathematical formalization, and algorithmic innovation. Without the mathematical groundwork, the advances in computing power and the biological neuron inspiration, this new field of research would not exist today.

By interweaving explanation, robustness and trust in the context of uncertainty, in this thesis, we aim to pave the way for engineering practical systems that are both reliable, interpretable, and ultimately trustworthy. In these three areas, we focus on empirical approaches and solutions for important research challenges. The first part of this thesis focuses on visualizing uncertainties of high-dimensional data in an unsupervised setting using the novel proxy for local intrinsic dimensionalities.

Furthermore, we show limitations of popular explainable AI methods using a newly constructed open-source dataset that focuses on an ambiguous classification task. We use the proxy of local intrinsic dimensionality as a proxy for the likelihood of adversarial attacks, connecting uncertainties with robustness metrics.

In the second part of the thesis, we delve more deeply into the robustness domain by proposing certainty attacks and discussing the independence of adversarial transferability to topological changes in the datasets. We discuss the origin of transferability and possible

research directions for future work. The main motivation between the importance of uncertainties stems from the need to calibrate human trust—for the successful application of machine learners, we have to align the trust levels of humans according to their actual performance. This motivates the third part of the thesis, where we discuss trust in AI systems with a special emphasis on uncertainty quantification. Finally, we discuss open challenges regarding uncertainty quantification and outline future work in this particular domain, with special emphasis on explainable AI and robustness.

PUBLICATIONS

This dissertation is based on the following publications. The main supervisor of all my work is Emmanuel Müller; this includes the peer-reviewed publications [NM22a], [NM23], and [NPM25]. One publication was led by main author Magdalena Wischnewski [Wis+24b]. In this work, I was responsible for providing guidance on what AI seals could realistically look like and other conceptual input in the study design, but the study was conducted and evaluated by her.

The thesis includes several papers, including collaborations that are still under review [New+25a; NV17; New+25b], on which I am the main author.

[New+25a] Carina Newen, Daniel Bodemer, Sonja Glantz, Emmanuel Müller, Magdalena Wischnewski, and Lenka Schnaubert. "Uncertainty Awareness and Trust in Explainable AI – On Trust Calibration using Local and Global Explanations." In: *2025 IEEE International Conference on Data Mining (ICDM)*. 2025.

[New+25b] Carina Newen, Luca Hinkamp, Maria Ntonti, and Emmanuel Müller. *Do you see what I see? An Ambiguous Optical Illusion Dataset exposing limitations of Explainable AI*. 2025. arXiv: 2505.21589 [cs.CV]. URL: https://arxiv.org/abs/2505.21589.

[NM22] Carina Newen and Emmanuel Müller. "Unsupervised DeepView: Global Explainability of Uncertainties for High Dimensional Data." In: *2022 IEEE International Conference on Knowledge Graph (ICKG)*. IEEE. 2022, pp. 196–202.

[NM23] Carina Newen and Emmanuel Müller. "On the Independence of Adversarial Transferability to Topological Changes in the Dataset." In: *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. 2023, pp. 1–8.

[NMN25] Carina Newen, Emmanuel Müller, and Albert Newen. "Trust and Uncertainties: Characterizing Trustworthy AI Systems Within a Multidimensional Theory of Trust." In: *Topoi* (2025), pp. 1–22.

[NPM25] Carina Newen, Sofia Vergara Puccini, and Emmanuel Müller. “Certainty Attacks using Explainability Preprocessing.” In: *DAWAK 2025*. Springer, 2025.

[Wis+24] Magdalena Wischnewski, Nicole Krämer, Christian Janiesch, Emmanuel Müller, Theodor Schnitzler, and Carina Newen. “In Seal We Trust? Investigating the Effect of Certifications on Perceived Trustworthiness of AI Systems.” In: *Human-Machine Communication* 8.1 (2024), p. 7.

ACKNOWLEDGMENTS

I would like to thank everybody who advised me about research during my Ph.D. and helped me get to the final version of my thesis, particularly Prof. Dr. Emmanuel Müller, Prof. Dr. James Bailey for his guidance in Australia, Prof. Dr. Xingjun Ma for his kind words about my XAI work. Additionally, I want to thank my family for their support. Furthermore, I would like to thank my interdisciplinary working group for keeping morale high, especially Simon Lutz for unwavering support during deadline stress.

To my colleagues in the RC-Trust building who shared the working space, you know who you are, thank you for keeping it interesting, especially Anna Neumann for being an office mate for a long time. Furthermore, I thank the members of the chair for their collaboration and encouragement, for shared lunch breaks and good discussions.

CONTENTS

1

INTRODUCTION AND THESIS OVERVIEW

Uncertainty Quantification describes the scientific process of mathematically characterizing biases and risks and provides a quantitative assessment of reliability [Hig+10] and other sources of error [Cou+12]. This consideration is especially important in high-risk domains [Sul15]: Beyond assessing the accuracy and alignment of a model with ground truth, it focuses on quantifying the relationship between pieces of information in machine learning models [Sul15].

While delving into the vast amount of literature on distinctions of epistemic and aleatoric or aleatory uncertainty [Sul15; Soi17] that explore biases, estimations, and other scientific inferences [HW21; KO01; Abd+21a; Smi24], we became aware of a noteworthy observation: Uncertainty quantification is closely aligned to robustness research in literature: For example, Schweighofer et al. [Sch+23] show that training adversarial models leads to more realistic estimates of epistemic uncertainty. Birrell et al. [Bir+19] established a theoretical and computational connection between distributional robustness and uncertainty quantification (UQ), particularly in rare event scenarios, by quantifying the maximum risk of the rare event in all plausible models.

Further publications highlight an implicit connection between robustness and the uncertainty quantification domain [Kab+19; Par13], which leads to intriguing research directions. Beyond this connection, we observe a similar pattern between the field of explainable AI (XAI) and robustness: Empirical results indicate that the interpretability of model gradients is crucial for adversarial robustness [Noa+21].

Other works describe that the representations learned by robust models align better with salient data characteristics and human interpretability [Tsi+18a]. This relationship between robustness and interpretable AI has been explored from several different angles [ALG19; RD18; Cha+20a; Etm+19] and offers a fresh perspective along with novel problem statements to explore.

In this thesis, we bridge the research areas by proceeding as follows: In the explainability domain, we tackle the visualization of uncertainties in combination with robustness estimations for unsupervised learners. We further take a look at the limitations

of explanations in ambiguous settings: We specifically construct a new optical illusion dataset to highlight constraints and illustrate boundary cases. We use this to propose future directions for explainable AI methodologies.

In the robustness area, we target the question of the transferability of adversarial examples. We mitigate common adversarial detection metrics by attacking the certainty of a model using explainable AI. This naturally leads to our last challenge tackled: Considering that we have explanations, appropriate uncertainty quantification and robustness of the model established, how are we then able to appropriately calibrate trust [Mir+16]? What is trust in AI, especially with regard to uncertainty quantification?

At its core, uncertainty quantification is the end-to-end study of the reliability of scientific inferences [Ene09]. Typical uncertainty quantification tasks involve carefully evaluating the assumptions made during the formulation of the problem, confidence estimation, data assimilation, and many other tasks. Intuitively, it is a common misconception that quantifying uncertainties means highlighting only the weaknesses of a system or that the presence of uncertainties is always necessarily an indicator of bad model quality.

However, uncertainty is not inherently negative—many systems are naturally stochastic. For example, when rolling a die, randomness is expected and necessary to accurately describe the process. Similarly, introducing uncertainty in models helps us better represent and understand such variability. In essence, machine learning algorithms prove to be an enormously powerful tool, but their potential and capabilities are currently overshadowed by several challenges: their susceptibility to adversarial attacks, [Noa+21], the black-box nature of prevalent models leading to interpretability difficulties that therefore hide potential uncertainties [Gil+18], and the challenge of calibrating user trust appropriately [Mir+16].

When we describe uncertainty quantification, we commonly divide it into epistemic and aleatoric or aleatory uncertainty [Sul15; Soi17]. Epistemic uncertainty is an uncertainty that arises from a lack of knowledge. Aleatoric uncertainty refers to inherent randomness, [Sul15]. Furthermore, highlighting the weaknesses of machine learners can be essential in high-risk applications. After all, we wish to understand if a learner is not sure about whether the classified skin patch is cancerous [Abd+21b]. The same applies to an autonomous car, which is uncertain about its decision to stop [Mic+20].

1.1 OVERVIEW AND CONTRIBUTIONS

In this work, we unravel the misconception of uncertainty being purely detrimental and focus on three main areas from a fresh, cross-domain perspective, namely: visualizations of uncertainties (Part I), robustness and uncertainties (Part II), and an interdisciplinary perspective regarding uncertainties as a trust calibration tool (Part III). In Part I, Chapter 3 concentrates on unsupervised global explainable AI of uncertainties. We then extend this to a discussion of the limitations of explainable AI in Chapter 4, particularly in contexts where visualizations alone may fall short of providing true interpretability.

This is an important research area due to potential hidden biases, and for ensuring fairness and accountability of models [Min+22]. Rather than narrowly focusing on the conventional division of aleatoric and epistemic uncertainty, in Part II, we focus on the question: What makes a model uncertain in real-life high-stake applications? One of the most persistent challenges and weaknesses of even the most sophisticated machine learning models is adversarial vulnerability [GSS14; Sze+13; Pap+16a; Don+18], a widely discussed and interesting phenomenon.

We suggest that ideally, uncertainty should be able to reflect the likelihood that a given input is adversarial in nature, especially since current methods are unable to fully prevent such attacks [BR18]. The most we are currently able to do is verification techniques of networks, which are costly and often limited in application [Kat+17; Hua+17]. We argue that there is an intuitive link between uncertainty and robustness, in the sense that both can establish trust calibration [NMN25], which we discuss in Chapter 9 in Part III of this thesis, linking our interdisciplinary work with our technical contributions. In Chapter 10 we ask ourselves- what does the perfect explainability algorithm help us if it remains human uninterpretable?

This interpretability goal was already identified in the very first work on explainable AI [RSG16], and remains relevant today. However, in many works, this is validated with very small sample sizes of human subjects [RSG16], and simply claims to be human interpretable by definition. In this thesis, we validate our proposed explainability algorithm with a study of the trust calibration abilities of the algorithm in Chapter 10. We maintain that interdisciplinary approaches are not only beneficial but essential: technical developments in explainability can be enriched by per-

spectives from philosophy, cognitive science, and human-computer interaction and vice versa.

Table 1.1: Overview of the Thesis Contributions

	Technical Aspects		Human-Centered Aspects	
	Visualizations of Uncertainty	**Robustness and Uncertainty Quantification**		**Trustworthiness and Interpretability**
Method	Chapters 3	Chapter 7	**Theory**	Chapters 9
Application	Chapter 3	Chapter 7	**Empirical Study**	Chapter 10, 11
Analysis / Limitations	Chapters 3, 4, 10,	Chapters 6	**Analysis/ Limitations**	Chapter 10, 11

Chapters 2 and 5 and 8 provide, respectively, the introduction and the related work necessary to Part I, II and III. A general overview of our major contributions and the general structure of the thesis can be found in Table 1.1, summarizing our contributions. This thesis can be divided into technical contributions and human-centered aspects, visualized in Table 1.1.

1.1.1 *Part I: Visualizations of Uncertainties - Global Explanations and Limitations of Explainable AI*

In the first part of the thesis in Chapter 3, we propose a novel global uncertainty visualization method for high-dimensional data. We focus on a classification setting. Our starting point is the definition of unsupervised quantifiable uncertainties in the unsupervised domain by the unique proxy of local intrinsic dimensionality [NM22a]. This novel function bridges confidence scores as an uncertainty metric with a score that reflects, but is not strictly mathematically, a likelihood of adversarial attack. [Ma+18] argue that there is a correlation between local intrinsic dimensionality and identifying adversarial examples.

We visualize the unsupervised quantifiable uncertainties by a combination of prior uncertainties in the form of confidence scores and using the knowledge that adversarial subspaces can be identified by this proxy [Ma+18; Ams+15]. This allows the determination of whether the model will be uncertain about its decision for a specific dataset.

Additionally, it visualizes if the data contains malicious instances that lead to the model potentially misclassifying instances. In our empirical evaluation, we demonstrate the usefulness of this method on multiple datasets, detecting misclassification despite not using

the labels as an identification metric. We also discuss potential limitations [LCY18] of this methodology, as well as possible future work.

In Chapter 4, we take a more general look at current explainable AI methodologies [Sel+17; Nau+23; Gho+19] and critically examine their limitations. To do so, we start with a prominent optical illusion, the rabbit and duck image, which challenges human visual perception. With the help of this illusion, we identify a key weakness of current methods: The underlying focus of the important features or areas in the image is always pixel-based.

However, it is known in the psychological domain that human understanding is not; we construct what we see from context [Gre15; Gre70]. Our minds actively adjust what we see, often relying on high-level concepts rather than raw sensory input [Gre97; NV17].

These perceptual biases and context-dependent phenomena are not yet encoded in current AI systems, which typically learn and explain at the level of pixel arrangements or activation gradients. This motivates our argument for the development of truly concept-based, rather than purely pixel-based, explanations. Guided by this perspective, we construct a novel dataset that aims to capture visual ambiguity in an animal classification context.

We did this using the image construction model of ChatGPT [Ope24], DALL·E. The goal was to generate additional animal optical illusions which incorporate shared features of two animals such that both are perceptually visible, often with one hidden within the other. This concealment of one animal within another allows us to probe how both humans and machine learning models handle conceptual ambiguity.

We discuss how this dataset helps us identify key concepts for animals, such as gaze direction or eye coordinates, that help learning accuracies when included in the training procedure and provide more meaningful explanations. The dataset thus provides a foundation for further exploration, opening up numerous directions for future research in concept-based learning, human-AI alignment, and interpretable AI under perceptual uncertainty.

1.1.2 *Part II: Robustness and Uncertainties*

Part II (Chapters 7 and 6) extends our bridge between uncertainties and robustness. Chapter 7 constructs an attack specifically on

the certainty (in this case, the confidence score) of the network by using insights from explainable AI. The idea was to introduce more than just accuracy as a key metric: We take a look at prominent attack mitigation techniques, which rely on factors such as certainty [SG18].

The key metrics we focus on are success rate, confidence in the misclassification, transferability of attacks to other models, and the image quality of the generated adversarial. We decide that, on top of success rate and certainty, taking a look at transferability increases the potential of an attack in real-world settings due to its extended applicability to other models. Meanwhile, image quality is vital for human imperceptibility: adversarial examples are more dangerous when they retain high visual fidelity. We want to test whether using explainable AI can improve those key metrics and, therefore, lead to stronger, more generalizable attacks.

Building on this, we then extend the robustness discussion by discussing adversarial transferability in Chapter 6. Inspired by numerous papers attempting to understand why and how adversarial examples transfer [Ily+19; Pap+16a], we endeavor to find a correlation using topological data analysis in Chapter 6 between the topological distance of two datasets and the transferability of attacks for models trained on the altered and original dataset.

If such global hidden features are indeed shared, topological data analysis (TDA) is one of many tools that offers a way to compare them, as it promises to preserve structural invariants of a dataset and reflect the geometric properties that remain stable under deformation or noise [SMC+07].

In order to compare the graph structures output by the mapper algorithm, we use the proxy NetLSD [Tsi+18b]. NetLSD provides permutation-invariant, scale-adaptive, and size-invariant comparisons between graph structures, making it well-suited for evaluating similarity between graph structures. However, our empirical analysis reveals no observable correlation between transferability and topological changes of the dataset, within our discussed limitations due to empirical evaluation.

Nonetheless, this opens up a new aspect: Why does the expected correlation not manifest in our empirical evaluation? Are our current assumptions about adversarial transferability incomplete or misguided? Overall, Part II of this thesis provides an interesting angle on the construction of future attack methods. It invites a reevaluation of the metrics used to measure attack strength. We ask ourselves: When do attacks transfer?

1.1.3 *Part III: Trustworthiness and Human Interpretability of Uncertainties*

In Chapter 9, we characterize warranted trust in AI systems by defining a general theoretical framework of trust in AI systems [NMN25]. We note a conceptual gap in literature in key dimensions of trust, allowing for an extension of existing trust definitions.

We validate this claim by conducting a focused literature review. Additionally, we present our summarized key objective dimensions for trust and elaborate on these dimensions using case studies, including widely used systems such as ChatGPT [Ope24].

The proposed framework allows the evaluation of trust on a numerical scale. Each dimension is further decomposed into observable subdimensions; for example, under transparency, one such subdimension is explainability. The fewer subdimensions, and therefore dimensions, are fulfilled, the less a system should be trusted.

With this work, we aim to bridge the gap between the conceptual literature on trust and the technical components of AI systems that can serve as levers for trust calibration.

Our framework thus provides a structured basis for both analyzing trustworthiness and guiding design choices in AI systems intended for real-world deployment. We continue by evaluating the trust calibration abilities of our proposed method, Unsupervised Deepview [NM22a], by means of a psychological study in Chapter 10. We argue that this is often neglected, even though it gives valuable insights into potential mental overload for the user: For example, in our case, we created a two-dimensional overview of how a model performs on a dataset, marking each datapoint as either certain or uncertain.

The method allows for the examination of the original datapoints, enabling inspection of whether a point is justly marked as uncertain (for example, if the image instance is unclear for human classification as well). We found that while the background colouring of the method is an additional, perhaps even useful, estimate of whether a model performs reliably on unseen data, in practice, this additional feature just leads to confusion.

While we did not evaluate the exact cause—whether it is cognitive overload or not- users stated that the extra information overall did not always help them in their decision towards trustworthiness.

The lesson we learn from this in general is that sometimes less is more: Even if every additional addition of information might be useful to someone, there is a certain trade-off between information

gain and overcomplexity [New+25a]. This is often neglected as a consideration when explainable AI algorithms are designed from a computer scientist's perspective. And finally, in Chapter 11, we present a collaborative study led by Magdalena Wischnewski.

This study investigates the effect of AI seals on the perceived trustworthiness of AI systems. However, trust in the institution that issued the seal was positively correlated with system trust, suggesting the presence of epistemic trust. Even more notably, among the different seal types presented, the least understood—verification-based seals—was rated as the most desirable.

To summarize, in this thesis, we tackle the following research challenges: In Part I, we focus on explaining uncertainties. First, we discuss and design visualization methods that enable an understanding of both empirical estimates of uncertainties and robustness. This is an important contribution because it allows a comprehensive interpretable breakdown of uncertainty quantification for a classification task without the need for labeling [NM22a].

This is especially useful in scenarios where the labeled data is expensive or has limited availability.

An additional contribution is the discussion of limitations of existing explainability methodologies. This reflection is critical for framing future research directions and open problem challenges in the explainable AI domain. In Part II, we examine the transferability of adversarial attacks using empirical estimates. This is in itself a research area widely discussed, and if solved, could provide a new milestone for machine learning research [Liu+16; Ma+18; PMG16; Xie+19].

Additionally, we propose an adversarial attack strategy specifically designed to target model confidence scores, thereby circumventing common detection mechanisms. Given that the adversarial domain is characterized by a continual arms race between attack and defense strategies, this approach presents a compelling direction. The novelty of our method lies in its integration of concepts from explainable AI, robustness, and uncertainty quantification to effectively achieve this objective. Moreover, in Part III, we present a framework for characterizing and quantifying trust.

Besides that, we investigate the trust calibration ability of our proposed visualization method and discuss the role of AI seals for trust calibration. Trust calibration is an important research problem because trust is affected by many different aspects, for example the habituation and adoption [LR23] and the end-goal of reliability and interpretability of this thesis will only be achieved by the correct adoption of models and systems due to warranted trust,

and not over- or undertrust in systems [SL21]. If the capabilities of systems are not correctly estimated by users, this brings new dangers, however well-designed they are.
The goal of this thesis was to provide a more comprehensive connection between the research areas of robustness, uncertainty quantification and trust in AI, and therefore open future research directions by combining the domains. We contributed new insights across all three domains, and we argue that this integrative viewpoint enables us to identify and extend existing limitations, thereby uncovering new potential directions for future research within each field which we will discuss further in our future work.

Part I

VISUALIZATIONS OF UNCERTAINTY AND CURRENT LIMITATIONS OF EXPLAINABLE AI

2

INTRODUCTION AND RELATED WORK

Explainable AI is the field that is concerned with making the behavior and decision-making processes of complex machine learning models transparent, interpretable, and understandable to humans [Arr+20]. This has become more and more relevant with the increasing adoption and complexity of machine learning models [Dwi+23]. Initial approaches of distinguished visualization techniques that explain the quality of a model include LIME [RSG16], SHAPex [LL17b], Anchors [RSG18], Grad-CAM [Sel+17] and LIME-based extensions such as EXPLAIN-IT [MCM19] for unsupervised learning. In addition, other methods such as LEMNA [Guo+18] extend the explainability to recurrent neural networks or networks that are specifically needed for security-critical contexts. Furthermore, there are other approaches, such as causal explanations for supervised learning algorithms [SK19]. The field has expanded by introducing the more detailed terms comprehensibility, succinctness, understandability, interpretability and explainability [Min+22].

In many applications, especially when the applied model is black-box, we wish to understand and interpret the underlying decisions with regard to the trustworthiness of its decisions: when machine learners become involved, liability and responsibility become unclear, but the system has to remain justifiable [Hof+18].

This motivates the bridge to our interdisciplinary contributions of the thesis in Part III. In real-world scenarios, labeled data is often scarce or unreliable [Mas+12], but there remains a strong need to understand the decision boundaries of classification models.

Understanding where the decision boundaries lie is particularly important for identifying data points that lie near these boundaries, points that are especially susceptible to small perturbations, potentially resulting in adversarial misclassification. We address this need in Chapter 3.

Another problem statement we tackle is how learning concepts in addition to pixel-based features can improve model accuracy, but we further discuss the limitations of saliency-based explainable AI in general. The importance of focusing on more than just a mathematical viewpoint of calculating feature importance using gradients and pixels is motivated by the origins of neural networks

as an imitation of the neurological capacities of the human mind: If we wish to emulate the abstraction capabilities of humans in varying contexts, we should also consider that human perception is highly dependent on context interpretation [Bie87].

Perhaps, in order to achieve human-level performance in classification problems, we have to train machine learning models with more consideration of human-like perception methods, such as concept abstractions. We discuss this observation in Chapter 4.

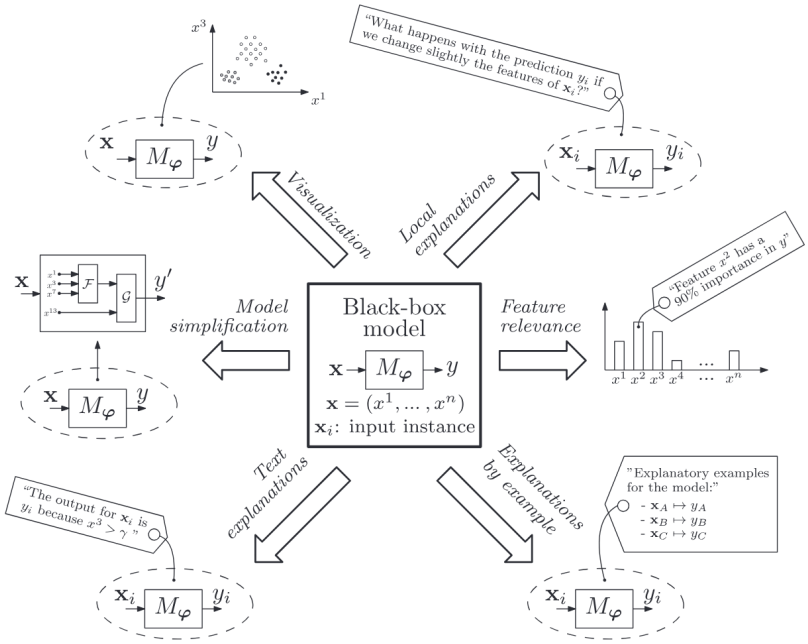


Figure 2.1: This portrays an overview of types of explanations from a conceptual overview review [Arr+20] and nicely summarizes recent developments in the explainability and interpretability domain.

To better contextualize the areas addressed in this thesis, we begin by outlining key methodological distinctions that help anchor the approaches under consideration. One key distinction by which explainability methods are classifiable is the design choice: transparency design means the model itself is explainable, such as decision trees [Xu+19]. In contrast, black box models require post-hoc explanations.

In addition, another key distinction is model-agnosticism versus model-specific explainable AI [Dwi+23]. Model-agnosticism denotes that the explainability algorithm is applicable regardless of the applied model [Arr+20]. Furthermore, explainable AI methods can be sorted into global versus local attribution or interpretability. Global interpretability aims to explain the overall model, whereas local explanations focus on single instances

[RSG16; Arr+20; LPK20], for example, image samples or other features. A summary of the distinctions within this domain is depicted in Figure 2.1.

When categorizing what explainable AI techniques are based on, the literature typically distinguishes feature-based, example-based, concept-based, rule-based XAI, surrogate explanations, counterfactual examples, and explanations by simplification [Min+22; Dwi+23; Hol+20]. Feature-based methods explain by attributing how much each feature contributes to a model output for a given data point [BWM20], including methods such as Shapley additive explanations [Mos+22], Grad-CAM [Sel+17], saliency maps, and more.

Example-based methods include prototypes such as PipNet [Nau+23] and ProtoPNet [Che+19], where the idea is to include a layer that aims to filter prototypical image parts via convolutions that are important for the classification task, showing us representatives of image-parts that were relevant for classification. Concept-based explanations include ACE [Gho+19], ProtoPDebug [Bon+22], CaCE [Goy+19], and many others, and aim to explain in more human-understandable terms by including concept-level abstractions rather than just pixel-based interpretation [Poe+23].

In this thesis, we will focus on a discussion of feature-based explanations and concept-based explanations in Chapter 4, highlighting the need for human interpretability rather than pixel-level interpretability. The following Chapters in this part are based on published work as cited below:

- Chapter 3, "Unsupervised DeepView: Global Explainability of Uncertainties for High Dimensional Data", is based on the peer-reviewed publication [NM22a];
- and Chapter 4, "Do you see what I see? An Ambiguous Animal Optical Illusion Dataset" is based on the work [New+25b].

2.1 EXPLAINABLE AI AND UNCERTAINTIES

Our main focus was the visualization of uncertainties. Uncertainty quantification is commonly divided into epistemic and aleatoric uncertainties [Soi17].

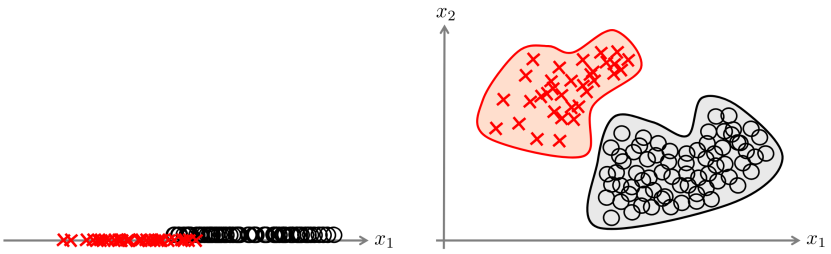


Figure 2.2: On the lefthand side, two classes overlap, causing aleatoric uncertainty. On the righthand side, embedding the data in higher-dimensional space allows for separation of the two classes, and the uncertainty is resolved. This example by Hüllermeier et al. [HW21] highlights how the distinction into aleatoric and epistemic uncertainty is highly context-dependent and can change the nature of aleatoric uncertainty into epistemic uncertainty.

Definition 2.1.1 (Aleatoric Uncertainty)**.** *Aleatoric, or sometimes denoted aleatory uncertainty—from Latin alea, meaning a die—refers to uncertainties that are inherently random by nature, for example, due to a process being nondeterministic. It refers to irreducible uncertainty that cannot be captured by more data [Soi17; GHO+17].*

Definition 2.1.2 (Epistemic Uncertainty)**.** *Epistemic—from Greek πιστημη, meaning knowledge—refers to uncertainty arising from lack of knowledge about the system [Soi17; GHO+17]. Epistemic uncertainty is often further subdivided into model form uncertainty, in which one doubts that the model is structurally correct and parametric uncertainty, where one assumes that the structure is correct, but it is uncertain what the exact parameters are [Sul15].*

The exact distinction between epistemic and aleatoric uncertainty is an imprecise one: The nature of uncertainties is highly context-dependent, as visualized in Figure 2.2. Furthermore, from a Newtonian standpoint, one could always argue that a dice roll does not reflect aleatoric uncertainty- perhaps it could be calculated with all complete information of the initial conditions of the die roll, such as the hand angle, wind, the force applied etc. [Sul15]. Therefore, in this thesis, we do not begin by dissecting the precise distinction between uncertainty and calibration, but rather focus on visualizing both aspects. Visualizing uncertainties provides critical insight into model reliability, and improvement strategies [RBP18; Tan+19], for example when challenged with poor quality data or small quantity datasets. It enables informed decisions [BBK19], fosters user trust [Abd+22], and is a cornerstone of reliable, interpretable AI [Lam+22; Deu+24].

2.2 METHODS FOR QUANTIFICATION OF UNCERTAINTY AND CALIBRATION

The starting point of this thesis is the following problem statement by Deuschel et al. [Deu+24]: We consider a supervised learning setting, specifically the task of classification in the image domain. Let $x_i, y_i \in (X, Y)$ be a data-label pair from the set of all training data X and labels Y. Let $X = \{x_i\}_i^N$ be the set of N training data points with corresponding labels $Y = \{y_i\}_i^N$, $y_i \in \{1, ..., K\}$ for K classes. Further, let $f_\theta : X \to \mathbb{R}^K$ be a deep learning function with learnable weights θ mapping to a K-dimensional logit space. That is, for a sample $x_i \in X$ we obtain logits $z_i = f_\theta(x_i)$, from which the probability for each class $k \in K$ can be derived after applying the softmax function σ [Deu+24]. The probability values $\hat{p}_i$ are calculated by

$$\hat{p}_i(k) = \sigma(z_i)^{(k)} = \frac{\exp(z_i^{(k)})}{\sum_{j=1}^{K} \exp(z_i^{(j)})} \tag{2.1}$$

and the maximum class probability is $\hat{p}_i = \max_{k \in 1,...,K} \hat{p}_i^{(k)}$. We now wish to visualize uncertainties in this setting. Consider the method Deepview [SHH20] as a starting point: We have a supervised setting for a visualization methodology of classification boundaries of deep neural networks using dimensionality reduction. We now wish to extend this intuitively to an unsupervised setting, where we only assume that we have the data points $x_i \in \mathbb{R}^d$. We further assume that we have some mapping $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ that outputs probability values $\hat{p}_i$ that reflect confidence scores. Our contribution is to eliminate the need for class labels $y_i \in Y$ by translating the K class classification problem into a two-class problem: uncertain and certain and introducing a quantifiable uncertainty score function using the estimated intrinsic dimensionality of the local as a proxy [Ma+18].

DeepView [SHH20] was the first attempt to visualize global decision boundaries of a deep neural network, in contrast to previous methods that explain single examples or their classification features. Other methods produce a projection of the data and generate global explainability; however, they either do not depict decision boundaries [Lap+19] or are not applicable to deep neural networks [SGH15] due to their density estimation approach in the input space. Furthermore, these methods cannot detect uncertainties by themselves as they rely on a human evaluation of the algorithm. In contrast, we aim at an unsupervised approach that lets the algorithm itself highlight where the uncertain regions are. We enable

an intuitive uncertainty estimation using the certainty of the unreliability in our process rather than the certainty of the class label, which is used by supervised approaches, as the background color to depict the robustness of the model in the form of adversarial examples.

3

UNSUPERVISED DEEPVIEW- GLOBAL EXPLAINABILITY FOR HIGH DIMENSIONAL DATA

In recent years, more and more visualization methods have been proposed for the explanation of artificial intelligence that focus on untangling black-box models for single instances of the data set [Sel+17; RSG18]. Although the focus often lies on supervised learning algorithms, the study of uncertainty estimations in the unsupervised domain for high-dimensional data sets in the explainability domain has been neglected so far. As a result, existing visualization methods struggle to visualize global uncertainty patterns on whole datasets. We propose Unsupervised DeepView, the first global uncertainty visualization method for high-dimensional data based on a novel unsupervised proxy for local uncertainties.

In this paper, we exploit the mathematical notion of local intrinsic dimensionality. As a label-agnostic measure of model uncertainty in unsupervised machine learning, it shows two highly desirable features: It can be used for global structure visualization as well as for the detection of local adversarials. In our empirical evaluation, we demonstrate its ability both in visualizations and quantitative analysis for unsupervised models on multiple datasets.

3.1 INTRODUCTION

Visualization of raw data, pre-trained models, and uncertainties of these models are essential methods for explaining machine learning models. In the area of supervised models, such visualizations of a classification model and its uncertainties are widely studied [Ant+20a; Zha+19; Lap+19], this has only been tackled less extensively for unsupervised learning methods [MCM19]. However, it has been neglected entirely for the combination of unsupervised models such as clustering [AR14] or anomaly detection [AY01] and global explainability.

This is due to the fact that unsupervised models do not have labeled training data that allow us to directly evaluate misclassifications or model errors. Similarly, uncertainties are challenging to quantify for a model without known (labeled) patterns. Furthermore, it is easier to depict uncertainties for specific data points

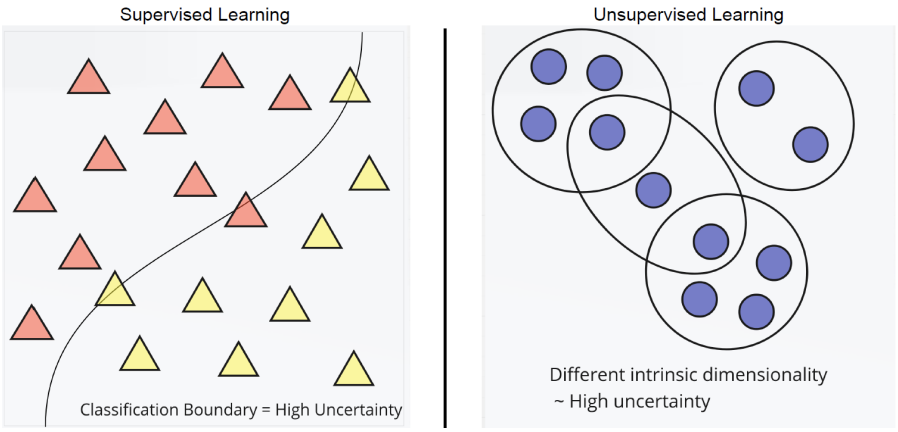


Figure 3.1: This Figure shows the difference between uncertainties for supervised models versus uncertainties in the unsupervised domain. Data areas along the classification boundary are most likely misclassified in the supervised domain. In the unsupervised domain, the uncertainty is not as clear. Here, we define our uncertainties by measuring the local intrinsic dimensionality of the data. When this dimensionality varies considerably from its neighbors, we define this as an uncertain area for our visualization scheme.

rather than reliably approximate uncertainties for the entire model. Recent existing visualization methods for high dimensional data [Tha+17; SHH20; VH08] rely on labeled data and are not able to visualize the hidden (unknown) patterns of data in case of unsupervised learning.

The key challenge is the lack of an objective function that, in the case of supervised learning, describes the separability of one or multiple classes, as depicted in Figure 3.1 on the left. In contrast to this, unsupervised measures such as cluster validation [Liu+10], clusterability scores [AB09], or anomaly scores [AA17] are designed to evaluate either the quality of clustering (globally for the entire dataset) or the detection of rare and exceptional outliers (locally for individual objects).

In contrast to unsupervised measures, we aim at both a global visualization of raw data structures and a local uncertainty of specific data areas. Neither the one nor the other should rely on labels but are purely based on the given pre-trained unsupervised model and a proxy for local uncertainty quantification.

Our research focuses on unsupervised knowledge discovery and defines novel proxies for uncertainties. In this paper, we exploit the mathematical notion of local intrinsic dimensionality as a proxy for the complexity of the data distribution. As depicted in Figure 3.1 on the right, we define local divergence in intrinsic dimensionality

as the label-agnostic measure of high model uncertainty in unsupervised machine learning. It allows for both local assessments of uncertainty for a specific data area and, at the same time, for global visualization of data structures vs. adversarial examples in less certain data areas.

We have implemented the first unsupervised DeepView approach that visualizes data structures and adversarials without any prior labels. It requires only a pre-trained unsupervised model and its uncertainties. This framework calculates the local intrinsic dimensionality of the data points and creates a mapping of uncertainties using dimensionality reduction [MHM18] on the combination of the data, their prediction uncertainties, and their dimensionality. We then use outlier detection algorithms on this data embedding to depict our uncertainties with a high empirical precision score.

Our paper solves the task of visualization of global and local uncertainty for unsupervised learning. In contrast, all other known methods either solve global supervised explainability [Tha+17; Lap+19; SHH20] or focus on local explanations [Zha+19; LL17c; RSG18] with few unsupervised methods [MCM19] only.

3.2 VISUALIZING THE UNCERTAINTIES OF AN UNSUPERVISED LEARNER

In this work, we consider uncertainties w.r.t. (A) properties of the raw data distribution as well as (B) properties of unsupervised machine learning models that have been pre-trained on these data. We define uncertainties as more than the misclassification of supervised models. In our definition, uncertainties are caused by the complexity of data, for example, uncertain areas following the phenomenon of empty space in high-dimensional data [LV07].

Similarly, based on a pre-trained model, uncertainties may be areas of high probability for adversarial examples. However, in our definition, adversarial examples are not depicted by wrong class labels. We observe unsupervised adversarial examples in the training of unsupervised models (e.g. adversarial examples in autoencoders [Böi+21]).

Hence our methods tackle two main challenges: First, our unsupervised approach's labels and amount of classes are unknown. We cannot assess traditional discrepancies between what a model predicts as uncertain and what is really uncertain or incorrect. Without supervision, we are forced to depict uncertainties using

novel unsupervised measures and propose a local comparison of this measure with the object's local neighborhood.

Second, existing supervised methods such as DeepView [SHH20] rely on prediction probabilities for classes, which are then used to visualize the uncertainty of class prediction. In our case, we assume the knowledge of those prediction uncertainties for our model and use them as input into the visualization method. Formally, we describe our abstract notion of unsupervised uncertainties as follows:

Definition 1
(UNSUPERVISED QUANTIFIABLE UNCERTAINTY)
We define the input space $\mathcal{X} \subseteq \mathbb{R}^d \quad \text{with} \quad d = H \times W \times C$, and the normalized hdbscan outlier function $\text{score}_{\text{HDBSCAN}}(z_i) \in [0, 1]$. The unsupervised quantifiable uncertainty denotes

$$\text{UQU} : \mathcal{X} \to \mathbb{R} \in [0, 1]$$

$$\text{UQU}(x_i) := \text{score}_{\text{HDBSCAN}} \left(\text{UMAP} \left([x_i, \hat{\text{LID}}(x_i)], \max \hat{y}_i \right) \right)$$

with $\phi(x_i) := [x_i, \hat{LID}(x_i)] \in \mathbb{R}^{d+1}$ being the feature augmentation of input x_i with LID. $\hat{y}_i := model(x_i) \in \mathbb{R}^c$ is the models softmax prediction for c classes. $\max \hat{y}_i \in [0, 1]$ describes the confidence score for the model visualized.

Intuitively, this function consists of two components: $l_{model}(x_i) := \max \hat{y}_i$ as a score for the likelihood that the sample x_i is out-of-distribution for the particular model, suggesting a misclassification. The second component captures a score function that correlates with the likelihood that the sample x_i is an adversarial example with the empirical proxy of local intrinsic dimensionality. This correlation was first introduced by Ma et al. [Ma+18]. The unsupervised quantifiable uncertainty function, however, captures both intuitions, because the outlier detection algorithms receive both the confidence scores and the LIDs concatenated to the original x samples.

An adversarial example is commonly described in literature as an optimization problem for a classifier $f : \mathbb{R}^m \to \{1...k\}$, which maps image pixels to a discrete label set [Sze+13]. f has a continuous loss function denoted by $loss_f : \mathbb{R}^m \times \{1...k\} \to \mathbb{R}^+$. An adversarial

example solves the following optimization problem for an input $x \in \mathbb{R}^m$ and target label $l \in \{1...k\}$:

Minimize $||r||_2$ subject to

1. $f(x+r) = l$
2. $x + r \in [0,1]^m$

Note that l is not the original label, so we want to solve for $f(x) \neq l$. On an intuitive level, the most common definition of an adversarial example is the smallest perturbation to an image that leads to a misclassification by a classifier. In practice, it is very hard to determine whether an image is simply a slight perturbation from another image in a dataset, especially if we generate an adversarial image from an image that is not part of the training dataset.

Mostly, adversarial examples differ from normal misclassifications in their certainty in the correctness of the prediction [ZBZ23], which is why we include the certainty for the calculation of the unsupervised quantifiable uncertainty.

We include local intrinsic dimensionality as an estimate of the likelihood of an adversarial attack. Although this does not constitute a formal likelihood function, outliers in local intrinsic dimensionality have been shown to correlate with a higher probability of being adversarial compared to non-outliers, as demonstrated in [Ma+18]. In practice, this means that we fail to distinguish between simple misclassifications that happen to have high certainty and real adversarials in the form of input perturbation attacks, but since we aim to approximate the uncertainty of a model as an end goal, this explicit distinction is not necessary.

Both are very dangerous for real-world applications and capture the essence of uncertainty, especially for high-risk scenarios such as autonomous driving. For our empirical evaluation, we simply compare whether our unsupervised approximation derived from our model is in line with the supervised misclassification (using labels). In real-world use cases, this would not be possible; however, our datasets have given labels that are ignored by our algorithm but useful for external evaluation.

3.3 UNSUPERVISED DEEPVIEW

We propose a generalized visualization technique to depict uncertainties and potential adversarial examples in unsupervised learning. Our framework consists of three components:

1. Local Intrinsic Dimensionality (cf. Section 3.3) as our main uncertainty proxy of the data distribution.
2. Dimensionality Reduction (cf. Section 3.3) that allows for 2D visualization of high-dimensional data.
3. Adversarial Detection (cf. Section 3.3) based on an outlier analysis of our uncertainty measure.

In the unsupervised framework, all three of these components can be exchanged in future work. In the following subsections, we present our first instantiation of each of these in the DeepView visualization [SHH20]. This allows for the first time a fully unsupervised analysis and visualization of uncertainties.

Please note that our method does not just show the quality of the decision boundary in a supervised learner. We extract and measure uncertainties on the basis of unlabeled data. To the best of our knowledge, this is the first uncertainty visualization method that focuses on global interpretability rather than explanations of individual predictions in the unsupervised learning domain that can visualize a smooth two-dimensional manifold of the uncertainties on high-dimensional data such as natural images.

ALGORITHM OVERVIEW Our estimation and visualization algorithm must be applicable to any unsupervised task or usable whenever the probabilities of an unsupervised model are given on a randomly sampled data set. We aim to highlight uncertain regions of samples that have a high likelihood of being adversarial examples but in an unsupervised setting, while also visualizing regions of high certainty and local uncertainty. To accomplish this, we propose the following algorithmic steps:

1. Calculate the local intrinsic dimensionality (LID, cf. Section 3.3) of each data point $\forall x_i \in S$ in the sample S compute $LID(x_i) \in \mathbb{R}_{\geq 0}$,.
2. Apply dimensionality reduction (UMAP, cf. Section 3.3) on the given input of LIDs together with unsupervised model uncertainty. We use the confidence score for the model as a supervision for UMAP. We project these three values together as x_i into two dimensions, resulting in $y_i = \pi(x_i)$.
3. Create a tight grid of samples r_i in two-dimensional space and map this to high-dimensional space.
4. Outlier detection (cf. Section 3.3) algorithm on uncertainty measures, resulting in binary detection of certain areas and

uncertain areas. We define the uncertain area by a high LID mean.

5. Visualize the outlier scores and interpret them as unsupervised uncertainties.

In contrast to our unsupervised algorithm, the supervised DeepView method [SHH20] is composed of four algorithmic steps that enable visualization of decision boundaries:

1. Application of the dimensionality reduction technique Fisher UMAP [MHM18] to project data points x_i to two dimensions, yielding $y_i = \pi(x_i)$.
2. Creation of a tight grid of samples r_i in two-dimensional space and mapping to high-dimensional space.
3. Application of the network to this mapping to obtain predictions and certainties.
4. Visualization of the labels together with the entropies of the certainties.

Ma et al. [Ma+18] already showed that local intrinsic dimensionality measures help characterize adversarial subspaces. While there has also been research on the limitations of these features for the characterization of adversarial examples [LCY18], one of the main critique points was the non-transferability of the LID features to other deep neural networks. This point does not affect the quality of our visualization method, considering that it is used on a specific model and does not require transferability to other models.

The second criticism expressed in this article was that the quality of LIDs as features for adversarial attacks with the trained detector algorithm suggested by Ma et al. [Ma+18] varies with the confidence parameter, and the training of ensembles of adversaries with different confidence levels did not help detection performance.

However, our approach does not rely solely on LIDs. Instead, it uses the dimensionality reduction technique based on UMAP [MHM18] suggested by the DeepView algorithm [SHH20] and searches for outliers on this mapping of the data points, LIDs and prediction certainties. It is important to note that the specific outlier detection algorithm used within the framework is interchangeable and can be replaced with any method that demonstrates optimal performance for the given dataset.

This means that we take into account the uncertainties as well as the LIDs, giving us a better chance of extracting adversarial

or uncertain regions. The focus of our method is not restricted to adversarial detection but rather includes the broader identification of uncertain or ambiguous prediction regions. Furthermore, there is currently no other implementation of a global visualization scheme of the decision boundary of uncertainty in the unsupervised domain. We later show that our approach gives us good approximations of the uncertainties within an unsupervised learner.

LOCAL INTRINSIC DIMENSIONALITY The intuition behind this metric is to measure the increase of data objects encountered, estimating the dimensionality of the structure of the data. Transferring the idea of expansion of dimensions to distance distributions gives a formal definition of LID [Hou17].

Definition 2 (LOCAL INTRINSIC DIMENSIONALITY)
Given a data sample $x \in X$, let $R > 0$ be a random variable denoting the distance from x to other data samples. If the cumulative distribution function $F(r)$ of R is positive and continuously differentiable at distance $r > 0$, the LID of x at distance r is given by:

$$LID_F(r) \triangleq \lim_{\epsilon \to 0} \frac{\ln(F((1+\epsilon) \cdot r)/F(r))}{\ln(1+\epsilon)} \tag{3.1}$$

whenever the limit exists [Ma+18]. The maximum likelihood estimator (MLE) of the LID at x given a reference sample drawn from the representation of the data distribution P is defined as follows:

$$\widehat{LID}(x) = -\left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{r_i(x)}{r_k(x)} \right)^{-1} \tag{3.2}$$

Where $r_i(x)$ denotes the distance between x and its $i-$th nearest neighbor within a sample of points drawn from P, and $r_k(x)$ is the maximum of the neighbor distances. Although this computation can become computationally expensive with the increase of the neighborhood, Ma et al. [Hou17] show that discrimination of adversarial and non-adversarial examples turn out to be possible for minibatch sizes of 100 and neighborhood sizes as small as $k = 20$, rendering our computational estimation feasible. Therefore, these are the parameters we also use to implement our visualization scheme.

DIMENSIONALITY REDUCTION The goal of dimensionality reduction techniques for visualizations is to find mappings $\pi : (S, d_s) \to \mathbb{R}, d = 2, 3$, where (S, d_s) is a metric space and π ide-

ally preserves the information encoded in a set of data points $x_1, ..., x_n \in S$. The paper is based on a dimension reduction technique called UMAP [MHM18], which performs at least equally well as the state-of-the-art non-linear dimensionality reduction method t-Distributed Stochastic Neighbor Embedding (t-SNE) [VH08], but allows for the inverse projection suggested by the original DeepView implementation. It allows for our two-dimensional visualization. Theoretically, this method could be exchangeable as long as inverse mappings can still be established with other algorithms.

ADVERSARIAL DETECTION USING LOCAL INTRINSIC DIMENSIONALITY It is important to note that detecting unsupervised adversarial examples cannot be effectively achieved through standard clustering algorithms, primarily due to class imbalance in the data. In well-performing models, uncertain predictions are considerably less frequent than confident ones.

Instead, our goal is to identify outliers—data points whose local intrinsic dimensionality (LID) values and prediction confidences deviate significantly from the overall distribution, indicating potential anomalous or adversarial behavior. As Ma et al. state in their paper [Hou17], adversarial examples have noticeably higher LID characteristics than normal examples. This means that we can distinguish adversarials using outlier detection algorithms. We follow the intuitive definition of an outlier as given by Hawkins [Haw80] and search for outliers that deviate so much from the other observations as to arouse suspicion that they are generated by a different mechanism.

We detect outliers based on density-based clustering algorithms that allow us to distinguish certain and uncertain areas, but also consider that adversarials or uncertainties have to be detected for both single-point outliers and cluster-based outliers. We refer to the definition of cluster-based outliers of LDBSCAN [Dua+09] and its extension HDBSCAN [MHA+17]. Compared to other cluster-based outlier detection methods, the advantage of these two algorithms is the quantifiable outlier score, which intuitively corresponds to the degree of the outlying object.

In the case of HDBSCAN, it is referred to as an outlier score. These scores can be incorporated into our visualization of the decision boundary as the certainty of the outlier. The difference between HDBSCAN and DBSCAN is that HDBSCAN performs a hyperparameter search of the ϵ parameter, namely the radius from at least one cluster point to another, of LDBSCAN without having to preset it, therefore only needing a minimum cluster size as an

input parameter. This is why our final implementation uses the HDBSCAN algorithm without comparing the results of LDBSCAN.

RUNTIME ANALYSIS The runtime of our algorithm, including the three mentioned steps, is similar to the supervised DeepView implementation. The extra effort in calculation consists of a one-time precomputation of first the LIDs over the dataset with a neighborhood size of 20 each, dimensionality reduction via UMAP over the inputs, LIDs and the outputs of the model for all inputs, followed by outlier detection over the embedding.

The extra run-time depends on 1. the run-time for the computation of the empirical LID estimate, 2. the computation time for the dimensionality reduction method chosen, and 3. the outlier detection method applied. The dominating terms in this case for Unsupervised Deepview are the runtime for UMAP [MHM18], which is $\mathcal{O}(N_1 \log N_1)$ according to the authors [MHM18], and the worst-case runtime for HDBSCAN, which is $\mathcal{O}(N_2 \log N_2)$ on average and $\mathcal{O}(N^2)$ in worst case [MHA+17]. N_1 is the number of input vectors that UMAP must embed, and N_2 denotes the total number of input points for the HDBSCAN algorithm. However, this is a one-time precomputation step, and then the run-time of DeepView and Unsupervised Deepview are the same, with the number of classes for Unsupervised Deepview limited to the two classes uncertain and certain, regardless of the amount of classes for the predictions. Preliminary experiments have shown very similar run-times for Deepview and Unsupervised Deepview for the datasets that we empirically tested in practice. Hence, we focus on qualitative and quantitative evaluation without deepening the topic of runtime evaluation.

3.4 EXPERIMENTS

In this section, we apply the new Unsupervised Deepview implementation[1] and evaluate how well we capture uncertainties. We measured uncertainties as either supervised misclassifications or adversarial examples and showed exemplary applications of our method to CIFAR-10 and Fashion-MNIST. We used models based on the resnet50 convolutional architecture with some additional convolutions and a linear layer. As you can see later in Table 3.1, the algorithm also performs excellently for models and data sets with very high precision scores over several runs.

1 The code for the project can be found at our chair's collective repository https://github.com/KDD-OpenSource/Unsupervised-Deepview/.

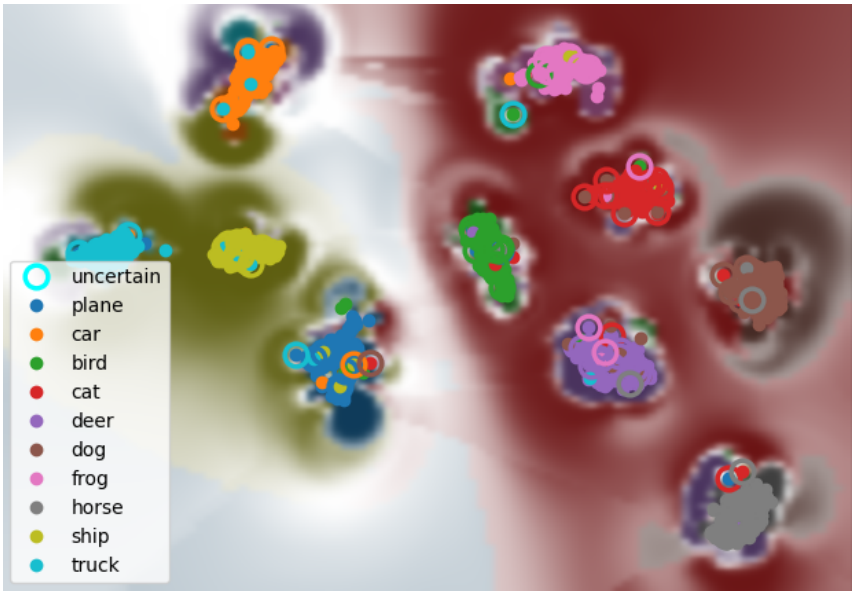


Figure 3.2: The original Deepview implementation achieves a two-dimensional visualization of the decision boundary of a supervised classifier. This picture shows DeepView, but for evaluation purposes, we inserted and marked the points that were detected by our method Unsupervised Deepview as "uncertain" using blue circles. As we can see, we were perfectly able to distinguish an adversarial example labeled as a bird, even though it was actually a frog (green isolated dot near the pink cluster), also visualized in Figure 3.3 from the CIFAR-10 data set. In the same picture, we also see the edges of the truck cluster marked, where a misclassified car and plane are nearly hidden (bright blue cluster with dark blue and orange spots). The other point marked as uncertain is a correctly labeled plane with low uncertainty values.

To evaluate our visualization scheme, we look at the following questions: (i) Does our combination of outlier detection algorithms, LID features, and prediction probabilities capture the same or similar uncertainties to the supervised version of the DeepView implementation without using labels? (ii) Are the LID features necessary for our implementation to perform well, or could we also just detect outliers using the uncertainties of the prediction alone? (iii) How well are we able to distinguish model uncertainty or adversarials without the knowledge of the label?
Addressing our first question, we compare the original DeepView visualization technique with labels with our unsupervised DeepView method by marking the points recognized as uncertain in the original DeepView plot in Figure 3.2. Here, we can see that while not all uncertainties were detected, the points marked as uncertain turned out to be either misclassifications, adversarials, or one point with just low model certainty. We were able to quan-

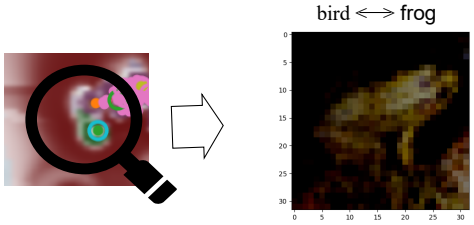


Figure 3.3: When selecting the point instance given as uncertain, the original instance of the image will be shown. On the left-hand side, the model-predicted label is shown, and on the right, the actual label is given. For the Unsupervised algorithm, we will only output certain or uncertain labels, for instance. The goal is to allow the user to directly assess the instances visually so as to better understand why the model classified them as uncertain.

tify misclassifications despite not using the labels to generate our uncertainty predictions. However, we knew the actual labels of the dataset because we evaluated over a labeled data set to test the quality of our model. A zoomed-in view of the uncertainty and its original image point can be found in Figure 3.3. It is visualized in the implementation when tapping on the data point in the plot given.

Furthermore, we compare the points detected as outliers using only the data points and prediction uncertainties, without the LIDs as features in Figure 3.4. As you can see here, the points marked as uncertain no longer map to adversarial or misclassifications. This gives us an intuition that the LIDs work as features to compensate for the actual class labels and shows that we can correctly identify actual adversarial examples within the dataset.
Addressing (iii), our precision score is consistently high. Although the method does not detect all uncertainties, the uncertainties we do detect are either misclassifications or adversarial examples with the following percentages (cf. Table 3.1).

Table 3.1: Precision scores over 10 runs

Dataset	**Precision**
CIFAR-10	98.4 %
Fashion-MNIST	97.3%

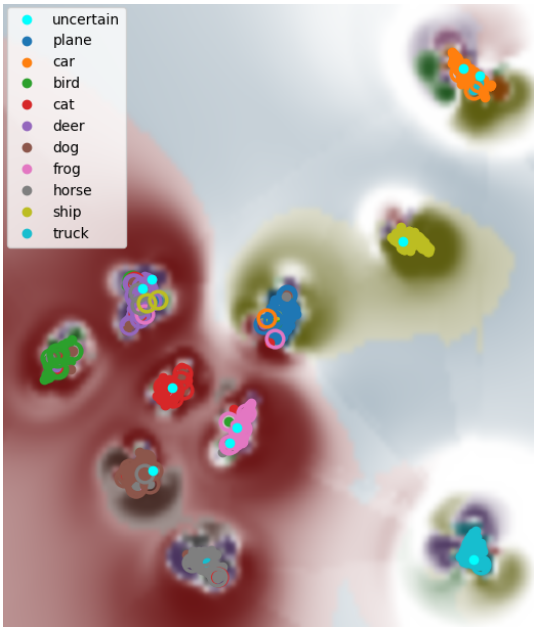


Figure 3.4: For evaluation purposes, we also tested whether the LID features were essential for visualization of uncertainty without knowledge of the labels. Here, we just marked the points detected by the HDBSCAN [MHA+17] algorithm without the LID features and without knowledge of the labels in the picture generated by the original DeepView algorithm. As we can see, the uncertainties are not captured well. Often, the centers of the class clusters are marked, whether they are uncertain or not.

The final output of our visualization method can be found in Figure 3.5. Here, we can see data points denoted as certain and uncertain, as well as our decision boundary colored in by the strength of the blue note. Darker background areas indicate more certain areas in the data for the specific model.

3.5 CONCLUSION AND FUTURE WORK

In this paper, we propose Unsupervised Deepview, a method that allows the depiction of a smooth dimensional manifold of uncertainties for high-dimensional data. To the best of our knowledge, it is the first method generally applicable to all unsupervised learning algorithms that provide uncertainties in the unsupervised domain. In contrast to recent methods such as GLAM-CLUE [LBW22] our method does not generate counterfactuals but shows the decision boundary of the uncertainties of the model. We do not require any labeled data as we exploit the mathematical concept of local intrinsic dimensionality as a local proxy. Our global visualization of data structures versus adversarial examples provides for the first time

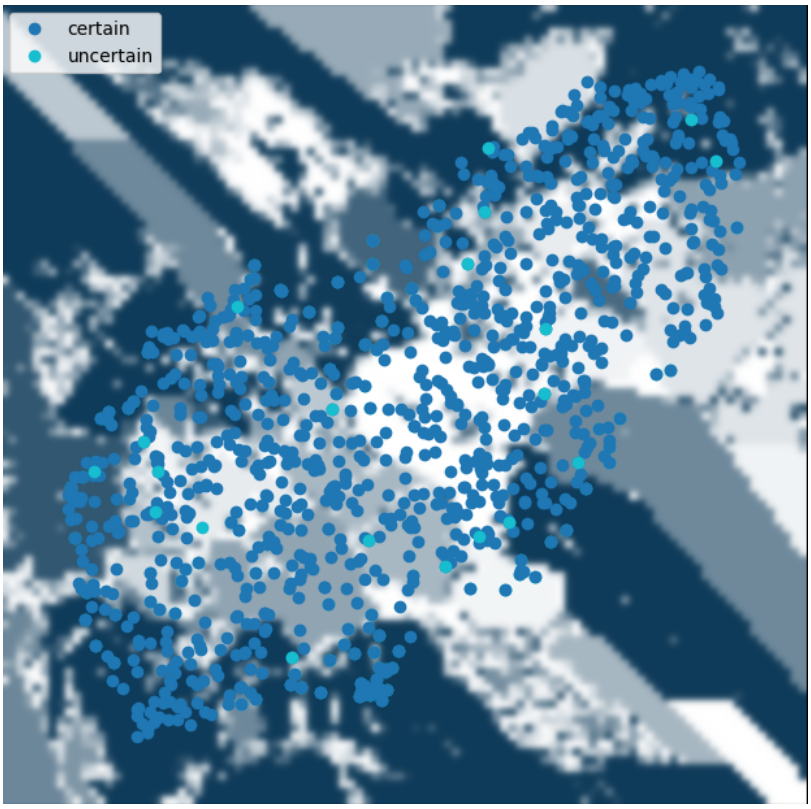


Figure 3.5: Our Unsupervised DeepView implementation achieves a two-dimensional visualization of the decision boundary of uncertainties. We do not claim that it identifies all model uncertainties. Still, those it detects are either uncertain in their prediction, misclassifications, or adversarials which we can evaluate with a very high precision score. The background coloring shows estimations over unseen data. Furthermore, our visualization method does not aim to identify them directly but rather outputs the areas in which predictions become uncertain or where the classifier performs well. Dark areas denote particularly certain areas, whereas whiter areas are particularly uncertain. Our uncertain data points are colored light blue.

unsupervised insights into the uncertainty and vulnerability of machine learning models. It depicts the whole decision boundary rather than the single decisions of the model.

In our empirical evaluation, we present the precision for CIFAR-10 and Fashion-MNIST. We evaluated it using the pre-trained residual network with 20 layers discussed in the original paper to show that the algorithms predict similar uncertainties despite the lack of labels. Secondly, we use a different convolutional neural network pre-trained on Fashion-MNIST to confirm our results.

As a first approach in this area, we see future work to extend our method to data beyond high-dimensional vector spaces. For example, graph data that has inherent local and global structures would benefit from our methodology but requires specific graph measures as local proxies of structure and uncertainty. Furthermore, we see improvement potential for the usability of the tool or practicability of trust calibration—Overall, this explainability method is designed not just for a technical audience but should be used to provide a better estimate than just accuracy estimates or other common proxies for the reliability of a model on a specific data set.

4

DO YOU SEE WHAT I SEE? AN AMBIGUOUS OPTICAL ILLUSION DATASET EXPOSING LIMITATIONS OF EXPLAINABLE AI

From uncertainty quantification to real-world object detection, we recognize the importance of machine learning algorithms, particularly in safety-critical domains such as autonomous driving or medical diagnostics. In machine learning, ambiguous data plays an important role in various machine learning domains. Optical illusions present a compelling area of study in this context, as they offer insight into the limitations of both human and machine perception. However, optical illusion datasets remain scarce.

In this work, we introduce a novel dataset of optical illusions featuring intermingled animal pairs designed to evoke perceptual ambiguity. We identify generalizable visual concepts, particularly gaze direction and eye cues, as subtle yet impactful features that significantly influence model accuracy. By confronting models with perceptual ambiguity, our findings underscore the importance of concepts in visual learning and provide a foundation for studying bias and the alignment between human and machine vision. To make this dataset useful for general purposes, we generate optical illusions systematically with different concepts discussed in our bias mitigation section. The dataset is accessible in Kaggle via `https://www.kaggle.com/datasets/anonymac12i3/ambivision`.[1]

4.1 INTRODUCTION

The motivation for this particular optical illusion dataset stems from a novel problem in the explainability domain (XAI). Explainable AI uncovers the black-box nature of models and visualizes the internal workings of a machine learner for image data. So far, research has focused on highlighting the pixels that are the most important or influential in the decision-making process [RSG16; Sel+17]. Other approaches highlight important regions [RSG18] or target the most essential features [LL17c; STY17].

1 The source code for this project can be found at https://github.com/KDD-OpenSource/Ambivision.git

However, these methods often fall short when confronted with perceptual ambiguity—situations where even human interpretation is uncertain. Take a look at Figure 4.1.

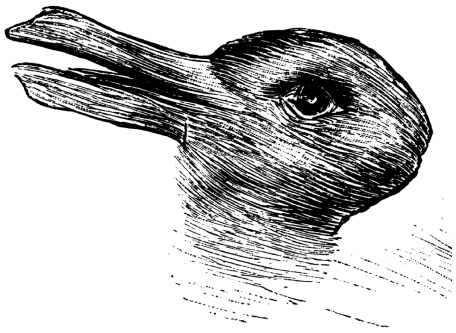


Figure 4.1: In this optical illusion, you can see both a rabbit and a duck. Common XAI methods that highlight important pixels could output exactly the same explanation for either of those classes without improving human understanding of which class was chosen why. This is a critical research gap in explanations—pixel highlighting is simply not enough.

Depending on the viewer's perception of the eye's direction, the image may be interpreted as either a rabbit or a duck. If you were to use any current XAI algorithm to generate explanations of why it is a rabbit or a duck, the explanations could look exactly the same. They would keep the actual reason behind this optical illusion secret. This is because current XAI methods highlight important pixels or areas, but in this case, all features are shared by both classes.

This suggests that standard XAI methods are currently inadequate when semantic interpretation relies on abstract perceptual cues rather than pixel-level interpretation. Examples of this phenomenon in the provided dataset can be seen in Section 4.3. Clearly, our internal decision process goes beyond what we see—resolving ambiguity by assigning direction, intent and anthropomorphizing [WC21]. One such so far neglected concept is the viewing direction of the animal, and not just where we look at as the supervisor.

In this paper, we introduce three major contributions: First, we expose the existing methodologies of XAI by showing limitations: Highlighting pixels or areas alone, at least in the image domain, is not enough. Second, we introduce a new open-source optical illusion dataset featuring animal images labeled with bounding boxes, gaze, and viewing direction annotations. And third, we demonstrate that integrating gaze direction and eye coordinates into the learning process improves model performance, even when all other aspects (architecture, epochs, learning rate, dataset struc-

ture, optimizer) remain the same. Our novel dataset is generated with sophisticated ChatGPT [Ope24] models and considerable computational power. It incorporates the understanding of a generative AI model in generating optical illusions while achieving images that are difficult for humans. However, we address how we mitigated potential biases in Section 4.7. Furthermore, the difficulty of generating convincing optical illusions makes it hard to provide large datasets that are drawn by human artists: We argue that a machine generated set offers a valuable perspective into human vs. AI perceptual learning.

4.2 RELATED WORK

Explainable AI, often shortened to XAI, is broadly considered to be split into two categories: transparency design and post-hoc explanations [Xu+19]. We criticize that the current approaches of visualization methods in the image domain highlight pixels or areas of images that show the importance of specific features. Saliency-based methods, such as LIME [RSG16], Grad-CAM [Sel+17], or Anchors [RSG18] offer local explanations by attributing predictions to pixel importance or localized regions.

The broader landscape of explainable AI encompasses countless methods [LL17a] from counterfactuals [MST20; Ant+20a], prototypes [Nau+23; Che+19], to global explainability methods [Set+21; MCM19; NM22a]. We highlight the need for concept-based XAI in ambiguous settings. One type of concept-based XAI works is based on predefined concepts and relies on human supervision [Kim+18; Bon+22; Yeh+20; Goy+19].

Hence, concept-based explanations were extended to automatic concept-based extraction, which relies on segmentation strategies that then employ importance scores to eliminate outliers [Fel+23a; Gho+19; Zha+21; Fel+23b]. However, this type of approach also comes with its limitations: We argue that all of these approaches still segment pixel-based logic. We apply ACE [Gho+19] to our dataset as a prominent representative of the field and show with examples that it also has problems extracting useful concepts.

We demonstrate in this work that general concepts for domains exist, such as gaze and the eye coordinates, that help the overall performance on a whole domain rather than just a specific type of animal. These concepts were not spotted when applying ACE. In our experimental section in Section 4.5, we show the improvement

when those concepts are considered versus learning without the concepts on our ambiguous dataset.

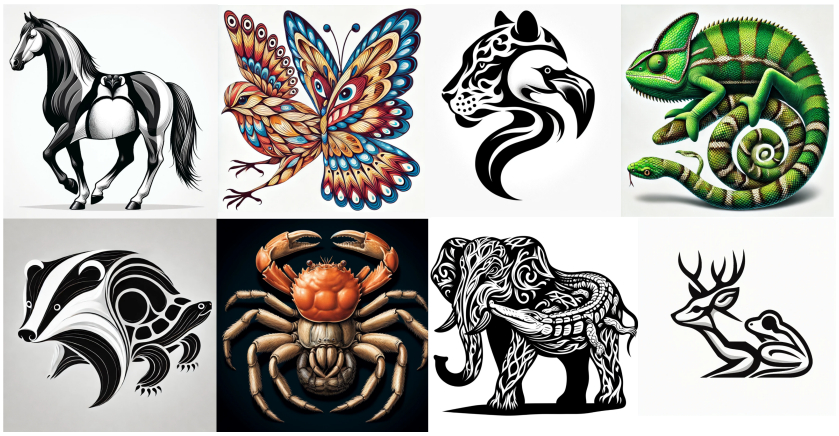


Figure 4.2: This Figure features several examples of our dataset: On the upper left side, a penguin can be seen hidden within a horse, depending on the direction we consider the animal to be looking. All of these examples have two animals distinguishable by the eye coordinate and the gaze vector, meaning they might be looking in the same direction, but their right eye (if more than one is visible) is positioned somewhere differently. The goal of this dataset was to test whether the gaze and eye coordinates prove to be useful general concepts in ambiguous settings but can, in general, be used for the evaluation of XAI as baseline for classification performance with optical illusions. More example pictures that show the diversity of illusions within one class are included in Figure 4.7.

4.3 LIMITATIONS OF CURRENT EXPLAINABLE AI ALGORITHMS

While current explainable AI algorithms do a phenomenal job explaining attributions of models in general, our dataset is intentionally constructed to challenge local attribution methods by presenting overlapping or intermingled visual features from multiple animals. As we can see in the following, we provide some example areas where the highlighted features of the dataset mark features that do not belong to one of the animals specifically.

In the following, we show instances where common XAI algorithms fail. We chose popular representatives from the saliency-based XAI area, such as Grad-CAM [Sel+17] and integrated gradients. We then present an example using a prototypical XAI method, namely PipNet [Nau+23]. We do the same for a representative of the automatic concept-based XAI area, ACE [Gho+19]. We show that these methods fail to distinguish the two animals well due to the shared features of the animals: These limitations underscore a

critical shortcoming of local attribution techniques and emphasize the need for concept-level reasoning.

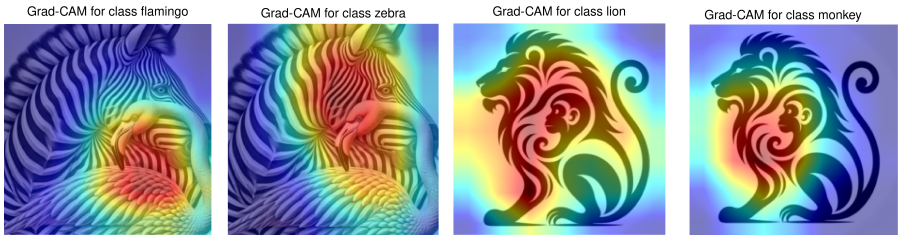


Figure 4.3: Pixel-based attribution explanations like Grad-CAM struggle to distinguish between the intermingling areas of the two animals. We later show that adding a single feature significantly enhances performance on ambiguous data.

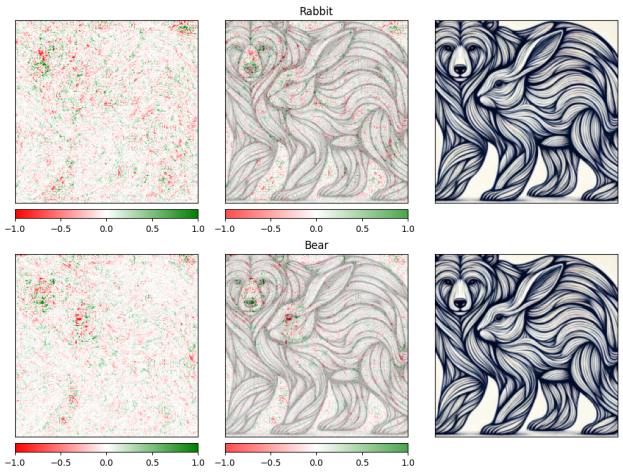


Figure 4.4: The same can be spotted for example using Integrated Gradients: The attributions for the classes are very similar, and often not very sensible. The explanation for bear clearly marks the rabbits' head in the picture. Both the rabbit and bear explanation include markings on the bear face area. Clearly, the model struggles to distinguish the two animals, and the explanations are limited in their meaningfulness and clarity.

Additionally, we show the same for PipNet [Nau+23], a prototypical XAI algorithm. In Figures 4.5 and 4.9, we can see that the prototypes marked by the algorithm contain all kinds of animal examples, including eagle feathers (which were intentionally designed to look similar to confuse). One of the prototypes even includes tiger eyes, if you look closely. The explanation of PipNet marks both the cheetah fur and the eagle patterns because they admittedly look very similar. However, this proves our point: As humans, we draw an invisible boundary in our head between the cheetah and the eagle, because we think in concepts [Bar04]. The similarity of the fur does not lead to misclassification when the

human is asked, he or she can still distinguish the two animals. Furthermore, we then tried to evaluate how concept-based explain-

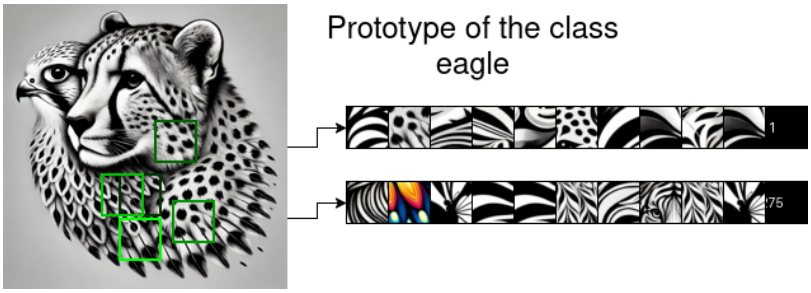


Figure 4.5: In this image, we see prototypes extracted via PipNet [Nau+23] for the eagle class. PipNet also struggles to distinguish the cheetah fur and the eagle feathers. The darker the box, the more important it is for the classification. Again, we argue that this is due to the area-based approach, a clear limitation for ambiguous data. Next to the original image, we see example concepts extracted that should show similar features.

able AI methods, such as ACE [Gho+19], perform in the automatic detection of concepts in these ambiguous settings. Although concept-based explanations generally require human-annotated concepts, ACE promises to automatically detect concepts using segmentation and clustering techniques using convolutional neural networks. When applying ACE to our dataset, it does not extract really meaningful concepts: For example, looking at the bird class in Figure 4.6, ACE does not extract anything useful from the leftmost image; the middle detected the dots on the bird wings, but not all birds have dots on their wings.

The rightmost image might be the wings, but it also highlights something insignificant at the bottom of the image. Furthermore, ACE fails to find the two general concepts that we suggest work well: gaze direction and the eyes. We argue that this is because, again, ACE employs segmentation on a pixel level and derives concepts from there. This general setup is shared among automatic concept-detection methods [Fel+23a; Zha+21; Fel+23b]. We argue beyond pixel-level segmentation and towards concept-level extraction. However, current literature that calls itself concept-based XAI is still essentially pixel-based, and we show the limitations in Figure 4.6.

IMPROVING CLASSIFIER PERFORMANCES USING ADDITIONAL KEY FEATURES SUCH AS GAZE ANNOTATIONS Apart from being beneficial in this specific example, gazes already prove useful

Figure 4.6: Example concepts extracted by ACE [Gho+19] for the bird class. We argue that ACE cannot find abstract concepts, such as gaze direction, because it clusters segmentations on a pixel-based level. We argue that we are currently missing concept-based XAI that goes beyond the grouping of pixels.

in enhancing object detection model performances. It is important to note that we still provide a unique angle: For example, [Saa+19] takes advantage of passively collected gaze information to decrease the number of training examples needed for the effective performance of a learner.

However, we consider the gaze direction of the animal's gaze in the image, not where a person is looking when observing the image. In the literature, gaze annotations for humans often capture the direction of the person's gaze within the image. In contrast, for animals or inanimate objects, annotations tend to reflect points of interest identified by human observers rather than attempt to label the gaze direction of the animal or object itself.

However, we focus on where the animal is looking and is being classified. Various literature supports that emphasizes that using additional features other than the image itself can create more effective learners with fewer training data: [WTC17] validates that gaze annotations can help improve the accuracies of classification approaches. However, both authors focused on annotating what humans look at and deem important. [Kel+19] use gaze annotations to increase the generalizability of their models compared to other benchmark datasets.

Our work incorporates the direction of the gaze that this animal looks at. [WTC17] evaluated their claims on two public datasets and published their own dataset in which food is annotated using important gaze points. The key difference between these ideas and our new contribution is that instead of focusing on saliency and repeating the concept of certain pixels or features that are most important in an image itself, we do not highlight where the person gazing at the image looks, but where the animate being contained in the image is looking. We give an overview of related datasets

and their contributions in Table 4.1 for further context clarification.

Figure 4.7: An overview of example instances from the "bird" class to demonstrate the diversity of the illusions

4.4 METHODOLOGY

One of the contributions of this paper is to show the usefulness of generalizable concepts such as eye coordinates and gaze direction. The direction of the gaze is a problem statement that is often considered in various different settings in the literature [MR15; Liu+19]. In general, it has already been recognized as an important factor in social interactions [Wan+21a] or even object detection itself [Bâc+17]. The goal of this new dataset is to provide an ambiguous dataset baseline that could help evaluate future XAI algorithms on their ability to provide meaningful explanations for ambiguous data.

A key distinction for this proposition of including the gaze, in contrast to existing ones, is that we track where the animal in the picture is looking, not what the person looking at. We consider gaze following as follows: The gaze direction starting from the eye coordinate and then considering the head tilt, the direction of the gaze. The mathematical definition consists of the following components:

- **The Eye Position (e)**: A 2D point representing the position of the eye in the plane, denoted as a vector $[e_x, e_y] \in \mathbb{R}^2$.
- **Head Looking Direction (d):**: A unit vector representing the normalized direction in which the head is oriented, given as $[d_x, d_y] \in \mathbb{R}^2$.

We can then define the gaze direction as $g = e + \alpha \cdot d$, with $\alpha \in \mathbb{R}$ a scaling constant to ensure a set length. For explanation purposes, we aim to make it intuitively long enough to be well visible to the human eye. α is given a length based on the given image size. We normalize the gaze, ensuring a unified vector notation. For the sake of this evaluation methodology, we define that looking straight ahead is annotated as (0.0, 0.0). We always annotated the position of the right eye if two were present.

4.5 EXPERIMENTS: BENCHMARKING THE CONCEPT OF GAZE DIRECTION

Despite having literature supporting our claims that specific additional features help the learning process [Saa+19; WTC17], we wanted to provide an additional experimental evaluation on learning improvements when including the gaze vector. In this work, we focus not on the gaze of the observer (in contrast to eye-tracking literature), but on the depicted gaze direction of the object in the image, e.g., which way the animal is looking.

This distinction is critical because it provides a novel perspective. For evaluation purposes, we downloaded several state-of-the-art pre-trained Imagenet classifiers, namely Resnet18, Resnet34, Resnet52, VGG13 and VGG16 [He+15]. We fine-tuned the networks on learning rates 0.0001, 0.00001, 0.000005 and over various amount of epochs (0-1000). For the direction, we included arrows in the image that annotated the gaze direction in the training set, and validated it, allowing either animal to be classified.

We gather the same results when only one animal is allowed to be correctly classified. Figures 4.10, 4.11 and 4.12 show the accuracies when allowing both classes. Despite allowing both classes as correct, in all cases, including the direction leads to significantly higher accuracies with otherwise the same hyperparameters.

As a baseline check, we tested the same by annotating random other areas in the image, to double check that the concept incorporated was meaningful. We also include one experiment where we tested how eye annotation alone performed in our dataset. Our experiments show that both gaze direction and eye annotation alone lead to meaningful improvements, raising accuracy rates by more than 20% in a setting that allows up to 1000 classes. Our results shown here were generated using the ADAM optimizer [KB17], because the initial experiments revealed to us that this was the optimizer that produced the best accuracy results in practice. ADAM is known to be state-of-the-art in practice [Cho19]. All

experiments were performed using a NVIDIA A100-SXM4-80GB GPU. Calculating lower epochs took seconds, the overall plotting of all accuracies up to 1.000 training epochs took several hours.

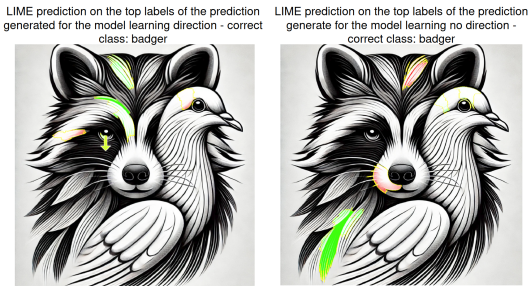


Figure 4.8: We see here that the LIME [RSG16] explanation of the badger trained exactly the same as the other model with the exception of the direction vector is able to include fewer features shared by badger and pigeon, such as more area around the pigeon's eyes and where the fur of pigeon and badger overlap. The model also directly accounts for the eye gaze as well, which shows the usefulness of the feature in the learning process. We also made sure the gaze of the animal varied during the training. So, the animal cannot be classified simply because the arrow always points in the same direction.

Marking the eye outperformed all other tests; however, it should be noted that while the eye is a meaningful concept, this dataset did not necessarily insure uniqueness of looking direction, but always a unique combination of eye coordinate and gaze vector. The interesting part of this experiment was that this eye annotation did not have to be human-level accurate. A marking close to the eye sufficed to increase detection accuracies in ambiguous settings. An overview of those results can be seen in Figures 4.10, 4.11 and 4.12.

We evaluated two different approaches for the improvement of the optimization problem $\arg\min_{g \in G}[\mathcal{L}(f, g, \pi_{x'}, e, d) + \Omega(g)]$: Our first approach was to optimize the task loss and then concatenate the gaze coordinates before evaluating the last softmax layer, then learn the data using a multilayer perceptron [Pop+09], essentially learning the following gaze information loss:

$$\mathcal{L}(f, g, \pi_{x'}, e, d) = \mathcal{L}\left(\mathcal{L}_{task}(f, g, \pi_{x'}) | (e, d)\right) \tag{4.1}$$

This particular setup, however, did not lead to a detectable increase in accuracy. We managed to increase the accuracy by including the gaze vector directly into the image information. However, we tested the accuracy improvement over several different types of architectures, using different amounts of epochs and learning rates,

showing that the learning process can be significantly improved using the concepts eye and gaze direction. As can be seen in Figure 4.10, both eye coordinates and gaze direction led to significant improvements over the baseline of no annotation and random annotations.

We preprocessed our dataset such that the shape of the images was adjusted to $3 \times 224 \times 224$ and normalized using the mean. Furthermore, we took a look at what exactly our network learns if we directly incorporate the direction arrow as well using LIME [RSG16]: As you can see in Figure 4.8, the network then highlights normal features but also the gaze vector as a feature, compared to worse features that overlap with the rival class and no highlighting of the eyes. We included another example of this as an interesting additional observation for PipNet [Nau+23] in Figure 4.9.

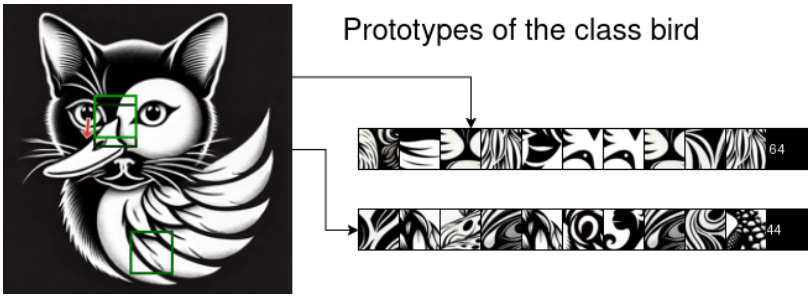


Figure 4.9: This image shows us the top prototypes marked by PipNet for the class "bird". Again, PipNet marks areas in the image that are important to the classification process. We can see that one of the boxes could either be the beak or the eye with the arrow, and one is focused on the feathers/wing patterns. Both are good features for the recognition of the animal. We also show some of the example patterns that PipNet gives us for the feather pattern by which it classifies as bird. Even though this works well, this is yet another XAI method that focuses on areas instead of abstract concepts like viewing direction.

4.6 AMBIVISION: ANIMAL OPTICAL ILLUSIONS- OUR DATASET

In order to underline our problem statement, we include an entirely new self-generated dataset that was constructed using iterative prompt engineering with ChatGPT-4 and ChatGPT-4o [Ope24] of animal-based optical illusions. Creating meaningful optical illusions is inherently difficult. First, note that we include psychological principles for good prompt engineering, such as Gestalt theory [Kof13], which we elaborate together with an example prompt in our discussion of bias mitigation strategies in 4.7. Each image in

Ambivision depicts one animal hidden inside the body of another animal, creating an intentionally ambiguous perceptual boundary.

The dataset consists of over 200 images annotated with the class label, eye coordinates, gaze vector, and bounding boxes for both animals—providing rich, concept-level labels beyond standard object annotations. For every successful image generation, we had to enter approximately 150 prompts, resulting in roughly 30.000 ChatGPT prompts [Ope24]. Of the resulting dataset, 41 images are RGB, while the majority are black and white. This design choice reflects the increased difficulty in generating convincing color illusions. Ambivision is, to the best of our knowledge, the first dataset of its kind to systematically encode perceptual ambiguity with fine-grained concept annotations, making it a valuable baseline for evaluating both classification performance as well as explainability in ambiguous visual settings.

We provide four different versions of the dataset for the convenience of the user: We provide the dataset with the label of the animals, followed by the eye coordinates (e_x, e_y), the normalized direction vector (d_x, d_y) and the bounding boxes (x_1, y_1), (x_2, y_2). We provide the same dataset with direction arrows drawn directly into the image as a baseline for this work. An additional version is the dataset with random markings in the image and one baseline where just the eye is encircled.

4.7 BIAS MITIGATION STRATEGIES

Because this dataset was generated with ChatGPT [Ope24], we had to make sure to achieve diverse results and mitigate biases. For each animal class, we ensure that the gaze and eye position vary. This applies equally to both animals in the image to avoid introducing unintentional class-specific cues. We include one exemplary overview of all eye positions in Figure 4.13 and the spread of the eye positions for the bird class as plots in Figure 4.14.

We varied artistic representation by prompting specifically for more or less realistic art styles. We specifically prompted the animals to be in alternating positions and poses, for example moving, sitting, flying, and eating. Another principle we applied for prompting is a principle borrowed from the psychological domain.

The design rules for optical illusions go back to fundamental problems of psychology, such as Gestalt theory [Kof13]. These laws of conceptual organization give us an easy overview of what design

concepts can trick our minds or make it difficult for our brains to correctly organize and interpret visual data.

One such example is the concept of proximity: When something is in close proximity to something else, we are more likely to interpret it as belonging together than when they are further apart. The Gestalt principles provide us with useful guidelines for our prompts. Establishing this baseline dataset allows us to explore perception differences between AI learners and humans. One example of an initial bias that we mitigated was that we had immense trouble making sure illusions from the owl class did not always look straightforward. This might be due to the fact that owls can turn their neck 270 degrees in real life as well and have quick reaction times, which means that owls will, in fact, face you more often than not. However, by asking for a flying owl, for example, and specifically asking for different poses, it was possible to get more diverse illusions.

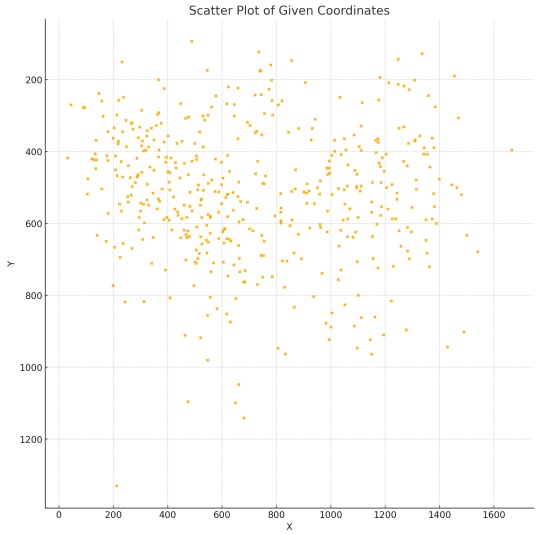


Figure 4.13: This scatter plot shows the spread of all eye coordinates from all classes. It is pretty diverse, considering the eyes are placed within the body of the animal and thus do not appear at the edges.

PROMPT EXAMPLE: An example of a written input is "Generate an artistic black-and-white image featuring a fusion of a tiger and a falcon.". ChatGPT itself then extended that prompt and output the final version of the prompt that it internally used. The final result as an example was:

Generate an artistic black-and-white image featuring a fusion of a tiger and a falcon. The design blends the tiger's powerful stripes and muscular build with the falcon's sharp beak and impressive wingspan. This creates

an optical illusion where, from one angle, it appears as a tiger crouching to pounce and, from another, as a falcon swooping down to capture its prey. The image emphasizes the shared features of predatory prowess and agility, highlighting the ferocity and grace of both animals.

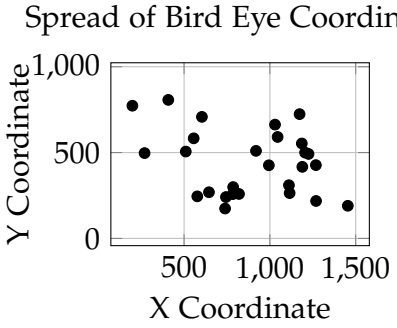


Figure 4.14: Scatter plot showing the distribution of bird eye coordinates as an example class that shows diversity of eye positions.

4.8 LIMITATIONS AND FUTURE WORK

First, one major limitation is that the dataset is generated using ChatGPT, which is itself vulnerable to internal biases. However, this is why we explicitly focused on addressing potential biases in our Section 4.7. Furthermore, it is very hard to generate any kind of optical illusion, and our approach here allowed us to generate a dataset of more than 200 images in a feasible manner. We also note that it is very interesting to have a dataset that is an optical illusion dataset for humans, but generated by a machine, and that large optical illusion datasets are scarce in general [Sha+24].

Furthermore, in this dataset, we explicitly focus on two animals distinguishable by their gaze vector and eye coordinates. Although this was outside the scope of evaluation, during our various and extensive prompt attempts, we generated other interesting concepts. Such as: examples of animals where more than one animal is hidden in the body of an animal, but also humans and animals in a mixed manner. We give access to the images that did not meet our criteria for evaluation purposes for exploratory and research purposes. Examples can be found in Figures 4.15 and 4.16.

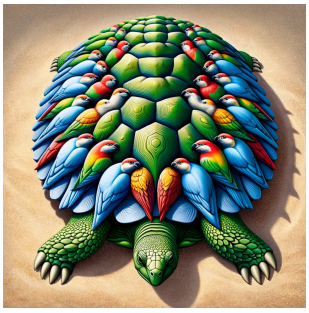


Figure 4.15: This shows us how easily optical illusions can be extended to more than just two animals within each other. We include several of these images generated accidentally when trying to generate as many different optical illusions as possible.

We must also consider what happens when the gaze is not the distinguishing feature, as in Figure 4.16. Overall, this paper aims to broaden our perspective: What if highlighting pixels was the wrong approach for explainable AI in the image domain? What concepts should we really learn? Can we tackle more comprehensive learning strategies with the use of optical illusions and knowledge from psychological domains? These extra images will be included in the open-source dataset in a separate folder.

Figure 4.16: This is an example of a generated image where the gazes are completely shared and do not help in distinguishing which is the correct feature. Is there a more prominent answer on which one is seen with a higher likelihood, and what is it dependent on? The outer one? Do we prefer the color black? Is it the animal whose head "looks complete"? Does this change when we turn the angle of the picture, so another kind of "gaze direction"?

We were able to generate pictures where neither gaze nor eye coordinates was a distinguishing feature, but there were still two animals visible, as in Figure 4.16. Ultimately, this work invites a broader perspective: What if pixel-wise saliency is not the most effective approach to explainability in the image domain? What kinds of concepts should models truly be learning? Can we pursue more holistic learning strategies inspired by optical illusions

and insights from cognitive psychology? These questions form a foundation for future exploration.

4.9 CONCLUSION

In this paper, we challenge the prevailing paradigm in explainable AI (XAI) for visual data, which primarily revolves around pixel-based attributions and saliency maps. Although such methods offer useful insights in many domains, they fall short when confronted with perceptual ambiguity——situations in which even human observers struggle to resolve competing interpretations.

Inspired by classical optical illusions like the rabbit-duck example, we propose that meaningful explanations in these cases must go beyond pixels and capture abstract, semantic concepts such as gaze direction and eye position. While there have been some efforts in the concept-based domain, automatic generation of concepts again relies on pixel-based methodology and fails to capture concepts such as the viewing direction.

To address this research gap, this paper introduces a novel dataset, Ambivision, which presents visually merged animal optical illusions. Each image is annotated with the animal classes, their right eye coordinate, the normalized viewing direction and bounding boxes for the animal. Through extensive experimentation across multiple state-of-the-art architectures and training regimes, we demonstrate that including such concept-level annotations, specifically gaze and eye location, leads to significant performance improvements on classification tasks in ambiguous settings.

Additionally, we show the limitations of popular existing XAI algorithms on this particular dataset due to its ambiguous nature. Ambivision represents a step toward rethinking how we evaluate and design explainability in AI. It opens up new directions for building more human-aligned explanations, ones that take into account not just what is seen, but how it is perceived.

Table 4.1: An exemplary Overview of existing Types of Gaze Annotation Datasets and their purposes.

Dataset Name	***Type of Data***	**Type of Annotation**
Cathegory: Human Gaze Annotations		
1. MPIIGaze [Zha+15]	large-scale dataset of Human Images of 15 subjects	Focus on different lightings, variable places and day times, as natural settings as possible
2. EyeDiap [FMO14]	human gaze estimation dataset of 16 test subjects changes in ambient and sensing conditions	head pose variations, gaze poses ground truth given by the 3d poses of the visual target
3. ETH-XGaze [Zha+20]	human gaze estimation dataset of 110 participants	head pose variations and different lighting
4. Gaze360 [Kel+19]	human gaze estimation dataset of 238 subjects	wide range of lighting conditions
Cathegory: Animal Gaze Annotations		
5. Animal Kingdom Dataset [Ng+22]	Video Dataset of Animals	annotated for relevant animal behavior, no eye gaze annotation but pose estimation
6. PET [Gil+15]	classes bird, cat, cow, dog, horse, and sheep tracking eye position for visual points of importance from 40 users	eye movements recorded of human points of interest for free vision task and visual search
7. AnimalWeb [Kha+20]	animal faces collected from 350 species	annotated with 9 land-marks on key facial features
Our Dataset: Ambivision Optical Illusions)	Drawn Animal Images generated by ChatGPT4 and ChatGPT4o consisting of always two animals, one merged/hidden in the other animal	Gaze vector in 2D format, animal labels for both animals animals were distinguishable by their right eye coordinate and normalized direction vector. If it stared straight ahead, gaze vector was assigned as 0.0, 0.0. annotated bounding boxes

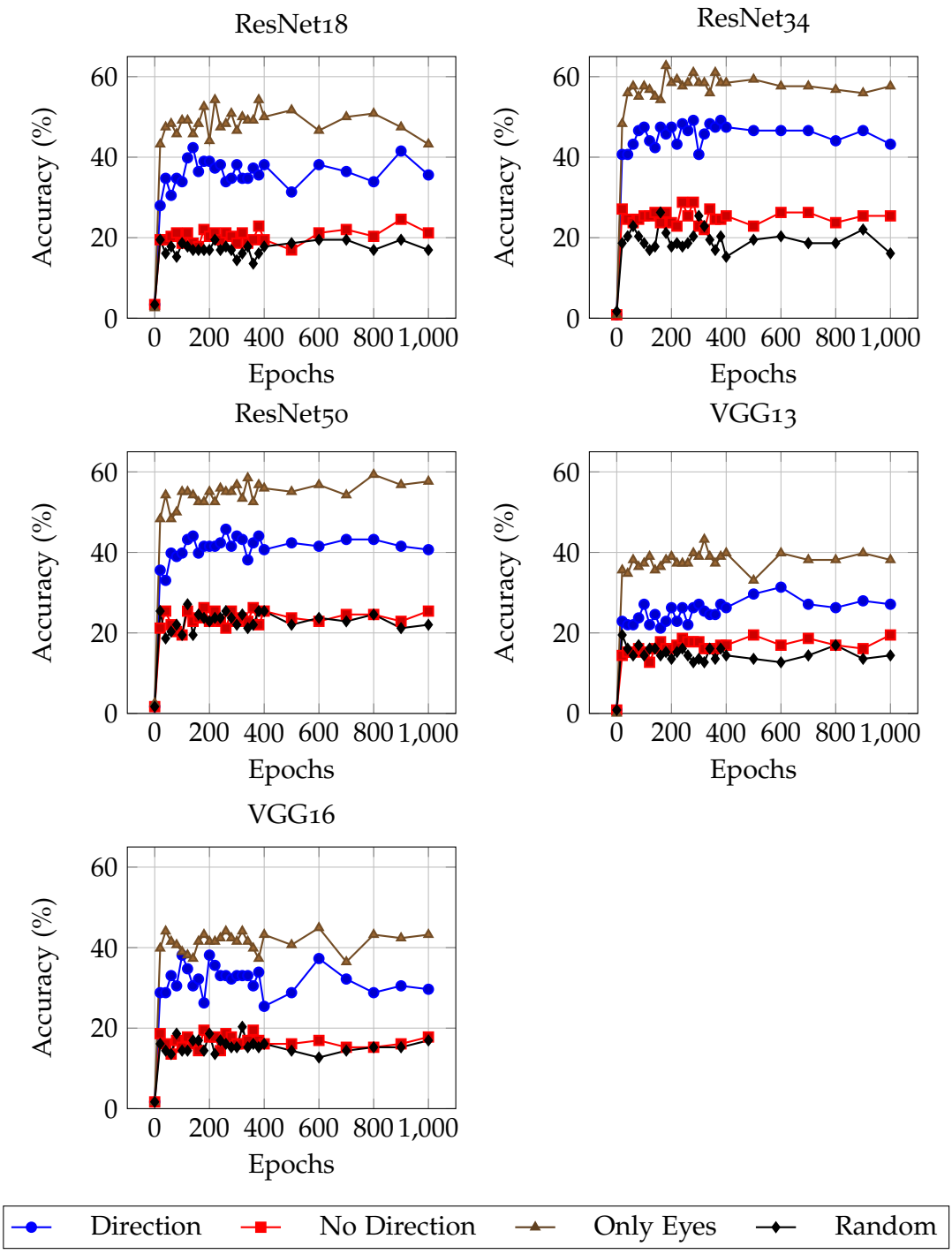


Figure 4.10: Accuracy vs. Epochs at LR=0.0001 for ResNet and VGG models across different annotation conditions.

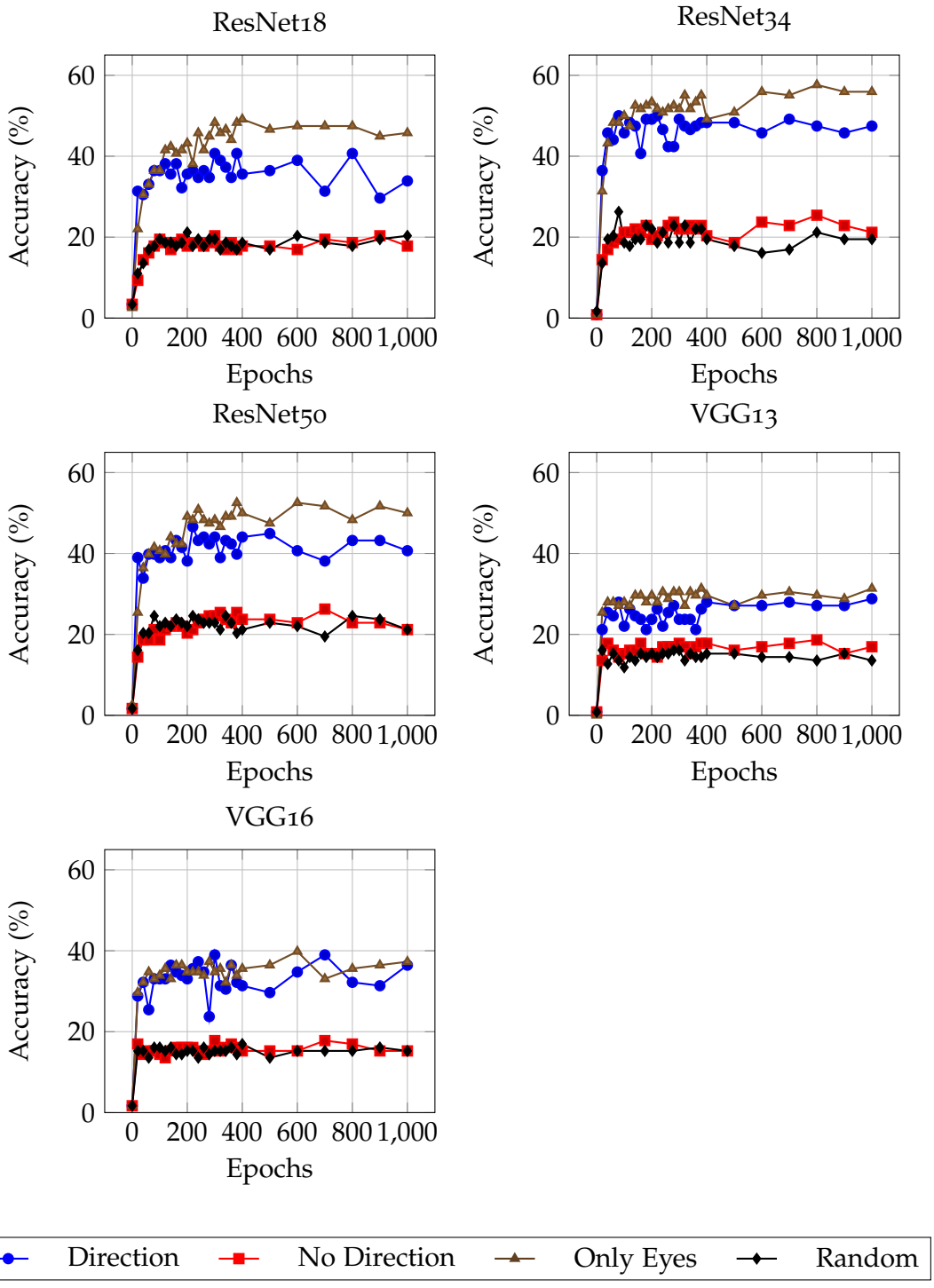


Figure 4.11: Accuracy vs. Epochs at learning rate 1×10^{-5} for ResNet and VGG models under various annotation strategies.

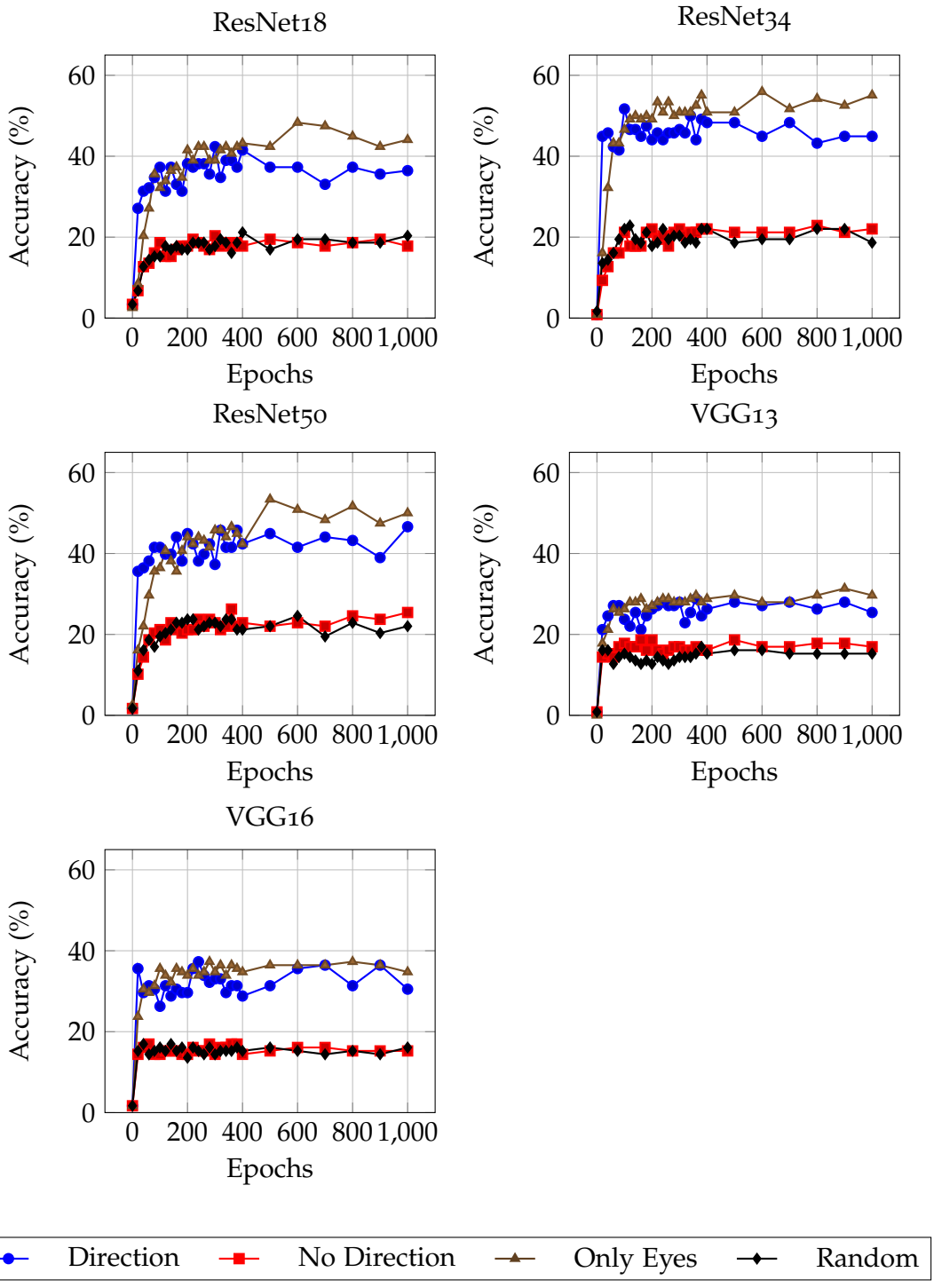


Figure 4.12: Accuracy vs. Epochs at learning rate 5×10^{-6} for ResNet and VGG models under various annotation strategies.

Part II

BRIDGING THE GAP BETWEEN UNCERTAINTIES AND ROBUSTNESS

5

INTRODUCTION AND RELATED WORK

An in intriguing phenomenon in machine learning research is the vulnerability of models to calculated attacks. Machine learners have been perfected originally according to their performance on unseen data without taking into consideration that the data to be classified could be actively manipulated by an adversary in order to cause misinterpretations. In real-world scenarios, though, if we apply neural networks to more important classification or observation tasks such as airport security, it is essential to account for active adversaries.

In the past, this has led to shocking results: When a sophisticated machine learning algorithm can be fooled into classifying a turtle as a rifle, this raises the question of whether robust classification with machine learning is possible at all and has lead to a new field of research, namely robustness [Ath+18]. The first formal definition of adversarial examples as an attack method was provided by Szegedy et al. in 2013 [Sze+13].

Definition 5.0.1 (Formal definition of Adversarial Examples for Images)**.** *Let $f : \mathbb{R}^n \to \{1, \dots, K\}$ be a trained classifier mapping the image value vectors to a discrete label set. We assume f has a continuous loss function f_{loss}. For a given $x \in \mathbb{R}^n$ and label $l \in \{1, \dots, K\}$, an adversarial example is a solution of the optimization problem [Sze+13]: Minimize $\|r\|_2$*

$$\begin{aligned} \textit{Subject to} \quad & f(x+r) = l \\ & x + r \in [0,1] \\ \textit{With} \quad & f(x+r) \neq f(x) \quad \textit{(misclassification)} \end{aligned}$$

Why do we care about adversarial attacks and the capabilities of a potential attacker? Especially in domains where the application is critical to safety, such as autonomous driving or health care applications, attacks can have grave real-life consequences that could harm human lives [Fin+19]. The problem of adversarial robustness has, to this day, not been satisfactorily set aside and solved.

Open problem statements include the transferability of adversarial attacks, which we will address in Chapter 6 and detection and mitigation strategies, which we will discuss in Chapter 7. We will now briefly give an introduction to adversarial attacks before ad-

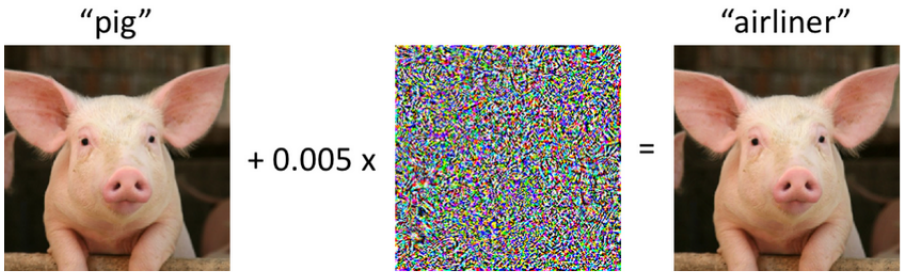


Figure 5.1: This figure shows an adversarial example [MS18]. Normally the classifier would classify the image as a pig. By adding a specific noise pattern to the data of the image the classifier afterwards misclassifies it as an airliner with high confidence even though the image has not changed for the human eye.

dressing why transferability and adversarial detection are essential milestones to solve for future machine learning research in Sections 5.1 and 5.2.

CAPABILITIES OF THE ATTACKER Adversaries are distinguished by their different attack capabilities and by how far we allow them access to our model and the training data. Taxonomies distinguish training data control, model control, testing data control, label control, source code control, and query access [OV23].

For each of these possibilities, the attacker gets full—or at least-partial—access. As an example, training data control means that the attacker is allowed to take control of a subset of the data by modifying the training samples. Query access describes that the attacker is allowed to receive predictions for a limited amount of interactions.

The primary distinction in adversarial taxonomy is between white-box and black-box attacks [Tab+19]. White-box attacks assume that the attacker has full knowledge of the model, the model parameters and the training data. This setting represents a worst-case scenario and is commonly used to evaluate a system’s robustness under the assumption of maximum adversarial knowledge. Black-box attacks only grant the attacker limited interaction in the form of access queries. The adversary has no knowledge of the structure or parameters of the network.

Furthermore, the adversary also does not have access to any large training set but is only provided the labels for classes specifically requested by a classifier oracle [Pap+17]. In real-world scenarios such as automatic spam detection, defeating a machine learning algorithm can reap ample benefits, leading to a high motivation to defeat those algorithms. On the other hand, it is unrealistic that

the adversary trying to pass spam email through filters knows exactly how the filter operates and what features it looks out for [LM05]. This is why research also explores the possible ways an adversary can actively learn enough about a classifier in order to still effectively generate adversarial examples through queries.

Endless combinations exist for any in-between capability scenarios, which then fall into the category of gray-box attacks. In order to measure the success of an adversary, we will have to take into account the cost of creating a successful misinterpretation.

Ultimately, the goal of an adversary will be to find a feature-changing strategy that will maximize its own expected utility. Applied to the most common use case for adversarial examples, namely misclassification attacks, we define the following problem: Let us define a classifier with model parameters θ that minimizes an empirical loss function $L(x_i, \theta)$ for a given set of samples $x_1, x_2, ..., x_n$ [Dal+04]. This means that for a fixed model θ and input x, the adversary aims to find δ such that

$$\max_{\|\delta\|_p \leq \epsilon} \mathcal{L}(f(x + \delta; \theta), y)$$

δ is constrained because we do not want to allow the perturbation to be too large so that it remains as inconspicuous as possible. A large amount of the research on adversarial examples focuses on active misclassifications of images. This is not the only area of research where the concept of adversarial examples is applicable. In speech recognition, hidden and inaudible voice commands can be integrated [CW18].

Furthermore, the specific loss functions vary with the problems faced. When applying adversarial examples to other domains, such as comprehension systems, the meaning behind the attack becomes quite a different one. While adversarial pictures punish model oversensitivity to imperceptible noise, adversarial attacks on comprehension systems punish overstability [JL17]. Otherwise, the main concepts and training methods remain the same. However, in this thesis, we will mostly focus on image attacks in the classification context.

Intuitively, small alterations of the input that remain unrecognizable are better than large changes. This is why the goal of the adversary is to find adversarial examples that require minimal alteration of an input x that lead to a wrong classification output. Therefore, we need to model the costs required in order to generate an adversarial example by proposing adversarial cost functions.

ADVERSARIAL COST FUNCTIONS The idea behind a cost function is to weigh the likelihood of detection by altering the input too much versus the effectiveness of the change and how likely it is to fool the model. This follows the same intuition as a loss function, only for adversarial detection likelihood and attack potency. This is why the cost function has to be specifically fitted to the task at hand [CB16]. An intuitive example for an adversarial cost function would be to define a linear cost function $J(x)$ as the weighted absolute difference between the feature values in the base instance x^a, and those in the target instance x [LM05]

$$J(x) = \sum_i a_i |x_i - x_i^a|$$

For example, spam email x^a generates the most sales, but we have to account for the costs of changing the original email. While we now have the means to evaluate the effectiveness and practicability of the adversarial example, we still need to address how to generate them when no prior knowledge of the attacked learner is given.

We have now given an introduction to adversarial examples and define two main threat models for the attacker, namely black box and white box attacks. In the following, we will introduce interesting properties of adversarial examples, such as transferability, that we focused on in our research.

For the understanding of Chapter 6, and later our own construction of an attack method in Chapter 7, we briefly introduce FGSM to give a comprehensive intuition of attack construction methods [GSS14].

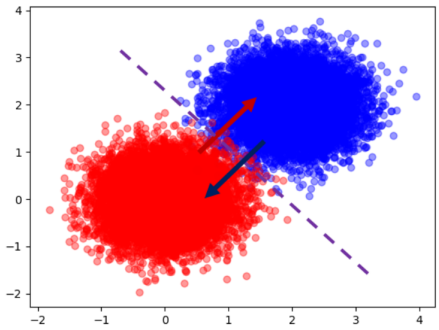


Figure 5.2: In this Figure, the Fast Gradient Sign method is visualized [Tsu18]. Noise is added such that the input moves closer to the regions where the classification decisions intersect.

Definition 5.0.2 (FAST GRADIENT SIGN METHOD (FGSM)). *Let θ be the parameters of a model, x the input, y the output, and $J(\theta, x, y)$ the cost function used to train the neural network. Then, Goodfellow et al. propose a method which constructs an adversarial example $\vec{x}^* = \vec{x} + \delta_{\vec{x}}$ with means of the optimal max-norm constrained perturbation [GSS14]*

$$\delta_{\vec{x}} = \epsilon \cdot \mathsf{sign}(\Delta_x J(\theta, x, y))$$

The gradient in this case can be efficiently computed using back-propagation, which is a common technique used for the computation of the gradient of a function. The factor ϵ can be interpreted as the perturbation's amplitude, measuring the detectabilty indirectly by limiting the amount of change that the adversary is allowed to apply [Pap+17]. The visualization in Figure 5.2 shows how adversarial attacks are designed to move samples in the direction of the decision boundary by a certain step size.

In the following Chapters, we will first bridge the gap by attacking certainty, and we will then expand the discussion to interesting phenomena such as adversarial transferability. To be more precise, we will now describe the contributions of the following Chapters and the papers which they are based on below:

- Chapter 6, "On the Independence of Adversarial Transferability to Topological Changes in the Dataset" is based on the publication [NM23],
- Chapter 7, "Certainty attacks using explainability preprocessing", is based on the collaborative publication [NPM25].

5.1 TRANSFERABILITY OF ADVERSARIAL PERTURBATIONS

Astonishingly, adversarial perturbations are not limited to working extremely well on targeted classifiers [Sze+13], but demonstrate an ability to transfer not only to other models trained on different subsets of training data, but also to different architectures [PMG16]. These findings imply that modern machine learning models often rely on incorrect or brittle internal representations learned from the data [GSS14]. There have been several attempts to explain these properties [Ma+18; Xie+19; Liu+16; PMG16], as well as research on the construction of universal adversarial perturbations [Moo+17]. One such explanation for assignments of the same class to adversarial examples applied to different classifiers is that neural networks trained with current methodologies all resemble one another.

More explicitly, their learning on different subsets still generate approximately the same weights and result in the stability of the adversarial example class.

However, this does not entirely explain this phenomenon. Another argument often discussed is whether the linearity and simplicity of models leads to adversarial examples occurring in broad subspaces, leading to higher transferability for less complex models.

This hypothesis can be supported by the observation that adversarial examples generated in deeper models exhibit lower transferability [Xie+19]. The transferability domain is often categorized into optimization-based and generation-based methods [Gu+23]. Optimization-based methods describe methods where one optimizes by introducing surrogate models, while generation-based methods introduce generative models.

Following [Gu+23], we define adversarial transferability as: Given an adversarial example x^{adv} of the input image x with the label y and two models $f_s(\cdot)$ and $f_t(\cdot)$, adversarial transferability denotes:

$$\arg\max_i f_t^i(x^{adv}) \neq y, \text{given } \arg\max_i f_s^i(x^{adv}) \neq y$$

There exist a lot of attempts of understanding this observation; however, they all come with their limitations: For example, for Ma et al.'s approach using local intrinsic dimensionality [Ma+18], it was immediately noted that LID adversarial detection is very dependent on the confidence parameter deployed by an attack [LCY18], and fails to then detect attacks in black-box settings.

This vulnerability underscores a deeper conceptual limitation of LID: it fundamentally captures data manifold properties in the feature space but is agnostic to model-specific characteristics, such as the shape of the decision boundary or gradients.

This leads to an intriguing paradox: How can a model-agnostic measure still perform reasonably well in detecting attacks that fundamentally exploit model-specific gradients? Furthermore, in all research directions, we come across limitations of these methods [Zou+20]. An example is the observation of limited input diversity when applying DI^2 -FGSM due to the application of simple padding and resizing strategies [Xie+19]. In the discussion by Demontis et al. [Dem+19], the authors highlight the intrinsic adversarial vulnerability of the target model and the complexity of the surrogate model as influencing factors.

One of the metrics they introduce and highlight is the cosine angle between the target and surrogate gradients to characterize adversarials, which has also been empirically confirmed by [Liu+16]. But, as all proposed proxies so far, they come with clear drawbacks: Their proposed method raises concerns for larger approximation

errors of the cosine similarity and only exploits a rigid optimization trade-off, as it requires tuning between maximizing gradient cosine similarity and minimizing loss with a fixed ratio.

In summary, all known methods so far show clear limitations and leave the universality and transferability of adversarial examples open to discussion. This makes the topic so full of potential: When fully understood, this could lead to breakthroughs for robust learning algorithms and a better understanding of neural networks in general. Another subarea of adversarial attacks relevant to our proposed work is adversarial detection as a mitigation attempt.

5.2 ADVERSARIAL DETECTION

Many research projects have tried to detect adversarial examples in the testing stage [LIF17; Met+17; Fei+17; Gro+17; Lin+17]. We will briefly focus on two of the approaches [Met+17]. Metzen et al. create a detector as an auxiliary of the original network. The task of this network is to predict the likelihood that the input is adversarial. In order to accomplish this, it is simply being fed images and their adversarial counterparts with the correct labels.

On the other hand, Grosse et al. define an outlier class which they include in their original deep learning method [Gro+17]. This broadly relates to our idea and the paper presented in Chapter 3, where we use local intrinsic dimensionality as a proxy for the likelihood of attack [Ma+18].

Using two statistical distance measures, the maximum mean discrepancy and the energy distance, adversarial examples can be identified with high confidence. Although this is a useful method, there is no guarantee that these detection algorithms will be able to identify adversarial examples generated by all attack methods. Furthermore, new attack methods could prove resistant to this countermeasure. In the past, there have already been successful circumventions of detection algorithms [CW17a].

Other detection approaches include detection attempts using confidence scores and uncertainty of networks [SG18; RWN17; Fei+17], which was the starting point of our research in Chapter 7, where we explicitly manipulate the confidence output of a network.

To summarize, in this thesis we first discuss an attack method targeting the certainty and transferability, as well as pertubation size, and then focus on the cause of transferability with a new

proxy attempt. The following Chapters are based on published work as cited below:

- Chapter 6, which is based on the peer-reviewed publication "On the independence of Adversarial Transferability to Topological Changes in the Dataset" [NM23]
- Chapter 7, based on the peer-reviewed publication "Certainty attacks using Explainability Preprocessing "[NPM25].

Part III

TRUST AND UNCERTAINTY QUANTIFICATION

8

INTRODUCTION AND RELATED WORK

Uncertainty Quantification and trust calibration go hand in hand [CBH23]. What use are machine learners to us if humans do not employ them for fear of trusting them? Trust becomes the gatekeeper of utility, especially in high-risk domains. Therefore, we discuss what makes an AI system trustworthy and how we achieve warranted trust [NMN25]. We address the need for a minimal unified characterization of trust for AI systems. Although there exist a lot of attempts to describe trust in AI, [Jac+21; SW18; Loc+21], we address current research gaps that exist in the unification of knowledge in the psychological domain with computer science, especially in the uncertainty quantification domain. At the same time, uncertainty quantification alone is not always sufficient to foster trust.

The following chapters show the interconnected relations between explanations, trust, uncertainty quantification, and other key factors that are essential for future safe deployment of AI models. It is essential that AI is not only employed and used, but also remains reliable and as impartial as possible [Rya20].

Moreover, the growing deployment of AI systems in sensitive sectors necessitates not just functional trust (i.e., that the system performs as intended) but institutional trust rooted in standards, oversight, and assurance. This motivates the emerging interest in AI seals [Wis+24a] or certification schemes that validate key trustworthiness properties such as robustness, fairness, uncertainty awareness, and explainability. Our contributions are summarized below.

1. In Chapter 9, we define a framework on how to assign trustworthiness to a machine learner and quantify it with a score. This work underlines all our interdisciplinary efforts in the following Chapters. It is based on [NMN25].

2. In Chapter 10, we perform a human-centered evaluation of interpretability by evaluating the actual trust calibration abilities and explainability of our proposed method Unsupervised DeepView [NM22a]. We derive some useful general guidelines for the design of explainable AI methods. It is based on the work [New+25a]

3. In Chapter 11, we investigate AI seals and their effect on perceived trustworthiness of a system. It is based on the peer-reviewed publication [Wis+24a].

9

TRUST AND UNCERTAINTIES: CHARACTERIZING TRUSTWORTHY AI SYSTEMS WITHIN A MULTIDIMENSIONAL THEORY OF TRUST

Trust is a basic feeling and attitude that shapes human relations and is the glue that holds groups and societies together. With AI increasingly present in our daily lives, the key question becomes how much we can trust these systems. This question can be discussed from a psychological person's perspective ("under which conditions are we inclined to trust AI systems?") or from an objective system's perspective ("under which conditions is a system worthy to be trusted and to which degree?").

In this work, we adopt a system-level perspective and thus abstract from subjective psychological conditions. Furthermore, our aim is to offer a general framework for a comparison of different systems, which is especially innovative by integrating two dimensions in the comparative framework, namely uncertainty (as a recent hot topic in AI) and commitment (rather new for AI systems). As a result, we make transparent in which dimensions some AI systems like ChatGPT or autonomous cars vary from rather established trustworthy systems like classical cars and (prototypical) democratic institutions.

This overview can be used both to understand specific features of AI systems and to disclose deficits of their trustworthiness that we need to overcome to make AI systems acceptable. Although the central question of whether and to what degree we can trust AI systems has already been intensively discussed, we are lacking a general framework for a systematic comparison of trustworthiness. Therefore, we propose a multidimensional framework of trustworthiness that is developed in three central steps.

First, we suggest six central dimensions of trust from a general perspective motivated by important prior articles: objective functionality, transparency, uncertainty quantification, embodiment, immediacy behaviors, and commitment. Second, we develop a more detailed perspective by unfolding several specific features for each dimension, partially realizing the dimension. In the third step, we exemplarily describe an evaluation for each feature of one dimension (as being implemented low, medium, or high) of a system such that we can calculate an average value for each dimension.

As a result, we receive our multidimensional trust account that enables us to compare the trustworthiness of different systems as a basis for future development of AI systems.

9.1 INTRODUCTION

Why is trust relevant? Trust plays a crucial role in our personal relations, and distrusting someone radically changes the social interaction. Furthermore, trust is also a crucial factor for societies: it is shown that "high-trust" societies have stronger economies and stronger social networks in general than "low-trust" societies [Fuk96; Ing99].

Since AI systems are about to interfere with our personal relations and are about to find their way into all types of work processes, we are faced with the question of whether we can trust AI systems. This paper aims to develop a framework to describe and analyze the trustworthiness of AI systems by comparing them with the trustworthiness of humans, classical machines, and democratic institutions: we call this a comparative perspective.

Such a systematic comparison should allow us to recognize and understand differences in trustworthiness for different systems, and especially to highlight deficits of existing AI systems. Our starting point is the conceptualization of Blöbaum [Blö21]: He distinguishes the trustor (the person who trusts someone), the trustee, i.e., the person or system someone is trusting, and the relation (or process) of trusting between trustor and trustee. We elaborate on this to constrain the focus of our endeavor.

All three elements of a case of trust can be characterized with relevant properties but can also vary a lot: a relevant aspect of the trustor is, e.g., the personality feature of readiness to assume risk: is the person risk averse, risk seeking, or having an intermediate level of risk propensity? This, of course, changes the tendency to trust another person or system intensely (from now on we use "system" to include persons unless we want to distinguish both explicitly).

The focus of this paper is to work out a description of the most relevant properties of trustworthiness of a trustee by focusing on the properties of the system we are warranted to trust; thus, we set aside subjective psychological features of the trustor like risk propensity as much as possible to reach a description of the most relevant intersubjective properties of trustworthiness of a system by focusing on the properties of this system; this can be understood

such that we presuppose standard psychological conditions for the trustor, e.g., having an intermediate level of sensitivity to risks, without discussing them as a factor in this paper.

We call this the system perspective of trustworthiness and will describe more details below. Concerning the relation of trust between the trustor and the trustee, we will elaborate that we are focusing on explicit and warranted trust when working out the properties of trustworthiness of a system. Thus, we leave aside the philosophical debate whether the trust relation is always a belief or some other category of mental state. For an attitude in a general sense, we can reasonably allow for the distinctions of implicit and explicit trust [MZ06].

9.2 OVERVIEW OF THE INNOVATIVE ASPECTS

We give in the following an overview of the innovative aspects of the multidimensional theory of trustworthiness. Although there has been a lot of research on the topic of trust in general, with a focus on humans and trustworthy AI in detail, we develop a general framework that enables us to systematically compare rather different candidates to be trusted, including humans, democratic institutions, and AI systems. The trustworthiness of a system is not a question of being either present or absent, and our comparison offers a justified gradual evaluation of the trustworthiness of the systems according to relevant dimensions.

Since we want to propose an adequate framework for such a general comparison, we want to highlight two dimensions to account for special aspects of AI systems, namely, uncertainty quantification and commitment. Why is uncertainty and uncertainty quantification so relevant for AI systems? It is a dimension that has long been underrepresented in the literature.

Recently, there has been an increase in discussion for specific AI systems, as we will describe in detail below (Table 9.2 at the end of the Chapter). Uncertainties are especially relevant when biases occur, as an intuitive example illustrates. When a dataset used for training itself is biased, for example due to gender and ethnic stereotypes, the AI reflects these stereotypes [Gar+18]. This makes uncertainty quantification a vital tool for identifying and mitigating these risks.

However, how can we tell exactly that this occurs and what type of uncertainty are we discussing? How can we even be sure that

uncertainty exists? Without uncertainty quantification and these biases being revealed, the person might just think the functionality is flawed or might not even notice that this bias exists in extreme cases. These types of occurrence are also well known in other areas, for example, recruitment [Nug+20], where bias in the data itself leads to potentially unethical/ unfair recruitment choices.

This example illustrates that uncertainty quantification is an especially important dimension concerning AI systems: Let us have a closer look, aiming for a working definition of this dimension of trust: Uncertainty quantification aims to measure all sources of error and uncertainty, including systematic and stochastic measurement error [Soi17]. It captures numerically the limitations of a model, and is commonly separated into aleatoric and epistemic uncertainty [Sul15].

Epistemic uncertainty describes uncertainty arising from a lack of knowledge and can therefore be improved by more training data. Aleatoric uncertainty refers to the uncertainty of an inherently random phenomenon such as throwing a die. Although uncertainty quantification is often discussed in AI, it is only more recently considered in discussions of trust and is lacking in general multidimensional accounts of trust.

To demonstrate how uncertainty quantification has been discussed so far, we provide a list of the top 45 most influential references in Table 9.2, where we list the paper, a brief description of their perspective on trust and a description of how they include uncertainty quantification as a component of trust calibration and to what extent. We show that most of this highly cited literature concerning "trust" in combination with "artificial intelligence" or "machine learning" and "uncertainty quantification" (according to Google Scholar) uses the term uncertainty to ambiguously describe risk without being aware of statistical or computer science methods that quantify uncertainty and its benefits (see Table 9.2).

This selection also includes some proposals of general frameworks of trust in AI systems, which we marked in bold: most importantly, none of those general accounts includes uncertainty or uncertainty quantification as a factor in the list of relevant aspects, but at best mentions the general aspect of risk without spelling it out. This gap should be overcome with our proposal. Furthermore, our literature review in Table 9.2 proves a systematically increasing sensitivity for the relevance of uncertainty quantification for the evaluation of trust in AI systems. This also supports the relevance of uncertainty quantification in a general concept of trust.

DESCRIPTION OF TABLE 9.2: We briefly describe our method with which we performed a brief literature review, which you can find in Table 9.2. We present an overview of the most influential articles based on two searches and describe them with their core content and the role of uncertainty quantification:

To demonstrate the importance of uncertainty quantification as a relevant aspect of trust in the existing literature, we selected the 45 most important papers on trust in AI systems and the role of uncertainty with the following criteria: We used google scholar and first searched for the keywords "trust" in combination with "artificial intelligence, AI", "machine learning, ML".

For the resulting articles we systematically checked whether the keyword 'uncertainty' was used at all (if not, there is a cross in the third column uncertainty dimension). Then we did a second search and added directly the keywords "LLM" and "uncertainty quantification" to the above list to especially find more relevant articles discussing this dimension. The result is 45 selected articles, which include the most relevant articles according to Google Scholar, connecting both search processes. We listed and discussed these articles describing the key content (2nd column). If uncertainty quantification was used in a paper (not guaranteed in the first search), we carefully read it and checked its role in the discussion of the article which is then shortly described (3rd column).

As a result, the list proves the important and increasing role of uncertainty quantification in specialized discussions of trust in AI systems, but it also demonstrates that uncertainty (quantification) is not yet integrated in any general framework of trust (literature in bold below).

To continue with our definition of trust: Overall, in the context of warranted trust calibration, uncertainty quantification should be a key milestone to account for the specific aspects of AI systems. Thus, we include the concept with the following working definition:

Definition 9.2.1 (**Uncertainty Quantification**). *Uncertainty Quantification (UQ) describes the scientific process of mathematically characterizing biases, risks, errors and other sources of error and is commonly split into aleatoric and epistemic uncertainties [Soi17]. Aleatoric uncertainties describe uncertainties arising from inherently variable phenomena, due to inherent randomness or measurement noise concerning the phenomena, and cannot be reduced, whereas epistemic uncertainties describe uncertainties occurring due to a lack of knowledge either of data or of the relevant model [Sul15]: It can be reduced with more information or*

better models [Cha+25]. We add uncertainties which result from external influences, such as adversarial attacks, in the case of AI systems.

Let us now have a look at the second innovative dimension, namely commitment: All theories of commitment are developed with the presupposition that only humans can be committed to doing or enabling something. And commitment is then described as a feeling or attitude of a human being to do or enable something. The relation of commitment between a human and the entity a human is committed to is described rather differently concerning the latter: Some analyze commitment to do (or enable) something as a relation of a human to either (a) a line of activity, (b) particular role partners, or (c) an organization (see overview by [BR91]).

One prominent theory, the identity theory of commitment [BR91] argues that the relevant relation is one of identity in the sense that all commitments of a human being are anchored in an identification of the person with certain self-defining beliefs ("a stable set of self-meanings"): a person is committed to what she thinks is constitutive for being the person she is. Independent of these varieties of relations, they are all held by human beings. This position is also held by Hawley [Haw14] who argues that "to trust someone is to rely upon that person to fulfill a commitment, . . . " (p. 1) and connects her view with the claim that being committed is only possible for humans.

We think that this perspective is too narrow since institutions and complex systems can have commitments too, e.g. democratic institutions are committed to guaranteeing free press and free elections. The commitment is anchored in the structural and organizational features of a (prototypical) democratic institution.

If our democratic rights are not secured, we sue the democratic state, not individuals. We propose to generalize this to AI systems as well, especially because they involve autonomous learning processes. To use this generalized notion of commitment, we need a working definition of it. In the case of humans, commitment is roughly described as a feeling or sense to realize a duty held by a human being A in intersubjective relations with a person B that needs support to realize a goal-directed action and B expects A to deliver this support [MSK16].

Given our system perspective, we abstract from a feeling or sense of commitment and instead focus on an objective commitment independent of any subjective feelings; this connects to the idea that commitment is like a contractual relation [Jac+21], but we allow it to be realized by non-human agents. Then, the question

of whether an agent is committed to supporting a person can, in principle, be asked for artificial intelligence agents.

Since AI systems such as robots are increasingly being evaluated as agents [San20], we want to clarify whether there is an objective commitment as part of the design of the AI system when interacting with a human. The intuitive idea is that an AI system has an objective commitment to support a human B with its goal-directed action H if it is able to recognize that a human wants to realize H and has further features (based on other core features of trustworthiness than commitment) and then pragmatically performs the supporting action in an adequate way when expected by B and finally has some (direct or indirect) legal responsibility for its activities. More precisely, we want to rely on the following proposal in which we implement our systems perspective (according to which institutions and AI systems are candidates for commitment) and do not even presuppose any demanding notion of agency.

Definition 9.2.2 (**Objective Commitment of a System**). *Definition of an objective commitment of a system to support a human being distinguishing (a) epistemic, (b) legal and (c) pragmatic aspects: An AI agent has an fully-fledged objective commitment in relation to human B concerning a desired goal-directed action H of B, if the system (a) has the epistemic abilities of recognizing that B wants H to be realized with a certain priority and of recognizing that B needs support to realize H and the agent metacognitively controls for other core conditions of trust (five other dimensions according to our account), (b) is (directly or indirectly) legally responsible for its actions and (c) has the pragmatic features that it is designed to do the relevant supporting activity that is necessary to realize H – even in changing or challenging circumstances – with a high priority and with adapted support if expected to do so by B.*

For our purposes, we only need a working definition that keeps the core aspects of the intuition of commitment but transfers them to operationalizable features with our system perspective. We think that this proposal can do the job, and we apply it with exactly the three features (a-c) in the comparison of different systems.

With these two working definitions, our aim is to integrate uncertainty quantification and commitment into our general comparative framework. Including them, we propose six general trust dimensions of a system: objective functionality, transparency, uncertainty quantification, embodiment, immediacy behaviors, and commitment. This will be justified and unfolded below in Table 2.

Based on this, we receive a trust profile in a system with evaluations of the six dimensions of the system (see our Figure 9.1).

Thus, the comparison of trust in classical machines, AI systems, humans, and democratic institutions can be based on comparing the profiles of trust for these systems.

THE FOCUS AND THE STRATEGY OF INVESTIGATING THE TRUSTWORTHINESS OF AI SYSTEMS Trust has become a highly studied phenomenon in the formation of philosophical theories [MZ06; Haw14], in psychological studies [Blö21], and in the machine learning literature [GW20; Loc+21; SW18]. However, bridging concepts of these areas together in such a way that the newest technological possibilities and concepts in trust models are linked to the corresponding keywords in the literature remains a challenge overall.

We want to contribute a more detailed insight into these concept connections. To contextualize our theory, we review existing distinctions and relate our main dimensions to established concepts and definitions of trust [SW18]. We propose to make use of two distinctions in the literature: implicit versus explicit trust [Hof06] and warranted versus unwarranted trust [Jac+21]. Our focus is on explicit and warranted trust.

The trust relation between a person and a system is explicit if it involves conscious considerations, questioning, or justifications of the probability that the person or object can or will do the supporting activity. Moreover, an attitude of trust is implicit if it does not involve such conscious consideration, questioning, or justification. One central form of implicit trust is habitual trust in a situation that is based on many repeated situations of relying on the system, such that the trust relation is automatic, non-inferential, and unreflected [BT19].

As adults, we often start with explicit trust in an object, e.g., in a plane, by reasoning why it is trustworthy to fly, and after a lot of unproblematic flights, we develop habitual trust and no longer think at all about the trustworthiness of planes but just use them. Here, we want to elaborate on the properties of a system enabling explicit trust. A second distinction to classify trust is the one between warranted and unwarranted trust [Jac+21; BT19; SL21].

We focus on warranted trust from a system perspective and thus constrain our theory to describe the key objective dimensions of a system itself as the basis for an adequate evaluation of how far we can trust it, more precisely, its trustworthiness. This system perspective is most relevant for trustworthiness in the long run, the degree of trust is determined (at least ideally) mainly by the objective properties of a system. This holds if two conditions are satisfied, namely, first that the person has epistemic access to these

objective properties, i.e. the person has knowledge of these properties, and second that she evaluates these properties in a non-biased way (i.e. she does neither strongly underestimate nor overestimate the role of the relevant properties).

Often these two conditions are not met, i.e. system properties are often not accessible or not recognized in a non-biased way. To account for those cases, one would need to integrate this subjective knowledge, perception, or evaluation, and this would need a psychological theory of trust in a wide sense. But we decide here to set this aside as an additional aspect that can be accounted for after the objective foundation is clarified.

Our central question is: What are the key dimensions and features that form the foundation of an explicit and warranted trust relation to a system? To answer this question, we presuppose that the two conditions are satisfied to a normal degree, i.e. a person is objectively informed and is not heavily biased in the perception or evaluation of system properties.

To further embed our proposal, we start with a widespread definition that trust is "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" ([MDS95], 712). This idea goes back to the philosophical debate initiated by Baier [Bai86], who explicates trust as the acceptance by one person of the dependence on the goodwill of another person to support her.

This idea had to be generalized due to the argument that we trust a professional doctor even if we do not rely on his goodwill but on this professional duty to provide his medical support [One02]. Despite these corrections, the central point of this discussion is that Baier [Bai86] triggered a notion of trust which can be called a "second-person reliance account of trust" [Gol20] with the constraint that trust relations should be reserved for interpersonal relations.

We think that this constraint is not fruitful: First, there are still important gaps concerning the relation of epistemic conditions of trust in these second person reliance accounts [Gol20].

Second, even if those challenges could be resolved, the constraint is too exclusive since intuitively we also trust e.g. institutions or companies. In Germany, many people have trust in the democratic institutions, in the legal system, in the police system [Mad24]. This trust is not directed towards the individuals involved, but rather towards the professional execution insured by the organizational structure of the legal system.

Trust in complex systems is used in science, e.g. decreasing trust into democratic institutions is discussed as a reason for political changes in western democracies. Trust into systems should be separated from trivial cases of reliance on simple tools, e.g. using a hammer to drive a nail, which obviously are not cases of trust. The concept of trust in the case of systems presupposes a certain complexity and autonomy in the range of behavioral or cognitive possibilities:

While the effect of the hammer is fully determined by the user, this is not the case by minimally autonomous systems like a democratic institution. Due to autonomous processes, the behavior realized by the institution is partially independent of the people who are doing it. Since modern complex machines and especially AI systems are candidates that meet these minimal complexity and autonomy, we include them into our comparative perspective of trust. This leads us to the following minimal characterization of trust:

Definition 9.2.3 (Minimal definition of Trust:)**.** *A person P trusts an entity E (person, minimally complex object or system) in a situation S given a certain task, e.g. realizing a goal-directed action GA if at least the following (necessary) condition holds: The person P wants to realize a goal-directed action GA (important for P) and has to or wants to rely on E to realize her goal GA in the situation and expects E to deliver a support which is essential to realize GA. Furthermore, P registers or is disposed to register a risk that E is not delivering the expected support, irrespective of the ability to monitor or control the other entity E in S and E has a minimal range of behavioural possibilities (including the ability to deliver the expected support).*

We aim to develop this minimal definition of trust into a general framework that enables us to make a fruitful comparison of trust in persons, on the one hand, and in institutions and AI systems, on the other. As a baseline we also include classical cars into the comparison even though they are a borderline case given the characterization of minimal complexity and autonomy.

To account for institutions and AI systems, the dimension of uncertainty needs to be integrated into the concept of trust and to account for persons. We also need to include the dimension of commitment. Starting with the highly influential definition by Glikson and Woolley [GW20], we will see that they do not involve these two dimensions. We justify in more detail that it is fruitful to include them in our multidimensional account. Concerning uncertainty: The minimal definition involves the registration of a risk not to receive the expected support for a goal-directed action. In

the case of institutions and especially AI systems this can best be captured by the concept of uncertainty.

Concerning uncertainty, Rossi [Ros18] briefly mentions bias and mitigation strategies, as well as explainable AI (XAI) as a trust calibration tool. We believe that due to trust being fundamentally defined by its risk factors, the focus of defining trust in AI should include the risks of uncertainties for many features, namely uncertainties in the data itself, in the model and other types of uncertainties, even those based on external influences.

While uncertainty and risks are mentioned in some of the existing definitions of trust [SW18], they are mentioned as overall base concepts without unravelling the types of uncertainties that can occur and provide a risk factor. We aim to offer a more specific guideline to characterize warranted trust in AI systems in one respect by unfolding how uncertainty contributes to trust in AI.

Previous work has often focused on the role of explainable AI for trust [FL22] which we think is an important contribution to the dimension of transparency. But the question of how transparent a system is should be distinguished from the question how uncertain the processing of relevant factors like data or models is. Thus, uncertainty should be distinguished from and is a useful addition to other key concepts of trust, including transparency but also reliability.

We integrate the idea of the reliability of a system by characterization the dimension of objective functionality of it in relation to the expected support. Thus, we need to distinguish transparency, objective functionality and uncertainty as dimensions of trust. Although reliability and transparency are often considered in theories of trust [FL22; GW20], uncertainty is so far not integrated in to general accounts of trust that can be found in the literature (Bold references in Table 9.2). Thus, we integrate it and propose to do this by specifying key features realizing this dimension in line with our working definition.

9.3 THE FRAMEWORK FOR A NEW DETERMINATION OF TRUST

Let us systematically unfold our account: Starting with the minimal determination of trust (s. above), we also find many helpful additional distinctions in the literature. We want to either integrate them or determine the space of them in relation to our multidimensional account.

According to our system perspective, we first exclude subjective psychological conditions since our objective is to work out the key properties of the system itself to determine its trustworthiness. We know that whether a relation of trust is implemented depends on many factors that we want to abstract from in our approach, e.g., the personality traits of a person (how anxious or skeptical a person is, often also discussed under the keyword intrinsic trust) [Hof06].

Furthermore, external influences like the type of situation (whether it is extremely important for the trustor or not) can influence the trust setting and the type of task (whether the task is very difficult or rather simple concerning the system's abilities [GW20]. We are aware that all these aspects are modulating trust, but we want to focus on the key properties of a system because we are interested in the question under which conditions warranted trust in new technological systems, focusing on AI systems, can be developed in a society in the long run.

Another influential background condition for trust is the general purpose the system is intended to have: whether it is constructed so that it is beneficial to humanity and the planet in the sense that it promotes the well-being of humans and the environment and respects basic human rights [Flo+18].

These background conditions are spelled out in detail as fostering beneficence, avoiding maleficence, supporting human autonomy and social justice, etc. [TLS21]. Again, while accepting this as relevant, we suggest to focus on the objective system properties.

To develop our own multidimensional account of trust, we start with the important work of Glikson and Woolley [GW20]. They propose the following as basic dimensions of moderate warranted trust: task characteristics, tangibility, transparency, immediacy behavior, and reliability. We leave aside task characteristics since we want to focus only on properties in the system; thus, we are not discussing properties of the subject (the trustor), the specific situation, or the task. Their second aspect, tangibility, is introduced to distinguish embodied robots from virtual AI.

We suggest relabeling it as embodiment and keeping it. This represents that we trust something more if it is bodily present: we like to personify and be able to see and hear something that feels real. The more physically embodied something is, the more likely we are to trust it [GW20].

The third aspect proposed is transparency. We agree that it is one central dimension, but we think it needs to be spelled out in

four different features, contributing to transparency in different ways (see our summarized Table 2). They propose a fourth aspect of immediacy behavior that includes the observation that aspects of social responsiveness of the system support a trust relation.

Thus, we take over four of the six dimensions and extend the definition by two major contributions: uncertainty quantification and commitment. We also include new important features for each dimension that can lead to trust calibration, connecting existing work in the machine learning community with psychological research efforts.

9.4 DIMENSIONS OF TRUST IN AI SYSTEMS – A COMPARATIVE PERSPECTIVE

To develop a concept of warranted trust in AI systems, we are partially inspired by aspects that are relevant to interpersonal trust, but we also suggest that it is useful to compare trust in AI systems with classical but complex machines like normal cars and with (prototypical) democratic institutions. After already motivating the six central dimensions, we proceed as follows in the rest of the article: First, we describe the six central dimensions in detail by proposing the prototypical implementing features for each dimension.

To illustrate this, we first describe these features for classical machines like normal cars. In a second step, we apply these features to democratic institutions. These two steps fulfill two roles, namely to demonstrate that even for these traditional cases of discussing trust, our framework is intuitively plausible; furthermore, it is explanatorily fruitful since it shows quite some unexpected overlap of the level of realization of the six dimensions for these two rather different cases as well as differences to the case of trust in humans.

We then apply our framework in Section 9.5 to AI systems in general, including autonomous cars. Our focus in Section 9.6 is then the application of our framework to LLMs as a key AI application. Our evaluation of dimensions and implementing features is reported in detail in Section 9.5 and allows us to make the deficits of trustworthiness into LLMs visible in our comparative Figure 9.1.

We now elaborate on the six main dimensions of trust (namely, objective functionality of the system, transparency, uncertainty quantification, embodiment, immediacy behaviors and commitment) by describing some implementing features for each dimension.

To develop a general comparative perspective, we briefly illustrate

each feature of each dimension by looking at a classical mechanical system, namely normal cars, and compare them with specific implementations in AI systems (overview of the implementing features in column 1 in Table 9.1). Objective functionality has two underlying features, namely that the system has adequate properties to allow it to do the expected job, i.e. cars nowadays have the properties that allow us to drive safely on streets. Another feature is systematic long-term control of these features, which in the case of normal cars is secured by obligatory checks (in Germany by the TÜV).

In the case of a specific AI system, for example, convolutional neural networks for images, the objective functionality is realized by adequate AI programming [RN21]. Systematic long-term control is still in its infancy; it was only recently started by the EU AI Act, which aims to secure functionality, and to reach that aim, it demands some transparency in its legal rules [VZ21]. This brings us to transparency as the second dimension.

We think that we need to distinguish four different features enabling full-blown transparency to account for the specific challenges of AI system: transparency by observation of the system and habituation of use which is obvious for normal cars in everyday life, transparency by reputation which is based on the (long-term) reputation of a company which is easily accessible for normal cars; equivalently, the big AI players including Open-AI developed a good reputation up to now.

A third feature is independent reports, which are, e.g. delivered by easily accessible independent journals which exist for car reports everywhere. Finally, there is the feature of explainability which is in the case of normal cars highly fulfilled in the sense that nowadays almost everyone has a basic idea how to explain the core function of the motor of a car. These four features of transparency are already challenging and need to be elaborated on for AI systems (see below), and are especially informative for those. The same holds for the dimension of uncertainty quantification, which we unfold into two aspects of epistemic uncertainty, namely (i) of data and (ii) of the relevant model; furthermore, there is (iii) aleatoric uncertainty, as well as (iv) uncertainty because of external influences (lack of robustness).

For classical complex machines like cars, there is no uncertainty in any of these dimensions, while for AI systems, we are faced with uncertainties in all four respects, and only some systems are investigated in all four respects by now, such that there is a gap of knowledge about the range of the uncertainties (see next section).

Before we elaborate on the challenge of uncertainty quantification for AI systems, we briefly comment on the remaining dimensions and the features realizing them, namely embodiment, immediacy behavior, and commitment.

Embodiment is the dimension under which we account for the type of realization of a cognitive system: it can be completely embodied e.g. in the form of a robot that acts on the basis of an internally represented program, or completely disembodied, as in the case of interacting with a purely virtual character, e.g., communicating with an avatar, with Replica or with Chat-GPT [Ope24]; there are also mixed cases like acting in the real world, e.g., boxing, and thereby fighting against a virtual agent in the virtual world. The more physically embodied an AI system is, the more likely we are to trust it [Bai+11; Lee+06; RTM13; Sal+15; Rob+23].

In their extensive review, Glikson & Wooley [GW20] show that concerning the temporal unfolding of trust at the beginning, trust is lower for embodied AI in contrast to virtual AI, but then it reverses completely ([GW20] p.633). The long-term effect of embodiment on trust has recently been questioned or at least discussed [Van+17; KK16]. The dimension of immediacy behavior highlights two features of interaction with the AI system, namely the degree of illusion of control that is triggered by this interaction: this is rather high in normal cars: we think that we are perfectly controlled drivers although we know that accidents occur on a daily basis and thousands of people die every year. This is usually not involved in AI systems since they are constructed to be active as autonomous systems.

The second feature is the social responsiveness of the system, which is not given in the case of normal cars, but, e.g., with communicative robots like Sophia (based on Chat GPT [Ope24]), it is very high. Along the same line, increased interactivity with a robotic co-worker is shown to also increase the positive perception of the robot and thus its trustworthiness [Ois+16].

Finally, there is the dimension of commitment of the system in relation to the relevant task. We propose that it is implemented by three features, namely that the system is ideally constructed such that it comes with a high degree of commitment to implement dimensions 1 to 5, which we have already elaborated (meta-commitment: in normal cars, there are many control systems with lamps indicating if an important part is dysfunctional). Furthermore, it should ideally come with a high degree of legal responsibility (which, in the case of standard cars, is fully covered by insurance).

The final feature ideally involves a high degree of commitment to realize the expected support. This is an important aspect of interpersonal trust but can be applied to institutions and AI systems as well. This outline of a comparative perspective can be applied to the four relevant systems, while each system has its specific profile of trust with a degree of realization for each dimension, visible in Figure 9.1.

To make a comparative perspective most informative, the trustworthiness of AI systems is also compared to the trustworthiness of a prototypical, well-functioning democratic institution. We can use our dimensions and the implementing features to describe their profile of trust (the degree of realization of each dimension of trust). Let us shortly go through the six dimensions and their implementations (as presented in Table 9.1, column 1) mainly relying on Wille [Wil16] and the EU-principles of good democratic governance [Cou08]:

Well-functioning democratic institutions like a parliament of a country should have high-level objective functionality, including adequate properties such as efficiently making and implementing laws for the benefit of society. An essential aspect of this functionality is long-term control by the highest court, who provides checks on the parliament through its legal decisions.

Actually, this results in mutual oversight since the parliament controls the highest court by electing competent judges, while the court controls the parliament. Transparency is another crucial dimension. A democratic parliament should ensure high levels of transparency, which can be demonstrated through four features implemented in key practices: (i) regular public observations, e.g., by broadcasting all meetings of the parliament, (ii) the reputation for transparency, (iii) independent reports, e.g. by critical journalism, and (iv) explainability of the process and results of the activities of the parliament.

The latter is, e.g. realized by the public speeches justifying and explaining the laws and their functions. The third dimension of uncertainty quantification is typically only realized at an intermediate level, realized by the four implementing features: the uncertainty quantification (i) of the data and (ii) the relevant model, which is intermediate level because the parliament has, e.g. typically to make political laws concerning the economy of a society with quite some data and an economic model but the data are always incomplete, and the model always remains underdetermined and partial.

Thus, the uncertainty quantification of legal decision-making of a parliament comes at an intermediate level: Parliaments often

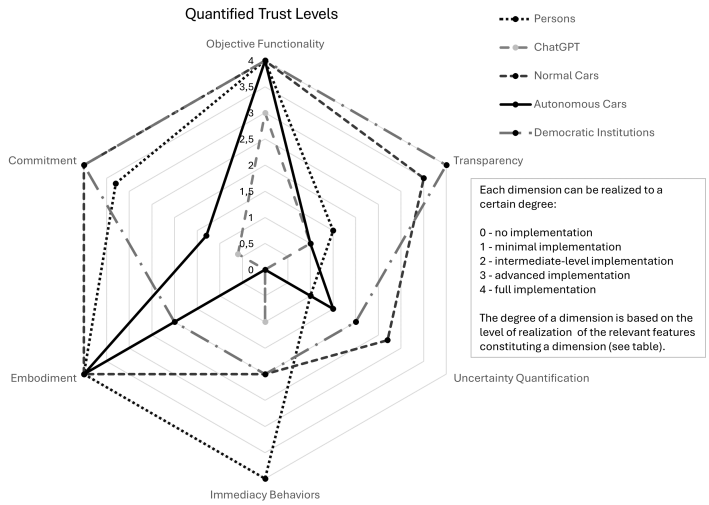


Figure 9.1: **Quantified Profiles of Trust:** For each dimension we describe the degree of implementation. This is done by estimating on the basis of commonly accessible knowledge for each feature of a dimension to which degree it is realized for a system (low = 0, intermediate = 2 or high = 4). The average value of the degrees of implementation for all the features of a dimension results in an average value for the dimension. Thus, we receive a profile of trust for each system. The resulting scheme allows us to see that if all levels of the profile of trust are relatively well fulfilled, the system itself is generally deemed trustworthy, such as with normal cars. The more trust dimensions are lacking or unfolded to a low degree, the more variety can be seen in an individual's trust levels and the less trustworthy it is.

legislate based on incomplete data and inherently limited models. Both data and models carry aleatoric (random) uncertainty and limited robustness——especially in the face of major economic changes or external attacks, e.g. by terrorists or other states. Therefore, even with good informational access, parliamentary decisions must be made under a significant degree of uncertainty. The fourth dimension of embodiment is also of an intermediate level because the members of the parliament are embodied, but the institution of the parliament is a disembodied legal institution constituted by the rules and regulations of the parliament.

Concerning the fifth dimension of immediacy behavior, a democratic parliament typically only realizes an intermediate level, namely concerning the feature of the illusion of own responsibility, the citizen actually has only partial control via the democratic elections, and the elected members have quite some independent room for decision making. Similarly, the social responsiveness of

the parliament is constrained: the members of the parliament do interact with the citizens, but only to a rather limited degree (except for the election times).

The sixth dimension of commitment is typically highly expressed in a democratic institution in all three implementing features, i.e. a democratic parliament has to guarantee the optimal handling of the first five dimensions just described. It is also legally responsible for the laws it makes, e.g., the parliament accepts the voters' decision in the next election, on the one hand, and, e.g., the decision of the Court of Human Rights, on the other. And, of course, a democratic parliament is committed to delivering the expected support for the benefit of society. This results in a high-level degree of realization of dimensions 1, 2 and 6 and an intermediate-level degree of realization of dimensions 3, 4 and 5.

This profile is visible in Figure 9.1. Interestingly, the levels of trustworthiness realized in the six dimensions for classical machines like normal cars have quite some similarities with the expected levels of trustworthiness in prototypical democratic institutions: There is quite a high level of objective trustworthiness. We now focus on the technological perspective to highlight the most relevant dimensions and features in more detail for AI systems, and especially elaborate on the special contributions in a general framework, including uncertainty quantification and commitment.

9.5 DIMENSIONS OF TRUST IN AI SYSTEMS- A TECHNOLOGICAL PERSPECTIVE

A further goal is now to specify profiles of trust in AI system by taking a technological perspective. As we will see, this allows us to highlight some dimensions that are specifically relevant to AI systems and also mark gaps in research that we need to fill to adequately describe the level of warranted trust in AI systems.

From a technological perspective, it is important to distinguish supervised and unsupervised systems: The more freedom of decision an AI learner is given, the harder the task to be solved becomes. The more priors we know about the data, the more we can tell whether the task was solved wrongly or correctly. Therefore, consider the distinction as an indication of the level of priors known beforehand and less as a strict distinction between learning paradigms. Supervised means that the task is well-defined enough that the input data has clear labels, so we know which output labels might occur [CCD08].

We distinguish supervised and unsupervised in order to highlight the difference between trust when we have a lot of known priors of the data, and when we give the algorithm a lot of freedom. We want to discuss these extremes and, therefore, do not additionally distinguish reinforcement learning, as we only want to distinguish the extremes clearly here. Reinforcement learning has a "scoring system", so some things are known about the data, such as in supervised learning. However, unlike supervised learning, there are no set outcomes.

For example, if a reinforcement learning system studies chess, it only gets points for gaining material, but the moves it is allowed are entirely up to the learner itself. Hence, we distinguish supervised and unsupervised learners as paradigms and consider reinforcement learning somewhere in the middle of this scale. In an unfolded version, this could be integrated at a later stage [CCD08]. The special contributions include uncertainty and commitment, and we will deliberate why these are such important additions in a general account. But let us start with the embodiment: For ChatGPT [Ope24], one of the most powerful AI tools, we have a case of a disembodied system. But if a large language model is, for example, implemented in a robot-like Sofia, then this can be embodied.

Uncertainty quantification itself is an important trust tool- knowing how much uncertainty is in the data, whether there are data biases present, and whether a model is uncertain about its predictions can really help a user calibrate warranted trust. Another aspect of uncertainty research is the inherent, the so-called aleatoric uncertainty. This can occur when the data itself is noisy, or we have other randomness as part of the data that cannot be learned with more training data. Furthermore, we include robustness as another uncertainty factor.

In the AI community, vulnerability to so-called adversarial examples is a huge research area [Don+18; Ma+18]. They have become so relevant because even the best machine learners can be attacked with known attack techniques under specific circumstances, which means that we always have to accept some kind of risk when using machine learners. The question is whether these risks are likely to be abused as these adversaries require an active attacker to be present. However, as long as these uncertainty aspects are unknown to a user, there will never be a chance of fully calibrating the warranted trust.

Beyond the minimal expectation of objective functionality of AI systems, e.g., that autonomous cars have adequate properties to drive safely in cities, an open challenge remains the systematic

long-term control of this functionality for all AI systems.

The EU AI Act is one of the first such efforts to define AI regulations, and will be followed by more and more of these regulations [VZ21]. There have also already been first efforts to investigate the effect of AI certificates on perceived trustworthiness [Wis+24a]. With fabricated seals, the trustworthiness does not increase with the seals given, but it leaves the open question of how the perceived trustworthiness would change if real companies or actual seals were involved. We currently face the problem that there are, as of yet, no official certifications/ seals [Win+21].

We distinguish transparency by observation, reputation, independent report, and degree of explainability. Observation means how frequently we observe an AI system in everyday life. As we know from normal cars, we have a very high level of trust simply due to habit. But autonomous cars are only running in San Francisco, and at the moment, only one company is left over.

The relationship between habit and trust has been highlighted by Farivar et al. [FTY17] or numerous other works. AI systems, as with any other new technology, need to overcome the barriers to familiarity and habituation. Thereby, a core aspect is the reputation of the company producing an AI system: The big international IT players have managed to offer services of standard IT and the internet that no one wants to miss, and many rely on: Google Maps is more reliable and informed for giving driving directions than an inbuilt car navigation system. Thus, Google, like all the other big players, can rely on its reputation and start to use it to introduce AI products. They already use AI systems to determine traffic patterns and estimate time of arrival, e.g., on Google Maps.

Furthermore, Google has started to test AI-generated answers in regular search results, i.e., they are not taking the content from other pages on the internet, but they create an answer by an AI system (and they mark it as AI-produced). In this context, independent reports on the use of AI in products would increase transparency.

They are still rare concerning AI systems and remain a desideratum. The most specific feature to enable transparency is explainability: Explainable AI is often defined as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans [Arr+20]. Explainable AI, in particular, is a rapidly expanding area of research with countless papers to its name [LL17c; RSG16].

Immediacy behaviors in our case consist of, on the one hand, socially oriented gestures intended to increase interpersonal close-

ness, such as proactivity, active listening, and responsiveness [GW20], which we call social responsiveness. However, we also include the consideration of whether a system is able to react to mistakes, whether it can adapt to mistakes or not with fallback mechanisms, or when it is pointed out to the system. Furthermore, we note that another implementing feature is the illusion of own responsibility.

If we think we are in control somehow, then we are more likely to trust an instance than not [Now21]. This is related to our self-efficacy, which can be characterized by the individual's belief in their ability to act successfully [Hsu+07]. Similarly, Pieters [Pie11] states that trust is a form of self-assurance and, therefore, can be used for the acquisition of trust.

The relationship between self-efficacy and trust has also been tackled in various works [TG09; OEK18], supporting the relevance of this dimension. The last dimension of trust that we include is commitment. This is obviously relevant for interpersonal trust. We argued that it is fruitful to transfer this dimension to trust in AI systems. We propose three commitments to a cognitive system from a pragmatic perspective. First, ideally, an AI system is constructed such that it involves the ability to guarantee that the other five dimensions of trust are fulfilled, i.e., it involves some meta-representation of all the dimensions guaranteeing its functionality in a wide sense.

Secondly, commitment can be assured by legal responsibility, which will also impact how trustworthy a system is. This demands that autonomous cars and other AI systems would come with insurance for damage they produce: now we observe legal trials concerning accidents with autonomous driving cars since the legal situation is unclear [DH13]. What if someone signs a contract on the basis of AI-generated information that may be partially false and misguided?

Third, we propose a commitment of the system to provide support in the expected way: e.g., ChatGPT is informatively communicating with us and we expect it to follow the rules of informative communication [Moo17]. This includes the expectation that informative communication aims at truth and does not invent stories. However, we all know that ChatGPT is not yet committed to these principles; it only aims to produce a plausible communicative response with its word probability evaluation system.

Thus, one radical difference is that humans and classical complex machines are committed to producing the expected support in the expected way, while AI systems so far are basically constructed and

arranged without commitments. At least, there is no commitment to truth. After this general overview, we now elaborate on the trust profile using the popular example ChatGPT.

9.6 CHATGPT- A CASE STUDY

ChatGPT is a well-known virtual interaction partner with astonishing linguistic abilities [Koc+23]. With regard to the objective functionality of ChatGPT, this application is often misunderstood. What most people are not aware of is that ChatGPT is a large language model, which means it is supposed to form comprehensive, meaningful sentences with regard to questions being posed.

It produces a sequence of words on the basis of the highest probability of the next word given the question and the words already chosen. The probabilities are predicted on the basis of enormously large language datasets. Considering this particular task, ChatGPT is spectacularly good, fulfilling adequate properties at a very high standard.

But a lay person might think that ChatGPT is also capable of performing mathematical proofs or citing scientific work appropriately, which it can only do to some degree. Although more and more tasks are trained using the newer ChatGPT versions, this application is obviously still not good at every possible task imaginable.

When people do not understand what particular task the machine learner is designed for and consider it to be a jack-of-all-trades, this might lead to over- or under-trust because the objective functionality of the learner cannot be specified correctly. ChatGPT so far has no systematic long-term control, quite the opposite. The model is constantly changing, training using the inputs and prompts given to it on a regular basis.

This means that overall, the objective functionality as a dimension is not completely fulfilled in all our defined subdimensions. With respect to transparency, the first feature is habituation. Habituation of ChatGPT is already happening and will be extremely "high" within a couple of years, and increases with frequent use. Considering the reputation, this ranges from very low to very high, depending on the level of trust in machine learning algorithms in general.

In terms of independent reports, ChatGPT is used independently by very different people every day, all of whom have their own

experiences with what ChatGPT excels at. However, it has only recently been introduced to the commercial market and, therefore, this subdimension is also not very high yet, but it will change over time. So far, the model has no means to learn how the output is generated when in commercial use. So far, no explanation for the outputs has been available. This means that regarding transparency, part of it is non-existent, and part of the dimensions are well covered. As an example, the system is now in very common use, so observation of the system is high, but explainability is not currently given.

When taking a look at uncertainty quantification, we have so far no means to uncover hidden data biases or predict a range of outcomes. Considering that ChatGPT is trained with, for example, a lot of religious text, there will definitely be underlying biases in the data that are not so easily uncovered. The model uncertainty is not displayed to a regular ChatGPT user at all, nor is any aleatoric uncertainty, which would be important for risk assessment.

Furthermore, we only know, due to some research experiments, that ChatGPT can be compromised using prompt injection attacks [Gre+23]. Prompt injects can lead to ChatGPT restrictions, such as ChatGPT not being able to output instructions on how to build, for example, a bomb. They can be easily circumvented [Kil25]. Even worse, prompt injections can lead to adversarially chosen or arbitrarily wrong summaries of documents, as well as to the output of malicious links that then pose security threats [Gre+23]. As for embodiment, ChatGPT is entirely virtual; therefore, this trust dimension is not fulfilled.

To continue with the dimension of immediacy behavior, ChatGPT does not give you any illusion of control whatsoever, but is capable of social responsiveness since it is designed as a communication tool. The problem is that it produces the illusion of a communication partner with real understanding. If you point out to ChatGPT that it made a mistake, it can correct itself.

But concerning commitment, there is a big deficit in ChatGPT: we expect a communication partner epistemically to be committed to telling the truth and not to invent evidence when we ask for information. ChatGPT is committed to neither: Given the predictive probability word selection algorithm, it is designed to produce a plausible communicative response given the large database of former communications. There is no guarantee or meta-control of the optimal handling of the other dimensions of truth (e.g. uncertainty quantification) such that it is usually intensely biased in an intransparent way due to biases in the data base.

So far, there is no legal responsibility if ChatGPT fails somehow and produces harm due to false information, e.g. Chat-Bots can be designed to isolate persons from their social environment, which recently resulted in the suicide of a teenager. Currently, no company has accepted legal responsibility.

Finally, there is no relevant pragmatic commitment since ChatGPT is only a communicative agent, and thus, e.g. it cannot commit to non-communicative aspects of promises, e.g. like keeping the promise to meet a person. In conclusion, you can see here how very few trust dimensions this tool actually fulfills. Following our methodology, this also shows which dimensions would have to be improved in order to make it trustworthy to a degree comparable to the trustworthiness of, e.g. normal cars.

9.7 CONCLUSION

We propose a multidimensional account of trustworthiness since this type of approach enables a systematic comparative perspective for rather different types of systems, including further advantages [DN25]. We highlight six dimensions with a special focus on uncertainty quantification and commitment.

AI systems are coming with a special challenge in these dimensions, which we visualize in our general framework of trustworthiness. We highlight that uncertainty quantification needs to be specified in detail to account for unknown epistemic uncertainties about data and models, as well as for aleatoric uncertainties intrinsic to the AI systems.

We also need to include the risk of external attacks modifying the processing, as in the case of adversarial attacks. Furthermore, AI systems – partially due to the characteristics of deep learning – are not constructed with a commitment to deliver the expected support (e.g. ChatGPT lacks an epistemic commitment of adequate knowledge of the needs of the interaction partner as well as a commitment to truth in communication, it lacks any direct or indirect legal responsibility and finally lacks pragmatic commitments like keeping a promise).

These distinguishing features allow us to compare warranted trust in AI systems with trust in humans and classical complex machines as well as democratic institutions. Thus, our multidimensional profile account is well suited for a comparative perspective to highlight the differences between the relevant cognitive systems and to see

what future AI engineers need to realize to form a trustworthy system. It helps existing or future certification processes, as well as defining a common standard that any AI system should fulfill. This could largely benefit the trust calibration and the quality of future AI standards.

Table 9.2: Related Work Overview: Important Trust Papers

Reference(in bold: general accounts of trust)	Content	Uncertainty dimension
1. In AI we trust? [Ara+20]	How personal characteristics can be linked to perceptions of automated decision making	✗
2. Perceived trust in AI technologies [Bit+20]	Focused on user's perception of trust and the correlation between task difficulty and perceived characteristics	✗
3. Personalized uncertainty quantification in AI [Cha+25]	Individual-level predictions and quantification of uncertainty, especially for underrepresented groups	Defines specific challenges for individual UQ, but lacks general trust framework
4.State of the art in enhancing trust in ML models using visualizations [Cha+20b]	Visual trust aspects and uncertainty awareness discussed	Mentions uncertainty awareness, no full trust framework
5. Impacts of attitudes, public trust in AI [CW21]	discusses whether personality traits influence trust in AI	✗
6. Trust and ethics in AI [CDR23a]	Ethical requirements for trust in AI, including accountability and transparency	✗
7. Trust and acceptance of AI [CDR23b]	Trust perception using AI voice assistants	briefly mentions that uncertainty and risk are connected, but does not mention uncertainty quantification as means for trust calibration
8. **The role of uncertainty quantification in AI [Deu+24]**	argue that uncertainty quantification is essential for AI-based systems	link factors such as explainability and uncertainty and trust, but does not integrate all relevant factors
9. Physician Understanding and Trust [Dip+20]	Local XAI linked to user trust and understanding	✗
10. Trust in AutoML [Dro+20]	effect of transparency on trust and trust in AutoML in general	✗
11. How explainability contributes to trust in AI [FL22]	philosophical explanation of trust and XAI relationship	✗
12. Attachment and Trust in AI [Gil+21]	association between attachment styles, trust based on affect, opposed to our approach, which focuses on trust in systems	✗
13. **Human trust in AI [GW20]**	Focus on cognitive trust, the role of AI's anthropomorphism specifically for emotional trust	✗
14. **Formalizing trust in AI [Jac+21]**	description of cognitive mechanisms of trust, warranted trust	describes uncertainty in the context of risk, without connecting it to uncertainty quantification from the computer science domain
15. ML based trust computational model for IoT services [Jay+18]	proposed trust metrics with numerical values such as co-work relationship- experience and relationship metrics	Mentions uncertainty in abstract, then never again
16. **To trust or not to trust a classifier [Jia+18]**	propose a new trust score as alternative to confidence scores	basically criticize the very naive approach of confidence score as metric for uncertainty quantification, but do not elaborate the vast field of options beyond the naive approach
17. **Trust in AI: Meta- Analytic Findings [Kap+23]**	Predictive trust factors split into human trustor, AI, trustee and shared context factors	✗
18. How do I fool you? Manipulating User Trust via XAI [LB20]	prove that malicious XAI can create overtrust	✗
19. Trust in AI: comparing human and automated trustees, bias [Lan+23]	distinguishes between ability, integrity and benevolence and focuses on interpersonal trust in comparison to trust in AI	✗
20. Generating with confidence: Uncertainty quantification for black-box LLMs [LTS23]	present a method for estimating uncertainty and confidence in black-box LLM	presents a method for UQ rather than providing a framework for measuring trust
21. **Uncertainty quantification and confidence calibration in large language models: A survey [Liu+25]**	Classifies LLM-specific uncertainty types and computational profiles	mention interpretability, robustness and reliability but lack our comprehensive summary of all relevant factors and focuses on specific LLM challenges

Reference (in bold: general accounts of trust)	**Content**	**Uncertainty dimension**
22. **A review of trust in AI [Loc+21]**	description of 5 trust challenges including key words such as transparency, explainability, accuracy, reliability, embodiment etc.	Mentions uncertainty only in context of risk, no elaboration
23. **Trust in AI: Foundational Trust Framework [LMS22]**	An effort to combine interdisciplinary surveys into one trust framework	✗
24. Trust in distributed AI [Mar92]	Trust in multiagent systems, and allowing interactions between agents	✗
25. **Confidence Meets Accuracy- indicators of trust in ML [RY22]**	show that different trust measures can capture different types of trust and seemingly contradict themselves	they state in their limitations that their results might change if uncertainty quantification is considered
26. Building trust in AI [Ros18]	Focus on explainability and bias/fairness	no connection of the fairness term to the vast field of uncertainty quantification in the research area
27. Towards trust in ML for healthcare and criminal justice system [RU18]	argues that black box models are no longer necessary	mention uncertainty, but do not include it in their calculations
28. **In AI we trust: Ethics, AI, reliability [Rya20]**	argues that AI trust is actually reliance due to lack of emotions, trust definitions based on interpersonal trust	✗
29. Quantifying interpretability and Trust in ML [SB19]	Measures of XAI quality, measuring trust using quality of explanation vs human trust	✗
30. Trust it or not: Effects of ML warnings [SXL19]	Warning labels discuss effect of warning labels, repetitions	mentions uncertainty in the context of familiarity only
31. A Survey on Uncertainty Quantification of LLMs: Taxonomy, Open Research Challenges, and Future Direction[Sho+25]	Overview for existing methodologies and clustering of methods of uncertainty from a computer science perspective	no general account of trust in AI systems
32. Building Trust in AI, ML, and Robotics[SW18]	Definition of Trust and Promoting Factors such as reliability, validity, utility, robustness and false-alarm rate	✗
33. Misplaced Trust in ML: Interference of ML in Human- Decision Making [SLL20]	discusses overtrust, and key metrics for trust calibration	✗
34. Should We Trust AI [Sut19]	Trust in the context of vulnerability, reliance, risks, and emotional factors	✗
35. Relationship between trust in AI and trustworthy ML [Tor+20]	Focus on fairness, explainability, audibility, and safety	✗
36. Student Perceptions of ChatGPT use [Tos+24]	Trust and reliance of ChatGPT: highlight the lack of confidence feedback of ChatGPT	Do not recognize the need for general uncertainty quantification, but are unable to name the explicit suggestions for the uncertainty dimension that we make in this paper
37. How to evaluate trust in AI assisted decision making [VBC21]	An analysis of review papers leading to a separation into vulnerability, positive expectations and attitude	Uncertainty in the context of risk and vulnerability is mentioned, but not uncertainty quantification specifically as the tool to quantify and calibrate trust
38. A survey on trust evaluation in ML [Wan+20a]	covers prerequisites of trust evaluation and then lists method types- no focus on trust definition but how trust can be evaluated	Mentions that trust can be evaluated using uncertainty, but does not list a complete list of components of trust
39. **User trust in AI: Framework [YW22]**	Literature review on user trust from 2015–2022	Mentions uncertainty avoidance and risk, no quantification
40. How do visual XAI and end user's Trust [Yan+20]	investigate effect of XAI on user's appropriate trust	✗
41. Understanding the effect of Accuracy on Trust in ML [YWW19]	found that accuracy was important for trust calibration	Mentions UQ as important but no full framework
42. **How transparency modulates trust in AI [ZBW22]**	draws the connection between explanations and transparency, discusses trust calibration and the harm of wrong explanations	mentions uncertainty and confidence scores in the connection of vulnerabilities without defining the types and just as an example of trust calibration. No overview provided of all the factors of trust calibration for a system.
43. Effect of confidence and XAI on accuracy and Trust calibration [ZLB20]	study the effect of showing confidence score and local explanation for a particular prediction	state the desire for essentially uncertainty quantification, do not mention how

Reference (in bold: general accounts of trust)	Content	Uncertainty dimension
44. **Reliability engineering, risk management, and trustworthiness assurance for AI systems [Zha+25]**	defines types of UQ and verification (so robustness measurements), which they name functional reliability of the AI	key focus is UQ and not the decomposition of trust into its factors, in contrast to our proposed trust framework
45. A ML based Trust Evaluation Framework, Social Networks [ZP14]	trust as a classification problem and proposition as an ML approach for social networks	✗

AI Systems: Examples. Column 1: Dimensions and their implementing features, Columns 2 and 3 Examples in the Supervised and Unsupervised Domain		
Dimensions of Trust	**Supervised**	**Unsupervised**
1. Objective Functionality		
(i) adequate	Convolutional Neural Networks [Li+21b] (Images) [Yu+21]	ChatGPT [Koc+23]
properties	Recurrent Neural Networks [MJ+01] (Speech Recognition) [Mak+19]	Autonomous Cars [Par+22]
(ii) systematic long-term	first efforts by the EU AI act [VZ21]	first efforts by the EU AI act [VZ21]
control	other ethical guidelines, i.e. by the UNESCO [UNE21]	other ethical guidelines, i.e. by the UNESCO [UNE21]
2. Transparency		
(i) observation of system	Decision Trees [SY15]	ChatGPT [Koc+23]
/ habituation	recommendation systems for movies [CRA]	Large Language Models [Cha+24]
(ii) reputation (chain of trust)	AI Seals [Wis+24a]	limitation studies of i.e. ChatGPT [Aza22]
	institutional trust in i.e. Google's networks	reputation of i.e. Deep Neural Networks [Sam+21]
(iii) independent	Neural Networks [SS97] in everyday use	Unsupervised Neural Networks [Dik+18] in everyday use
report	Convolutional NNs [ON15]	Netflix Algorithm [GH15]
(iv) explainability: experts or lay persons /	Layer-wise relevance propagation [Mon+19]	Label-Free Explainability for Unsupervised Models [CS22]
global or local	LIME [RSG16]	Unsupervised DeepView [NM22b]
3. Uncertainty Quantification		
(i) of data (epistemic)	Wang et al. [Wan+21b]	Taha et al. [Tah+19]
	CLUE [Ant+20a]	Kazemi et al. [KTW23]
(ii) of the relevant model (epistemic)	Zhou et al. [Zho+22]	Lee et al. [LL20]
	Hu et al. [Hu+19]	Kontolati et al. [Kon+22]
(iii) aleatoric (inherent randomness)	Sambyal et al. [SKB22]	Kazemi et al.[KTW23]
	Hüllermeier et al. [HW21]	Jiang, Minqi, et al. [Jia+22]
(iv) to external influences (Robustness)	Verification of Networks [TS+97]	Likelihood of Adversarial Attacks [Ma+18]
	Adversarial Examples (FGSM) [Don+17]	Quality Guarantees Autoencoders [Böi+21]
4. Embodiment		
(i) Physical or Virtual	Classification algorithm (virtual) [Nas17]	Virtual AIs such as ChatGPT [Koc+23]
	A robot using the supervised learning paradigm (physical)	Physical AIs such as a robot that uses ChatGPT for speech
5. Immediacy Behaviors		
(i) Illusion of own responsibility	Illusion of some control [Now21]	illusion of no control, i.e. ChatGPT [Koc+23]
(relates to concept of self-efficacy)	Supervised learning for trust assessment [Hau+13]	negative bias towards unsupervised learners [Wat23]
(ii) Social responsiveness	no responsiveness, only metrics like accuracy [ZLB20]	some responsiveness i.e. ChatGPT[Koc+23]
	no responsiveness, metrics like AUROC scores [Rei+21]	often also no social responsiveness, i.e. Deep Neural Networks [Sam+21]
6. Commitment		
(i) guarantee of optimal handling of 1,2,3,4,5 (certification)	first efforts in certifications of AI [Win+21]	certifiable unsupervised AI [Lan22]
	i.e. guarantee of adequate properties by evaluation metrics [Rei+21]	i.e. guarantee of some transparency by using explanations [CS22]
(ii) legal responsibility	responsibility gap AI [SM21]	Schuett et al. [Sch+19]
	health care responsibility [Nai+22]	Kemp [Kem18]
(iii) to realize the expected support	i.e. online forums like Towards Data Science	often non-existent (ChatGPT)
	Environmental Decision support system frameworks [Cor+00]	Youtube, i.e. Online forums (Towards Data Science)

Table 9.1: Trustworthiness of a system: distinguishing dimensions, implementing features and offering examples

11

IN SEAL WE TRUST? INVESTIGATING THE EFFECT OF CERTIFICATIONS ON PERCEIVED TRUSTWORTHINESS OF AI SYSTEMS

Trust certification through so-called trust seals is a common strategy to help users ascertain the trustworthiness of a system. In this study, we examined trust seals for AI systems from two perspectives: (1) In a pre-registered online study with N = 453 participants, we asked whether trust seals can increase user trust in AI systems, and (2) qualitatively, we investigated what participants expect from such AI seals of trust.

Our results indicate mixed support for the use of AI seals. While trust seals generally did not affect the participants' trust, their trust in the AI system increased if they trusted the seal-issuing institution. Moreover, although participants understood verification seals the least, they desired verifications of the AI system the most.

11.1 INTRODUCTION

Artificial intelligence (AI) systems are ubiquitous and have become integral to everyday professional and private life. AI systems such as Open AI's ChatGPT or Google's BERT can generate meaningful text [Feu+24], other AI systems are components in safety-critical applications such as those that enable autonomous driving [Gri+20], and even further, AI systems process highly sensitive health information such as echocardiograms [Mad+18].

Simultaneously, these systems and their underlying building blocks, such as deep learning models, have become very complex, aggravating the so-called black box phenomenon. Consequently, knowing when (not) to trust an AI system can be challenging for different stakeholders, from users to decision-makers and even developers.

While efforts to develop inherently trustworthy AI systems are much needed, approaches solely focusing on technical aspects are insufficient, as trust results from a system's perceived rather than its actual trustworthiness. Consequently, users sometimes perceive a system inappropriately, placing either too much or too little trust in an AI system.

To help users' trust calibration, different paths can be taken. One popular and well-researched example is explainable AI (XAI), which aims to increase an AI systems' intelligibility by providing explanations for the system's behavior, making internal processes visible, and increasing the overall transparency of the system [Arr+20].

Typical methods of XAI are, for example, visual explanations such as heat maps, which highlight areas of input data that were most influential for the system's output, or textual explanations which provide written or oral statements of the explainer. However, XAI is no panacea to cure a lack of trust, and concerns have been raised in terms of users' cognitive biases [Ber+22] and the cognitive burden that explanations pose on users when explanations are not designed with the end-user in mind [Mil19].

In this paper, we aim to counter the shortcomings of XAI and tackle the problem of trust from a different perspective. We empirically explore the effects of AI certifications, so-called AI seals of trust. Such seals are credentials which certify that software has been tested and validated to meet specific predefined criteria or standards in various dimensions. Theoretically grounded in works on epistemic trust, trust theory, signaling theory, and persuasion literature, we examined the effects of three different AI seals of trust in a quantitative online experiment.

To do so, participants of our study either viewed an AI system with (experimental groups) or without (control group) an AI seal of trust. In addition, in a qualitative part we asked participants in an open-ended format about their preferences for AI certification.

The importance of this work is underlined by initiatives such as the EU AI Act, which suggests certification as a central mechanism to communicate to the public the compliance with industry and legislative requirements. To date, however, empirical studies investigating the effects of such certifications for AI systems are scarce.

11.2 RELATED WORK

FROM TRUST IN AI TO CALIBRATED TRUST IN AI To describe and define trust in AI, previous work builds on thoughts from various disciplines, such as philosophy, sociology, and psychology that predominantly examine trust as an interpersonal judgment between two or more individuals. Moreover, choosing interpersonal trust as a starting point to examine trust in AI seems sensible as humans, at times, react socially to machines [NM00].

In fact, the most widely adopted definition of trust in automation originates in Mayer et al.'s [MDS95] dyadic model of organizational trust, in which trust results from a person's (the trustor) perceptions of another person's (the trustee) ability, benevolence, and integrity. While the direct application of an interpersonal trust conceptualization might be appropriate for certain occasions, this is not always the case [MW07].

Hence, emanating from Mayer et al.'s ability-benevolence-integrity framework, Lee and See [LS04] postulate that for a person to trust a machine, the person needs to assess the perceived reliability and functionality of an AI (ability = performance), the intentions with which it was built (benevolence = purpose), and the intelligibility of AI (integrity = process). Beyond these three trust antecedents, Lee and See [LS04] define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (p. 54).

Hence, users' trust must be appropriately calibrated to the system's actual trustworthiness [LS04; MW07; PR97]. As described above, users' trust depends on various factors, such as the system's overall performance or the perceived integrity of the system. However, cognitive and social psychology insights suggest that users' perceptions can be distorted, possibly leading users to place too little or too much trust in a system.

Such a mismatch of the perceived and actual system trustworthiness can result in either the system's disuse (i.e., resistance to use the system) or the system's misuse (over-reliance on the system). Both disuse and misuse pose serious consequences. In the context of semi-automated driving, for example, ignoring and over-relying on autopilot has led to deadly incidents. Hence, reaching calibrated user trust is essential.

To calibrate user trust, different approaches have been taken. Wischnewski et al. [WKM23] offer a systematic overview of previous approaches. In their work, the authors surveyed different empirical, human-centered interventions to match perceived and actual system trustworthiness for automated systems accurately.

Many of the interventions reviewed aim to increase a system's transparency, assisting the users' trust assessments by making the system more intelligible. While some interventions successfully calibrated the users' trust in a system, in some cases, the intervention also increased the users' workload [Kun+19] or led to overtrust [YW01].

In addition, adding, for example, explanations for increasing transparency had adversarial effects, eroding the users' trust, which

Kizilcec [Kiz16] explained by arguing that the additional information might have been confusing for users, reducing their understanding instead of increasing transparency.

Even though these transparency interventions have shown mixed effects, there are other reasons to question these approaches. First, many interventions are not developed for end-users but for developers themselves to make the inner workings of AI more transparent [Mil19]. However, explanations are likely to be less successful without the end-users in mind. Second, implementing additional measures such as explanations to increase users' trust shifts the responsibility of being trustworthy from the AI system and its developers to the users, who must determine whether the AI system is trustworthy.

Third, previous research has also shown that some users do not want to know how systems, in particular AI systems, work. They would rather stay willfully ignorant because they fear that knowing how a system operates might stop them from using it [NK22].

To conclude, while understanding- and transparency-enhancing approaches aiming to increase user trust indeed hold benefits, they also come with many downsides. In the next section, we suggest a different approach to user trust: epistemic trust through AI seals of trust.

EPISTEMIC TRUST IN AI AND TRUST IN AI-AS-AN-INSTITUTION

One of the main assumptions of understanding- and transparency-enhancing approaches to increase trust in AI, such as explanations or cues, is that users carefully assess the trustworthiness of AI to know whether they can trust it or not. Implicitly, this assumption often entails that users make rational choices about a system, that is, choices based on accurate perception and inference. However, as shown in the previous section, this assumption does not always hold.

We suggest that an alternative to such understanding-based trust is epistemic trust. Individuals show epistemic trust (see also, trust in testimony, Coady [Coa92]), whenever they accept communication or communicated knowledge from others as trustworthy, generalizable, and relevant [Spe+10].

In other words, when individuals trust what others tell them, they show epistemic trust. One could quickly assume that, as such, epistemic trust is equal to blind trust. However, individuals only assume information to be truthful and relevant when contextual or content cues like source credibility or plausibility evaluations do not indicate otherwise [GTM93].

In the context of AI systems, showing epistemic trust in the communication of especially experts can ease their trust assessments, as it is easier for them to ask "Whom to believe?" instead of attempting to understand the AI system. Examining epistemic trust in science communication, Bromme and Gierth [BG21] argue that, while from a classical logical perspective, to judge the trustworthiness of someone (or something) based on their expertise would be called an argumentum ad verecundiam (an argument from authority), a fallacious inference, it is indeed more accessible for individuals to assess the expertise of the scientists than to assess the veracity and scrutiny of the scholarship itself. Hence, establishing epistemic trust in AI systems could help overcome the burden of understanding the system.

Arguments similar to epistemic trust in AI systems also come from within the human-AI interaction community. Knowles and Richards [KR21] established the concept of public trust in AI. In doing so, they differentiate between trust in a specific, discrete, and identifiable AI from trust in AI as an abstraction, which they call trust in AI-as-an-institution.

Central, here, is the argument that "individuals do not develop trust in [AI] systems through careful and ongoing assessment of their trustworthiness; instead, one trusts that the system itself has appropriate mechanisms for ensuring trustworthiness" [KR21]. Knowles and Richards also make clear that the ensuring instances are not the developers of the AI systems but the broader ecosystem that determines the trustworthiness rules developers must follow. In other words, Knowles and Richards suggest that users develop epistemic trust in the ecosystem to ensure the trustworthiness of AI systems.

In their model of public trust, Knowles and Richards [KR21] also suggest a four-step process to reach public trust in AI, starting with (1) defining trustworthiness, followed by (2) specifying trustworthiness, (3) enforcing trustworthiness, and (4) reaching trustworthy AI.

In their model, the matter of trust calibration is taken over by the ecosystem, ensuring that AI development and outcomes are inherently trustworthy. However, how would an ecosystem communicate the trustworthiness of AI? One answer, included by Knowles and Richards [KR21] in the fourth step of their model, is by providing certifications which we discuss in the next section.

AI SEALS OF TRUST: THEORETICAL AND EMPIRICAL CONSIDERATIONS Certifications such as AI seals of trust generally "refer to a process in which a company's processes and services [here: AI] are evaluated against a predefined set of criteria via an audit by a third party, which formally acknowledges that the standard defined by the criteria is met" ([Lan+19], p. 4).

As such, certifications aim to reduce complexity and uncertainties about systems and make it easy for users to identify what is (not) trustworthy. To that end, certifications have been discussed and introduced in various contexts, such as cybersecurity, web assurances in e-commerce, or cloud services. For the context of AI, the EU AI Act suggests certification as a central mechanism to communicate compliance with industry and legislative requirements to the public (see Article 44 in Chapter 5 "Standards, Conformity Assessment, Certificates, Registration" 2).

To introduce seals of trust to the field, it is crucial to consider the effectiveness of such measures. Theoretically, arguments supporting seals of trust have previously predominantly been grounded in (1) trust theory, (2) signaling theory, and (3) persuasion literature, in particular, the elaboration likelihood model (ELM).

From the perspective of trust theory, seals of trust communicate to users through trust-assuring arguments that a system can fulfill the specific requirements laid out in the contract between trustor and trustee. In doing so, in trust theory, seals of trust become part of an institutionalized mechanism that ensures trust. In signaling theory, the main focus is on the communication process of one party to the other. Central here is the assumption of an information asymmetry wherein one party is less informed (the trustor) than the other (the trustee). Providing information in the form of seals of trust "are signals which are actions that parties take to reveal their true type" [KR00].

In contrast to trust theory and signaling theory, the ELM is more explicit in how seals are perceived. At its core, the ELM describes how individuals process persuasive arguments by following either a peripheral route of processing which requires less cognitive effort, or a central, more effortful route of information processing. Theoretically, seals of trust function as cues that can effortlessly be processed via the peripheral route. However, processing via the central route is also possible when seals of trust induce deeper elaboration [Low+12].

While all three theoretical approaches assume positive effects of seals of trust, empirically, previous scholarship has been inconclusive. On the one hand, some authors have found no effects. For

example, McKnight et al. [MKC04] found no effects of, what they called, privacy assurance and industry endorsement seals on trust in web business. The authors explain their results, suggesting that participants either did not notice the seal or did not know what it was supposed to signal. Similar results were obtained by Kim et al. [KFR08], who found no effect of seals on trust but also pointed to a lack of understanding and familiarity with the seal's meaning.

On the other hand, in a more recent study, Kim et al. [Kim+16] (2016) found that Web Assurance Seal Services (WASS) were effective instruments to increase users' trust and mitigate their concerns about e-commerce platforms. Moreover, results for the positive effects of seals on trust in the context of e-commerce are supported by findings from Mavlanova et al. [MBL16].

In doing so, the authors differentiated between internal (company's certification) and external (third-party certifications) signals. Their results indicate that, although both signals increased trust, only external signals also increased the perceived quality of the seller. Joining results against and in favor of seals of trust, Adam et al. [Ada+20] introduce the trust tipping point. Examining the effectiveness of seals of trust in the context of online websites, the authors found that below a certain trustworthiness threshold, seals effectively increased users' trust. However, with raising trustworthiness, the seals could not increase users' trust further.

Concluding from previous empirical findings, we know that seals of trust can effectively increase trust. However, the effectiveness might be reduced when (a) users do not notice the seals of trust, (b) users do not know the function of the seal of trust, (c) the seal of trust is granted internally, and (d) user trust is already at a high level.

11.3 THE PRESENT STUDY

Based on the theoretical and empirical findings elaborated above, for this study, we assume that:

> **Hypothesis (H1):** An AI system with an AI seal of trust is perceived as more trustworthy than an AI system without an AI seal of trust.

Moreover, we are also interested in how a seal of trust would affect each trust dimension (performance, process, and purpose). However, empirical differentiations between the three trust dimensions are rare. Hence, we did not formulate a directional hypothesis but instead posed the following research question:

> **Research Question (RQ):** How does a seal affect the three trust dimensions (performance, process, and purpose)?

Going beyond the mere presence (or absence) of a seal, we are also interested in the specific content of such a seal. What exactly should be certified? As it stands, trustworthy AI can refer to various aspects. While we hypothesize that any seal of trust would help to increase the users' trust perceptions (see H1), we also assume differences between different seals (H2), relating to how familiar users are with the seals' content (H3a) and how well users understand what the seal certifies (H3b). More formally stated, we hypothesize:

> H2: The three trust seals differ in their perceived trustworthiness, with certification of training data receiving the highest trust, followed by certification of transparency and certification through formal verification.

> H3a: The seals' perceived trustworthiness partly depends on the perceived familiarity with the seals' content. The more familiar users are with the content of the seal, the higher the perceived trustworthiness of the seal.

> H3b: The seals' perceived trustworthiness partly depends on the perceived understanding of users of the seals' content. The more intelligible seals are for users, the higher the perceived trustworthiness of the seal.

In addition, as the literature reviewed above suggests, trust in the certifying body will also affect how a seal is perceived. Hence, we assume:

> **H4**: The seals' perceived trustworthiness partly depends on the perceived trustworthiness of the certifying body. The higher the perceived trustworthiness of the certifying body, the higher the perceived trustworthiness of the seal.

Because the literature on the possible effects of AI seals of trust is scarce, we also included a more explorative approach to better understand users' needs and expectations. Hence, in addition to the directional hypotheses, we included a qualitative part in which we asked participants to elaborate on which aspects of AI systems should be certified through an AI seal of trust.

11.3.1 *Method*

The study received ethical approval from the ethics committee of the University of Duisburg-Essen. All hypotheses and analyses were pre-registered via OSF-Open Science Framework.

11.3.2 *Sample and Study Design*

To test our hypotheses and research question, we conducted an online study with a between- group design. To that end, we collected data from N = 453 participants who were randomly assigned to one of four conditions. The sample consisted of 220 females, 218 males, 12 nonbinary, and three participants who preferred not to disclose their gender identity.

All participants were recruited via the crowd-sourcing platform Prolific. Participants' mean age was 37.94 (SD = 12.69) and ranged from 18 to 80 years. The highest degree for two participants was a middle school degree, for 184 a high school degree, for 194 a Bachelor's degree, for 48 a Master's degree, for four a PhD, and 21 indicated to have received another degree.

11.3.3 *Manipulated Variable: The AI Seal of Trust*

The four experimental conditions reflected the different trust seals, in addition to a control group. To that end, we selected three certifications which correspond to archetypical levels of insight into the inner workings of AI systems: (1) The quality of the training data (n = 114)—that is, even if the AI system is a black box, certifications based on the input (i.e., training data) may assist in assessing the system's trustworthiness, (2) the transparency (e.g., explainability) of the AI system (n = 114)—as it relates the input and output of a black box approximate system behavior, and (3) the formal verification of a AI system (n = 113)—as it guarantees desirable behavior of the system by white-boxing it. In addition to these different certifications, we included one control group (n = 113), which did not receive any seal of trust.

In addition to a brief description about the respective trust seal (all detailed descriptions can be found in the online supplementary material C), participants saw an image of a seal (see Figure 11.1). Because the design of a seal likely affects the end-users' trustworthiness perceptions, we reduced this effect by adding the following statement to the visual representation of the seal: "Please be aware

that due to copyright reasons, we cannot represent the actual seal. The representation you see here is just a placeholder for this study."

Figure 11.1: Visualization of the AI Trust Seal that participants saw in the Study

11.3.4 *Procedure*

After agreeing to the informed consent, participants were introduced to a working definition of AI (see the online supplementary material A for details). We included this information to ensure that all participants understood the terminology similarly. Afterward, participants of the experimental groups were introduced to the concept of AI seals of trust with the following text:

> "Artificial intelligence (AI) is recognized as a strategically important technology that can contribute to a wide array of societal and economic benefits. However, it is also a technology that may present serious risks, challenges, and unintended consequences. Within this context, trust in AI systems is necessary for the broader use of these technologies in society. It is, therefore vital that AI-enabled products and services are developed and implemented responsibly, safely, and ethically. But how to know whether one can trust AI? One way to make this trust judgment easier for users are so-called AI seals of trust. Such AI seals of trust are granted by independent and neutral intermediaries who assess whether AI fulfills trustworthiness standards. Similar to food certifications and labels, these AI seals signal to users the state of an AI."

Next, participants saw the different seals of trust and were introduced to different AI systems certified with AI seals of trust. Participants of the control group were directly introduced to the AI system and did not view information on the seals of trust. After

viewing the AI systems, participants were asked to answer several questions about one of these AI systems. Before closing the study with a manipulation check and the debriefing, participants were informed about all three possible seals of trust, after which, in an open question, participants were asked to indicate which of the three seals they found most important (ranking question), and what they expect from an AI seal of trust.

11.3.5 *Stimulus Material*

Participants read short descriptions of four different AI systems and their functionalities. While modeled after real-world applications to avoid prior exposure effects, all systems were hypothetical and did not exist. The systems were: (1) CheckMySkin, a mobile application to check for skin cancer, (2) Drive Tek, an autonomous driving system, (3) Sound Shuffle, a music recommendation system, and (4) FindYou, a hiring system. The texts participants read can be found in the online supplementary material B.

To increase the generalizability of our results, half of the participants answered questions about the system CheckMySkin, whereas the other half answered questions about the system Drive Tek. Participants in the experimental groups saw both of these systems alongside an AI seal of trust. For the analysis, both conditions were joined. Moreover, to increase external validity, we added two additional systems, Sound Shuffle and FindYou, which were always presented without an accompanying seal of trust. Hence, all participants of the experimental groups saw two systems with and two systems without seals of trust, whereas participants of the control group only saw systems without seals of trust.

11.3.6 *Measured Variables*

All of the following measures were assessed on a 5-point Likert scale, ranging from 1 = "strongly disagree" to 5 = "strongly agree." For subsequent analyses, items of all measures were summarized to a final mean score.

Trust in a system. Because we wanted to assess trust as thoroughly as possible, we combined items from different scales to measure the three dimensions of trust (performance, process, and purpose) and mistrust.

The final measure included 15 items to measure the perceived performance of a system (Cronbach's $\alpha = 0.96$), 13 items to measure the perceived process (Cronbach's $\alpha = 0.9$), 10 items to mea-

sure the purpose of the system (Cronbach's $\alpha = 0.87$), and 12 items to measure mistrust (Cronbach's $\alpha = 0.94$). All items used to measure the trust dimensions and a supporting exploratory factor analysis can be found in the online supplementary material F.

Perceived familiarity and perceived understanding We used a three-item measure, adapted from Gefen [Gefoo], to assess the participants' perceived familiarity with a seal's content. The items were "I am familiar with the concept of [. . .]," "I have heard about the possibility to make AI systems better by controlling [. . .]," and "Media often report about controlling [. . .]." Depending on the group participants were allocated to, the blanks were filled by "the training data," "the concept of transparency," or "the concept of formal verification." For the analyses, all items were summarized in one mean score with Cronbach's $\alpha = 0.91$.

The construct perceived understanding was assessed through the following four items, which were developed following Ngo and Krämer (2022b): "I understand what the seal of trust means," "It is clear to me what the seal certifies," "I could explain in my own words what the certification does," and "I am uncertain about the meaning of the seal." For the analyses, all items were summarized in one mean score with Cronbach's $\alpha = 0.88$. Both constructs, perceived familiarity and perceived understanding were not assessed by participants of the control group who did not view a seal of trust.

Trust in the certifying body. Trust in the certifying body was assessed through seven items from corporate credibility scale of Newell and Goldsmith (2001). For the analyses, all items were summarized in one mean score with Cronbach's $\alpha = 0.95$.

Trust in artifical intelligence Because we did not want the individual's take on AI to interfere with our results, we also included individuals' attitudes toward AI as a covariate, using the ATAI scale of Sindermann et al. (2021), which includes five items on an 11-point Likert scale such as "I fear artificial intelligence" or "Artificial intelligence will benefit humankind." For the analyses, all items were summarized in one mean score with Cronbach's $\alpha = 0.78$.

11.4 QUALITATIVE CONTENT ANALYSIS

To better understand the participants' needs and expectations toward an AI seal of trust, we included a ranking question and an

open-ended question at the end of our online experiment. In the ranking question, having been introduced to all three possible seals of trust, we wanted to know which of the seals of trust participants found most important. To conclude, we asked:

> "Lastly, having seen now three possible AI seals of trust, we are curious whether you have your own opinion about what an AI seal of trust could certify. Below you have some space to let us know what you think would be important."

We analyzed all answers following Mayring's [May14] recommendations for qualitative content analysis (see results section for details).

11.5 RESULTS

All data can be accessed via OSF-Open Science Framework..

MANIPULATION CHECK A chi-squared test with the independent grouping variable trust seal and the dependent variable trust seal recall indicated that significantly more participants remembered correctly the seal they saw than those who did not remember correctly $\chi^2(12) = 747.05,\ p < .001$. In the control condition, 63.4% of participants remembered correctly (n = 71), in the training data condition, 67.5% (n = 77), in the transparency condition, 63.15% (n = 72), and in the formal verification, 79.6% (n = 90).

HYPOTHESES TESTING In the central hypothesis of this work (H1), we expected that participants trust an AI system certified with an AI seal of trust more than an AI system without certification. To determine the effect of a seal on the participants' trust, we conducted an ANCOVA with the trust score as the dependent variable and the four leveled factor AI seal of trust as the grouping variable. As the covariate, we controlled for participants' general trust in AI. The descriptive results of the variables trust and its subdimensions performance, process, and purpose, as well as mistrust grouped by the factor AI seal, can be found in Table 11.1.

Results of the ANCOVA indicate that there was no significant difference in the partici- pants' trust scores between the different groups, $F(3, 448) = 0.72,\ p = 0.54$. Moreover, we also had to reject H2 for which we expected that the training data seal would receive the most trust, followed by the transparency seal, and the formal verification seal.

While the result for H1 indicates that none of the three different seals of trust affected participants' trust perceptions, it could have

Table 11.1: Descriptive Results of the Dependent Variable Trust and Its Subdimensions by Experimental Group

		No Seal	Training Data	Transparency	Formal Proof
Trust	*M*	3.48	3.53	3.50	3.41
	SD	0.71	0.68	0.76	0.69
Performance	*M*	3.29	3.49	3.39	3.36
	SD	0.87	0.78	0.93	0.88
Process	*M*	3.22	3.24	3.22	3.08
	SD	0.85	0.82	0.94	0.87
Purpose	*M*	3.93	3.87	3.88	3.79
	SD	0.79	0.77	0.76	0.76
Mistrust	*M*	3.34	3.24	3.35	3.44
	SD	1.05	0.99	1.05	0.94

been the case that the seal affected only subdimensions of trust. For this possibility, we did not articulate a hypothesis but posed RQ1, asking whether the different seals affected the three subdimensions, performance, process and purpose differently.

In addition to the three subdimensions, we also included the measure for mistrust in RQ1 (note that mistrust was not included in the RQ in the preregistration). To assess RQ1, we conducted a MANCOVA with the subdimensions performance, process (integrity & transparency), and purpose, as well as mistrust as outcome variables and the four leveled factor AI seal of trust as the grouping variable. Similarly to testing H1, we also controlled for individual levels of trust in AI. Results indicate that the three subdimensions, as well as mistrust, were similarly affected by the trust seals, Pillai's trace $= .02, F(3, 448) = 1.09, p = .075$

Although we found no differences between the three seals of trust and participants' trust perceptions (see H1), an indirect effect of the seals on trust can still be expected. In H3a and H3b, we suggested that an effect of the seal is at least partly the result of the participants' perceived understanding of the seal's content and the participants' familiarity with the seal's content.

In addition, in H4, we anticipate that the effect of the seals might also be the result of the perceived trustworthiness of the institution which issued the seal. To understand these possible explaining mechanisms, we ran three separate mediation analyses with understanding, perceived familiarity, and perceived source trustworthiness as mediating variables. For this, we used the Process Macro version 4.3.1 for SPSS by Hayes [Hay17].

Furthermore, we used the variable AI seal of trust as the independent variable, which was dummy-coded. Participants who viewed the training data seal were entered as a reference category. Participants of the control group were excluded from the analyses as they did not answer questions about their understanding of the seal, their perceived familiarity, and the perceived trustworthiness of the source (see also the elaboration in the methods section).

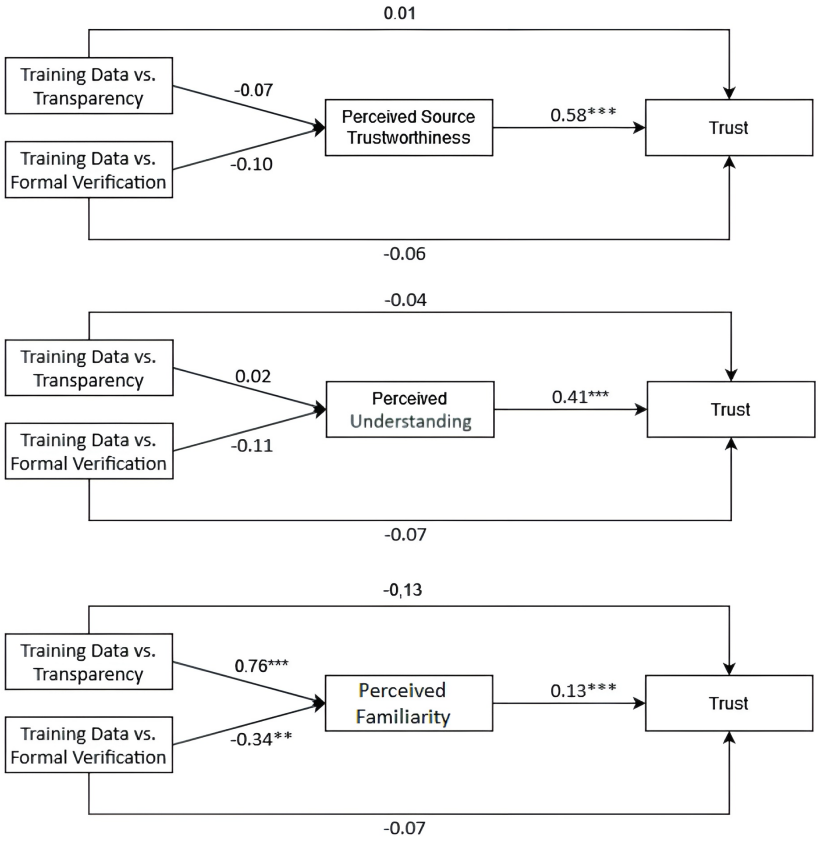
The outcome variable was again trust. We tested the significance of the effects using bootstrapping procedures, computing 5,000 bootstrapped samples with a confidence interval of 95%. All unstandardized path coefficients and significance levels can be found in Figure 11.2. The full results of the mediation analyses can be found in the online supplementary material D.

The mediation analyses revealed nonsignificant indirect effects for all three variables (understanding, source trustworthiness, and perceived familiarity). For understanding and source trustworthiness, the a-path was insignificant, indicating that the AI seal of trust participants viewed was neither related to the variable understanding nor source trustworthiness. However, the b-path was significant, indicating that both were very strong predictors of trust, with understanding explaining roughly 34% of the trust variance and source trustworthiness explaining roughly 72%. Not surprisingly, these results underline the importance of users understanding what a seal represents and the importance of the issuing source of the seal.

In contrast, we found a significant a-path for perceived familiarity, suggesting that participants were not equally familiar with all AI seals. In particular, we found that participants were more familiar with transparency than verified training data (positive coefficient) but were less familiar with formal verification than training data (negative coefficient). This result partly confirms what we anticipated in H2, suggesting that participants are not equally familiar with the different seal content. Beyond this, the significant b-path indicates that higher familiarity with a seal's content resulted in greater trust.

11.6 QUALITATIVE RESULTS

First, we asked participants to rank the three seals of trust they found most important. With one as the highest rank and three as the lowest, mean results indicate that participants found all three seals of trust similarly important, with formal verification scoring $M = 1.94$, transparency of the AI system $M = 1.94$, and



* $p < .05$, ** $p < .01$, ***$p < .001$

Figure 11.2: Visual Representation of Mediation Analyses with Unstandardized Path Coefficients

training data $M = 2.12$. While the mean ranks do not indicate a great difference between the three seals of trust, which reflects the results of our quantitative analysis, inspecting the absolute number that a seal was ranked first, we can see that participants found the formal verification and transparency of a system most important (see Figure 11.3).

Because the three trust seals we selected reflect our understanding of importance, we assessed the participants' answers with an open-ended question, asking what participants find most important in an AI seal of trust. We applied descriptive and in-vivo codes in the first coding cycle to capture the participants' answers (Saldaña, 2013). In the second step, all codes were abstracted and summarized into higher-level codes. Throughout both coding cycles, three independent coders worked on the answers. To ensure the quality of the final coding scheme, we calculated Cohen's Kappa on 25% of the answers. In the first round, all three coders reached an agreement of $K = .64$. To increase agreement, all three coders discussed and resolved cases of disagreement. Consequently, inter-rater relia-

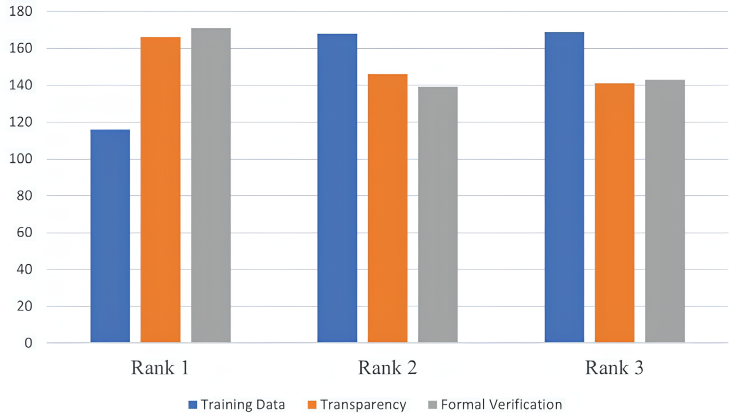


Figure 11.3: Absolute Numbers of the Ranking Data

bility increased to a sufficient $K = .82$ in a second round of coding on different sets of answers. In the following, we report the most important results of the qualitative content analysis. Overall, the final coding scheme identified seven different categories (see Table 11.4), which differ in the number of mentions as well as the level of abstraction (number of second-level codes).

While some participants voiced general support for trust seals, others rejected any certification as well as AI systems as a whole. For example, P84 stated, "nothing would really give me any trust in AI. I am very against the idea of anything AI." In addition to general mistrust in AI and certifications, participants also voiced concrete concerns about the seal-issuing institution. For example, P163 states, "I don't necessarily trust these seals of trust because they can always get bought". This reflects our quantitative results, which underline the importance of source trustworthiness. Moreover, the distrust voiced by our participants reminds of what Dietvorst et al. [DSM15] call algorithm aversion, a generally negative stance toward anything related to algorithms and AI.

Following trust literature, most participants, however, commented along the lines of the three trust dimensions are performance, purpose, and process, with performance-related comments being mentioned by far the most. Among those, most participants wanted a seal of trust to certify that the system does what it was set out to be (formal verification) and its safety.

Related to the issue of safety, a smaller group of participants also voiced the need for, what we call, a nondestructive AI seal of trust. For example, P136 stated that a seal "could certify that the AI can be trusted not to be evil and ruin mankind," and P389 who noted that a seal "could certify whether the AI's intentions are

1st level codes	#	Description	2nd level codes (#)
Trustworthiness	19	Verification of the general trustworthiness of a system without further specification	
Distrust	24	General distrust in the AI or the seals of trust	
Performance	206	Verification of AI's abilities and characteristics	(formal) Verification (60) Safety (55) Accuracy (23) Error-free (20) Re-evaluation (20) Extensive testing (20) Efficiency (8)
Process (transparency)	49	Related to how the AI operates and the intelligibility of its inner workings	
Purpose	66	Verification of the intentions of the AI's developers and the development process	Ethical compliance (13) Privacy (24) Training set quality (25) Copyright compliance (4)
Trustworthy AI seals of trust	36	Verification of the seal-issuing institution	Trustworthy origin of the seal (26) Transparency of the certification process (10)
Destructive AI	20	Verification that AI cannot develop its own agency and intentionally harm humans	

Figure 11.4: Results of the Qualitative Content Analysis

true—whether it wants to make humans safe or whether it wants to further its own goals regardless of our safety."

11.7 DISCUSSION

Through quantitative and qualitative data collections, in this work, we investigated the effect of an AI seal of trust on the users' trust assessments of AI systems as well as the users' expectations toward such seals respectively.

QUANTITATIVE RESULTS: ADDRESSING THE NULL EFFECT OF THE TRUST SEAL In a pre-registered online experiment, we tested three different seals of trust (certification of the training data, transparency, and formal verification) and their effects on user trust in an AI system. However, unlike hypothesized, none of the three different seals of trust could significantly increase our participants' trust in an AI system compared to a control group.

A more fine-grained analysis, differentiating trust into its sub-dimensions performance, process, and purpose, supported this null result. The seals of trust did not affect the trust dimensions differently compared to a control group. While previous results from different domains would suggest an effect of the certifica-

tion, this paper's null results echo previous null results. Examples include McKnight et al. [MKC04] and Kim et al. [Kim+16], who relate their null findings to users' not noticing the seal or users' limited understanding and familiarity of the seal's content.

We can rule out these explanations because we also assessed participants' understanding of and familiarity with a seal. In addition, the manipulation check indicated that participants remembered the respective trust seals. Instead, we suggest that our results relate to the findings of Adam et al. [Ada+20]. The authors suggest that if a system's trustworthiness is already high, an additional seal of trust cannot increase the trustworthiness any further. We find support for this speculation in the mean trust ratings of our study as we noticed that these fall within 3.41 and 3.53 points, significantly higher than the scale midpoint (2.5 points).

Following theoretical considerations of trust theory and signaling theory, an alternative explanation to the null results is that the trust seals did not signal the intended meaning. Indeed, our seals might not have communicated the trustworthiness of the systems because they are neither well established outside the experimental setting nor granted by a well-known institution (see also next section). Hence, they possibly lacked the epistemic authority to convince our participants.

Moreover, we found that the seals of trust were not perceived differently in terms of understandability but differed in familiarity, with transparency certification being the most well-known, followed by training data and formal verification certifications. Finding differences for familiarity but not understanding indicates that, while knowing of a specific certification method, this knowledge does not necessarily translate into understanding.

SUPPORT FOR EPISTEMIC TRUST We found that independent of which seal participants saw, the higher the participants' trust in the seal-issuing institution was, the higher was the trust in the AI system. In other words, if users trust the institution that grants the seal, trust in the system will increase. Consequently, this shifts the users' trust assessments from the system to the certifying institution. Hence, our result supports the idea of epistemic trust and trust in AI-as-an-institution [KR21]. It seems that it is easier for users to ask, "Whom to trust?" instead of attempting to understand AI systems.

Moreover, in line with predictions of the ELM, knowing a certifying institution might also function as a mental shortcut. Knowing that a certain institution is trustworthy, any communication orig-

inating from such an institution should also be trustworthy (see also, authority heuristic in Sundar [Sun+08]. For the present work, we could not rely on the authority of a specific institution as our seals might have been less effective because their origin was unknown to the participants. However, adding additional information such as a seal or a seal-issuing institution whose trustworthiness has to be assessed also comes with downsides discussed in the next section.

11.7.1 *Qualitative Results*

NEED FOR VERIFICATIONS WITHOUT UNDERSTANDING OF VERIFICATIONS In the qualitative part of this work, we asked participants to explain what they expect from AI certifications. Through a qualitative content analysis, we found that participant responses mainly fell within the three trust dimensions, performance, process, and purpose, with performance-related certification being mentioned the most. Among the performance category, participants indicated that (formal) verification, the certification that the system does what it was set out to be, was mentioned the most.

This is also supported by the ranking data that we collected. Here, formal verification was ranked first most of the time. However, in light of the quantitative results, which indicated that participants knew the least about formal verification compared to transparency and training data, the higher ranking of formal verifications is alarming. Participants found the greatest reassurance in something they understood the least and, in turn, maybe expected it to be most comprehensive and fail-safe. We speculate whether this might be due to participants having given up on other, more well-known methods.

SECOND-LEVEL TRUST CALIBRATIONS Interestingly, some participants mentioned the general need for a trustworthiness certification, whereas others voiced distrust toward any such certification and AI-related system. We relate these contradicting sentiments to what Wischnewski et al. [WKM23] define as second-level trust calibrations, where users have to perform an additional (second level) trust judgment (here: judging the trustworthiness of the seal) on top of the trust judgment concerning the AI system (first level), possibly increasing users' cognitive load. While following persuasion literature, which suggests that seals can reduce the users' cognitive load by offering trust cues, future studies should examine whether cognitive load can also be increased through the additional information that needs to be processed.

This is especially true in the context of calibrated trust. Suppose it is the aim that user trust is appropriately calibrated to the AI system's functionality. In that case, users must also find a way to calibrate their trust in the AI seal appropriately.

In addition, the distrust sentiment voiced by our participants also indicates the limits of approaching trust from an epistemic perspective. If the seal-issuing institution is not trusted, users will likely not trust the system. Hence, future studies should assess which cues make an AI seal of trust more trustworthy, and which user groups generally distrust AI.

11.8 LIMITATIONS AND FUTURE STUDIES

The strongest limitation to our study concerns its external validity. First, as currently no established, noncommercial certification body or trust seal exist, all material was hypothetical. Similarly, participants did not directly engage with the AI systems but read different vignettes.

Hence, we could not measure how participant trust translated into actual behavioural outcomes. Further, online data collection is limited for decisions in practice, as this problem type involves substantial cognitive effort that an online environment may not be able to replicate as well as decision-making often is a high-involvement task and online participants may not meet this criterion.

For future studies, we suggest integrating actual systems into the experimental setting. In addition, with AI systems based on large language models such as GPT-4 being commercialized, it could be interesting, for example, to include such a conversational interface and interactivity in general. Moreover, as participants likely did not know about AI seals of trust, we had to provide a definition of such. While we tried to be as subtle as possible, describing AI systems as "a technology that may present serious risks, challenges, and unintended consequences" (see Method section), we potentially biased participants to be more critical and vigilant than they initially were, raising participants' overall skepticism toward the presented system.

However, as we can see in the overall trust ratings across conditions, participants perceived the systems as relatively trustworthy (mean trust ratings > 3.41 points at a scale midpoint of 2.5 points). In addition, we statistically controlled for participants' general attitudes toward AI by including individuals' attitudes as a covariate in our analyses. Hence, even if a subgroup of users was affected

by our definition, it should not have changed our results. Lastly, as we suggested in the previous section, we speculate that our null results are related to all AI systems being equally trustworthy. To test this interpretation, future studies should experimentally vary the trustworthiness of AI systems by, for example, comparing different levels of system reliability (high vs. low) to investigate whether trust seals can increase the users' trust.

11.9 CONCLUSION

In this work, we investigated the effects of AI certifications, so-called AI seals of trust, on the users' trust in AI systems. We tested three certifications and their effects on global trust and the trust subdimensions performance, process, and purpose. Unlike hypothesized, we found that the trust seals did not affect users' trust in the AI system. Examining possible underlying mechanisms, we found that a higher understanding of the seal's content as well as familiarity with the seal's content, could increase users' trust.

Moreover, we found evidence of epistemic trust. That is, the more participants trusted the seal-issuing institution, the more they trusted the AI system. However, our qualitative results also indicated that some participants reject the idea of an AI seal of trust as they do not trust AI systems or any certifying party. Nevertheless, most participants said they would like to see a system's functionality be certified, specifically, its performance and safety.

Part IV

SUMMARY

12

SUMMARY AND FUTURE WORK

In this thesis, we both aim to extend methods and applications as well as tackle central points of interest in the research realm: Some fundamental challenges——such as understanding the general conditions under which adversarial attacks transfer——have been approached from multiple perspectives yet remain unresolved. A core goal of this thesis was to bridge research areas that are typically treated in isolation and to shed new light on such open questions. For all of our presented papers arise new questions which can be extended upon in the future. The thesis can be subdivided into the following fields: Explainability of uncertainties, robustness and uncertainty attacks, and trustworthiness of AI systems. Our work spans both empirical proxies, such as the local intrinsic dimensionality, as well as interdisciplinary work using psychological studies. In our interdisciplinary context, we highlight the role of uncertainty quantification for trust calibration and propose a characterization of trust.

In Part I, we focus on explainable AI: We start by presenting a novel unsupervised visualization method of both uncertainties in the form of confidence scores of a network and simultaneously a proxy for robustness of high-dimensional data in Chapter 3. The goal was to be able to build warranted trust in machine learners: We neither wish for humans to over- nor undertrust learners, and it is commonly known that the probability associated with the predicted class label does not always reflect its ground truth correctness likelihood [Guo+17].

As Guo et al. argue [Guo+17], models should signal when they are likely to be incorrect. This intuition is pursued by us in our published work Unsupervised DeepView [NM22a]. The challenge we faced here lay in designing a method capable of quantifying uncertainties in an unsupervised learning setting. In a supervised setting, visualizing the classification boundaries can be done using dimensionality reduction techniques [SHH20].

Our contribution was to employ the unique proxy of local intrinsic dimensionality to extend the visualization of classification boundaries to certain and uncertain regions instead of directly mapping to the class labels. A key limitation of Unsupervised DeepView [NM22a] is the need for confidence scores as output of

a network. We visualize confidence as well as the likelihood of an adversarial attack using the proxy of local intrinsic dimensionality [Ma+18].This implies that there is no intuitive implementation for architectures such as autoencoders, as they lack predictive uncertainty [YB22]. Although it is possible to include frameworks that estimate aleatoric and epistemic uncertainty for Bayesian autoencoders [YB22; Abd+21a], variational autoencoders [Abd+21a], or other specific approaches, what we would like to test in the future is whether our intuitive definition of a score function correlating with the likelihood of adversarial attacks can be extended without the additional need of prior confidence scores by using verification techniques.

Regarding *future work*, we would consider using formal quality guarantees of autoencoders [Böi+21] as an additional estimate for our quantifiable unsupervised uncertainty. This would be only one means of extending the method's capabilities: The focus of the method currently lies in its ability to capture a global overview; however, it is implemented such that local instances are still visible. This shows potential for combining local uncertainty visualization schemes with the global overview. Examples could be an extension of CLUE [Ant+20a].

Unsupervised Deepview is model-agnostic and can therefore also be applied to Bayesian neural networks. CLUE tackles the question "What is the smallest change that could be made to an input, while keeping it in distribution, so that our model becomes certain in its decision for said input?". This could be extended to the question "What is the smallest change hat could be made to an input, while keeping it in distribution, so that our model becomes certain or the local intrinsic dimensionality of the input sample becomes an outlier in its neighborhood?". With this add-on, we would again visualize our proposed unsupervised quantifiable uncertainty estimate.

We then further investigate on a more general scale the impact of ambiguous data on learners and the limitations of pixel-based explainable AI in Chapter 4, based on our work "Do you see what I see? An Ambiguous Optical Illusion Dataset exposing limitations of explainable AI" [New+25b]. This study aims to bridge the gap between technically rigorous explanations and the concept of "human interpretability". We highlight the challenges of combining both perspectives meaningfully and discuss their limitations in detail.

The dataset, Ambivision, points out examples of ambiguity: If two animals share characteristics, how can we meaningfully distinguish the two? As human brains interpret our visual data in a context-dependent manner [Bie87], perhaps a promising direction for the domain of automatic object detection will lie in including general concepts such as gaze direction, rather than only the inclusion of pixel-based features. The same can be utilized in the explainable AI domain: If in those specific cases, explainable AI is currently unable to provide human-interpretable explanations, future XAI methods should focus on true concept-based explanations rather than segmenting and grouping pixels [New+25b].

It remains an open problem how to design a specific scheme that fulfills these criteria: After all, current automatic concept-based extraction techniques also essentially rely on pixel-based segmentation and clustering, such as ACE [Gho+19], CONE-SHAP [Li+21a] or EAC [Sun+23]. Other popular concept-based XAI algorithms rely on human-annotated concepts instead [Kim+18; FV18; Yeh+20]. We incorporate human knowledge, but this does not address the more general open problem visible:How do we automatically construct concept-based explanations that go beyond pixel-level segmentation?

In *future work*, we would like to address this particular research gap by attempting to detect actual concepts, such as viewing direction, only automatically. Furthermore, the dataset could be used as an estimator for future explainable AI methods to measure their performance in purposefully ambiguous settings by testing the human interpretability of the extracted information. After all, the dataset highlights current limitations where prevalent methods fail [New+25b].

In Part II of this thesis, we dive deeper into the robustness of neural networks. Building on the concept of attack likelihood introduced in Part I, we explore adversarial vulnerability in greater depth. In Chapter 7, we extend our discussion by introducing an attack method that takes into consideration not only common metrics such as accuracy and distance [GSS14; WX18], but also other important parameters such as certainty and transferability. In doing so, our objective is to show that the metrics commonly used for detection and mitigation purposes [SG18] are vulnerable to this kind of attack setup.

We therefore aim to push the arms race discussion between new attacks and their detection strategies into consideration of additional metrics, especially those important for detection and mitigation attempts. Future adversarial attacks should take this

into account when designing new attack angles. Another direction could lie in attacking not just confidence scores of a network, but rather trying to optimize adversarial attacks with regard to other epistemic and aleatoric uncertainties captured in a less naive manner.

Transferability is an interesting phenomenon in machine learning: Attacks can transfer from one model to another, even when learners are trained on different datasets or using different model architectures, and it remains an open problem and is poorly understood [WZT+18]. In Chapter 6, we therefore tackle the correlation between topological similarities of datasets and the transferability of attacks. We expected that topologically distant datasets were less likely to transfer, but surprisingly, this was not the case. This opens up room for discussion.

Previous research suggested that there should be some global shared features present [Xie+19]. We were not able to find those with our suggested proxies, and while our empirical evaluation has its limitations, we still have to ask ourselves: Are we really looking into the right direction in this research field? Are there really shared features present that we are simply unable to detect as of now, or does the nature of adversarial attacks stem from completely different factors? If topological data analysis is not the correct tool for depicting similarities that lead to transferability, then what is? Promising new research areas to be considered include manifold learning [Ize12] or algebraic geometry [Har13].

In *future work*, this research question can be expanded using verification estimates for certain perturbation sizes, or we might have to go for entirely new mathematical proxies that represent this more proficiently. In summary, research on robustness provides endless opportunities: What other proxies can we test that might correlate with adversarial transferability? What could capture adversarial subspaces in a more reliable manner other than local intrinsic dimensionality, a factor that relies solely on the dataset? How do we include model factors? All of those questions remain open for future work.

In Part III of this thesis, we continue by defining a trust framework that explicitly incorporates uncertainty, as presented in Chapter 9. This chapter interconnects the core themes of robustness, explainable AI, and uncertainty quantification into a unified framework for trust. Chapter 9 motivates the need of our additional interdisciplinary work and serves as an introduction to the connection of trust in AI systems: In Part I of our thesis, we discuss the need for

human interpretability of explainable AI. However, testing the real interpretability of an AI scheme is often not performed rigorously or without a meaningful sample size. This lack of empirical validation can result in misleading conclusions about interpretability.

Analyzing our own proposed scheme Unsupervised DeepView in Chapter 10, we found that sometimes less information is more comprehensible. In our particular example, the background of the explanation, while depicting an interesting estimation of the uncertainty, was really only confusing to users and lead to less interpretability. This finding is not entirely new [Ten+19], but we were able to crystallize that it is not often foreseeable which part of the scheme is still useful and what is redundant without testing interpretability using psychological studies [New+25a].

In addition to defining this thesis' grounding in trust and evaluating the interpretability of our proposed explainable AI method, we also collaborated on a study led by Magdalena Wischnewski[Wis+24a], investigating the trust certifications through trust seals and their effect on perceived trust. We found that seals do not affect user's perceived trust, but this variable is influenced if the issuing institution is deemed trustworthy by the user. Many critical questions still lack a unified perspective between psychologists and computer scientists, as knowledge from their respective subfields has not yet been fully integrated.

Possible new problem statements include trust calibration for future teachers in order to improve the ability to handle the flood of new upcoming methods and research in that domain. Both overtrust and undertrust pose significant risks, especially for younger users unaware of the dangers that come with artificial intelligence. In conclusion, due to the rapid technological progress in machine learning there are many different promising research directions and interesting problem statements left to tackle .

BIBLIOGRAPHY

[Abd+22] Moloud Abdar, Abbas Khosravi, Sheikh Mohammed Shariful Islam, U Rajendra Acharya, and Athanasios V Vasilakos. "The need for quantification of uncertainty in artificial intelligence for clinical data analysis: increasing the level of trust in the decision-making process." In: *IEEE Systems, Man, and Cybernetics Magazine* 8.3 (2022), pp. 28–40.

[Abd+21a] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." In: *Information fusion* 76 (2021), pp. 243–297.

[Abd+21b] Moloud Abdar, Maryam Samami, Sajjad Dehghani Mahmoodabad, Thang Doan, Bogdan Mazoure, Reza Hashemifesharaki, Li Liu, Abbas Khosravi, U Rajendra Acharya, Vladimir Makarenkov, et al. "Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning." In: *Computers in biology and medicine* 135 (2021), p. 104418.

[AB09] Margareta Ackerman and Shai Ben-David. "Clusterability: A theoretical study." In: *Artificial intelligence and statistics*. PMLR. 2009, pp. 1–8.

[Ada+20] Martin Adam, Lars Niehage, Sebastian Lins, Alexander Benlian, and Ali Sunyaev. "Stumbling over the trust tipping point–The effectiveness of web seals at different levels of website trustworthiness." In: *Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference, June 15-17, 2020*. 2020, p. 1.

[AA17] Charu C Aggarwal and Charu C Aggarwal. *An introduction to outlier analysis*. Springer, 2017.

[AR14] Charu C Aggarwal and Chandan K Reddy. "Data clustering." In: *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra* (2014).

[AY01] Charu C Aggarwal and Philip S Yu. "Outlier detection for high dimensional data." In: *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*. 2001, pp. 37–46.

[Ams+15] Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E Houle, Ken-ichi Kawarabayashi, and Michael Nett. "Estimating local intrinsic dimensionality." In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, pp. 29–38.

[ALG19] Cem Anil, James Lucas, and Roger Grosse. "Sorting out Lipschitz function approximation." In: *International conference on machine learning*. PMLR. 2019, pp. 291–301.

[Anj+19] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. "Explainable agents and robots: Results from a systematic literature review." In: *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems. 2019, pp. 1078–1088.

[Ant+20a] Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. "Getting a clue: A method for explaining uncertainty estimates." In: *arXiv preprint arXiv:2006.06848* (2020).

[Ant+20b] Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. "Getting a clue: A method for explaining uncertainty estimates." In: *arXiv preprint arXiv:2006.06848* (2020).

[Ara+20] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. "In AI we trust? Perceptions about automated decision-making by artificial intelligence." In: *AI & society* 35.3 (2020), pp. 611–623.

[Arr+20] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." In: *Information fusion* 58 (2020), pp. 82–115.

[Ath+18] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. "Synthesizing robust adversarial examples." In: *International conference on machine learning*. PMLR. 2018, pp. 284–293.

[Aza22] Amos Azaria. "ChatGPT usage and limitations." In: (2022).

[Bâc+17] Mihai Bâce, Philippe Schlattner, Vincent Becker, and Gábor Sörös. "Facilitating Object Detection and Recognition through Eye Gaze." In: *19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI 2017)*. ETH Zurich. 2017.

[Bai86] Annette Baier. "Trust and antitrust." In: *ethics* 96.2 (1986), pp. 231–260.

[Bai+11] Wilma A Bainbridge, Justin W Hart, Elizabeth S Kim, and Brian Scassellati. "The benefits of interactions with physically present robots over video-displayed agents." In: *International Journal of Social Robotics* 3 (2011), pp. 41–52.

[Bar04] Moshe Bar. "Visual objects in context." In: *Nature Reviews Neuroscience* 5.8 (2004), pp. 617–629.

[BBK19] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. "The need for uncertainty quantification in machine-assisted medical decision making." In: *Nature Machine Intelligence* 1.1 (2019), pp. 20–23.

[Ber+22] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. "How cognitive biases affect XAI-assisted decision-making: A systematic review." In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022, pp. 78–91.

[BWM20] Umang Bhatt, Adrian Weller, and José MF Moura. "Evaluating and aggregating feature-based model explanations." In: *arXiv preprint arXiv:2005.00631* (2020).

[Bie87] Irving Biederman. "Recognition-by-components: a theory of human image understanding." In: *Psychological review* 94.2 (1987), p. 115.

[BR18] Battista Biggio and Fabio Roli. "Wild patterns: Ten years after the rise of adversarial machine learning." In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 2154–2156.

[Bir+19] Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Luc Rey-Bellet, and Jie Wang. "Distributional robustness and uncertainty quantification for rare events." In: *arXiv preprint arXiv:1911.09580* (2019).

[Bit+20] Olga Vl Bitkina, Heejin Jeong, Byung Cheol Lee, Jangwoon Park, Jaehyun Park, and Hyun K Kim. "Perceived trust in artificial intelligence technologies: A preliminary study." In: *Human Factors and Ergonomics in Manufacturing & Service Industries* 30.4 (2020), pp. 282–290.

[Blö21] Bernd Blöbaum. "Some thoughts on the nature of trust: Concept, models and theory." In: *Trust and communication: Findings and implications of trust research*. Springer, 2021, pp. 3–28.

[Blu20] Henry Blumberg. "Hausdorff's grundzüge der mengenlehre." In: (1920).

[Böi+21] Benedikt Böing, Rajarshi Roy, Emmanuel Müller, and Daniel Neider. "Quality guarantees for autoencoders via unsupervised adversarial attacks." In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part II*. Springer. 2021, pp. 206–222.

[Bon+22] Andrea Bontempelli, Stefano Teso, Katya Tentori, Fausto Giunchiglia, and Andrea Passerini. "Concept-level debugging of part-prototype networks." In: *arXiv preprint arXiv:2205.15769* (2022).

[BK05] Karsten M Borgwardt and Hans-Peter Kriegel. "Shortest-path kernels on graphs." In: *Fifth IEEE international conference on data mining (ICDM'05)*. IEEE. 2005, 8–pp.

[BT19] Inka Bormann and Barbara Thies. "Trust and trusting practices during transition to higher education: Introducing a framework of habitual trust." In: *Educational Research* 61.2 (2019), pp. 161–180.

[BG21] Rainer Bromme and Lukas Gierth. "Rationality and the public understanding of science." In: (2021).

[Bro+11] Alexander M Bronstein, Michael M Bronstein, Leonidas J Guibas, and Maks Ovsjanikov. "Shape google: Geometric words and expressions for invariant shape retrieval." In: *ACM Transactions on Graphics (TOG)* 30.1 (2011), pp. 1–20.

[BS98] Horst Bunke and Kim Shearer. "A graph distance metric based on the maximal common subgraph." In: *Pattern recognition letters* 19.3-4 (1998), pp. 255–259.

[BR91] Peter J Burke and Donald C Reitzes. "An identity theory approach to commitment." In: *Social psychology quarterly* (1991), pp. 239–251.

[CW17a] Nicholas Carlini and David Wagner. "Adversarial examples are not easily detected: Bypassing ten detection methods." In: *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 2017, pp. 3–14.

[CW17b] Nicholas Carlini and David Wagner. "Towards evaluating the robustness of neural networks." In: *2017 ieee symposium on security and privacy (sp)*. Ieee. 2017, pp. 39–57.

[CW18] Nicholas Carlini and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." In: *2018 IEEE security and privacy workshops (SPW)*. IEEE. 2018, pp. 1–7.

[Car09] Gunnar Carlsson. "Topology and data." In: *Bulletin of the American Mathematical Society* 46.2 (2009), pp. 255–308.

[CG20] Gunnar Carlsson and Rickard Brüel Gabrielsson. "Topological approaches to deep learning." In: *Topological Data Analysis: The Abel Symposium 2018*. Springer. 2020, pp. 119–146.

[Car+08] Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. "On the local behavior of spaces of natural images." In: *International journal of computer vision* 76 (2008), pp. 1–12.

[CMB19] Giuseppe Casalicchio, Christoph Molnar, and Bernd Bischl. "Visualizing the feature importance for black box models." In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer. 2019, pp. 655–670.

[CD14] Tianfeng Chai and Roland R Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)." In: *Geoscientific model development discussions* 7.1 (2014), pp. 1525–1534.

[Cha+25] Tapabrata Chakraborti, Christopher RS Banerji, Ariane Marandon, Vicky Hellon, Robin Mitra, Brieuc Lehmann, Leandra Bräuninger, Sarah McGough, Cagatay Turkay, Alejandro F Frangi, et al. "Personalized uncertainty quantification in artificial intelligence." In: *Nature Machine Intelligence* 7.4 (2025), pp. 522–530.

[Cha+21] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. "A survey on adversarial attacks and defences." In: *CAAI Transactions on Intelligence Technology* 6.1 (2021), pp. 25–45.

[Cha+20a] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. "Concise explanations of neural networks using adversarial training." In: *International conference on machine learning*. PMLR. 2020, pp. 1383–1391.

[Cha+24] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. "A survey on evaluation of large language models." In: *ACM Transactions on Intelligent Systems and Technology* 15.3 (2024), pp. 1–45.

[Cha+20b] Angelos Chatzimparmpas, Rafael Messias Martins, Ilir Jusufi, Kostiantyn Kucher, Fabrice Rossi, and Andreas Kerren. "The state of the art in enhancing trust in machine learning models with the use of visualizations." In: *Computer Graphics Forum*. Vol. 39. 3. Wiley Online Library. 2020, pp. 713–756.

[Che+19] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. "This looks like that: deep learning for interpretable image recognition." In: *Advances in neural information processing systems* 32 (2019).

[CW21] Yi-Ning Katherine Chen and Chia-Ho Ryan Wen. "Impacts of attitudes toward government and corporations on public trust in artificial intelligence." In: *Communication Studies* 72.1 (2021), pp. 115–131.

[CRA] Ng Shin Cheng, Nabilah Filzah Mohd Radzuan, and Mohd Norshahriel Abd Rani. "Movie Recommender System using Supervised Learning Techniques." In: ().

[Cho19] D Choi. "On empirical comparisons of optimizers for deep learning." In: *arXiv preprint arXiv:1910.05446* (2019).

[CDR23a] Hyesun Choung, Prabu David, and Arun Ross. "Trust and ethics in AI." In: *Ai & Society* 38.2 (2023), pp. 733–745.

[CDR23b] Hyesun Choung, Prabu David, and Arun Ross. "Trust in AI and its role in the acceptance of AI technologies." In: *International Journal of Human–Computer Interaction* 39.9 (2023), pp. 1727–1739.

[Coa92] Cecil Anthony John Coady. *Testimony: A philosophical study*. Clarendon Press, 1992.

[CBH23] Joseph Cohen, Eunshin Byon, and Xun Huan. "To trust or not: Towards efficient uncertainty quantification for stochastic shapley explanations." In: *Phm society asia-pacific conference*. Vol. 4. 1. 2023.

[Cor+00] Ulises Cortès, Miquel Sànchez-Marrè, Luigi Ceccaroni, Ignasi R-Roda, and Manel Poch. "Artificial intelligence and environmental decision support systems." In: *Applied intelligence* 13 (2000), pp. 77–91.

[Cou+12] National Research Council, Division on Engineering, Physical Sciences, Board on Mathematical Sciences, Their Applications, Committee on Mathematical Foundations of Verification, and Uncertainty Quantification. *Assessing the reliability of complex models: mathematical and statistical foundations of verification, validation, and uncertainty quantification*. National Academies Press, 2012.

[Cou08] Council of Europe Centre of Expertise for Multilevel Governance. *12 Principles of Good Democratic Governance*. `https://www.coe.int/en/web/centre-of-expertise-for-multilevel-governance/12-principles`. Accessed 14 April 2025. 2008.

[CS22] Jonathan Crabbé and Mihaela van der Schaar. *Label-Free Explainability for Unsupervised Models*. 2022. arXiv: 2203.01928 [cs.LG].

[CB16] Antonia Creswell and Anil A Bharath. "Task specific adversarial cost function." In: *arXiv preprint arXiv:1609.08661* (2016).

[CCD08] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. "Supervised learning." In: *Machine learning techniques for multimedia: case studies on organization and retrieval*. Springer, 2008, pp. 21–49.

[Dal+04] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. "Adversarial classification." In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, pp. 99–108.

[DCP20] Xolani Dastile, Turgay Celik, and Moshe Potsane. "Statistical and machine learning models in credit scoring: A systematic literature survey." In: *Applied Soft Computing* 91 (2020), p. 106263.

[Dem+19] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks." In: *28th USENIX security symposium (USENIX security 19)*. 2019, pp. 321–338.

[Den12] Li Deng. "The mnist database of handwritten digit images for machine learning research [best of the web]." In: *IEEE signal processing magazine* 29.6 (2012), pp. 141–142.

[Deu+24] Jessica Deuschel, Andreas Foltyn, Karsten Roscher, and Stephan Scheele. "The role of uncertainty quantification for trustworthy AI." In: *Unlocking Artificial Intelligence: From Theory to Applications*. Springer, 2024, pp. 95–115.

[DK22] Jürgen Dieber and Sabrina Kirrane. "A novel model usability evaluation framework (MUsE) for explainable artificial intelligence." In: *Information Fusion* 81 (2022), pp. 143–153.

[DSM15] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. "Algorithm aversion: people erroneously avoid algorithms after seeing them err." In: *Journal of experimental psychology: General* 144.1 (2015), p. 114.

[Dik+18] Happiness Ugochi Dike, Yimin Zhou, Kranthi Kumar Deveerasetty, and Qingtian Wu. "Unsupervised learning based on artificial neural network: A review." In: *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*. IEEE. 2018, pp. 322–327.

[Dip+20] William K Diprose, Nicholas Buist, Ning Hua, Quentin Thurier, George Shand, and Reece Robinson. "Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator." In: *Journal of the American Medical Informatics Association* 27.4 (2020), pp. 592–600.

[Don+15] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. "Knowledge-based trust: Estimating the trustworthiness of web sources." In: *arXiv preprint arXiv:1502.03519* (2015).

[Don+17] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Xiaolin Hu, and Jun Zhu. "Discovering Adversarial Examples with Momentum." In: *CoRR* abs/1710.06081 (2017). arXiv: 1710.06081. URL: http://arxiv.org/abs/1710.06081.

[Don+18] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. "Boosting adversarial attacks with momentum." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9185–9193.

[Dro+20] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. "Trust in AutoML: exploring information needs for establishing trust in automated machine learning systems." In: *Proceedings of the 25th international conference on intelligent user interfaces*. 2020, pp. 297–307.

[Dua+09] Lian Duan, Lida Xu, Ying Liu, and Jun Lee. "Cluster-based outlier detection." In: *Annals of Operations Research* 168 (2009), pp. 151–168.

[DH13] Sophia H Duffy and Jamie Patrick Hopkins. "Sit, stay, drive: The future of autonomous car liability." In: *SMU Sci. & Tech. L. Rev.* 16 (2013), p. 453.

[DN25] Leonard Dung and Albert Newen. "The multidimensional profile methodology (MPM) for comparative cognition: towards a universal strategy of understanding animal minds." In: *Philosophical Studies* (2025), pp. 1–30.

[Dwi+23] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. "Explainable AI (XAI): Core ideas, techniques, and solutions." In: *ACM Computing Surveys* 55.9 (2023), pp. 1–33.

[Dzi+03] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. "The role of trust in automation reliance." In: *International journal of human-computer studies* 58.6 (2003), pp. 697–718.

[Ene09] U.S. Department of Energy. *Annual Energy Outlook 2009*. Tech. rep. DOE/EIA-0383(2009). Washington, D.C.: U.S. Department of Energy, 2009.

[Est+96] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *kdd*. Vol. 96. 34. 1996, pp. 226–231.

[Etm+19] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola-Bibiane Schönlieb. "On the connection between adversarial robustness and saliency map interpretability." In: *arXiv preprint arXiv:1905.04172* (2019).

[FTY17] Samira Farivar, Ofir Turel, and Yufei Yuan. "A trust-risk perspective on social commerce use: an examination of the biasing role of habit." In: *Internet Research* 27.3 (2017), pp. 586–607.

[Fas+14] Brittany Terese Fasy, Jisu Kim, Fabrizio Lecci, and Clément Maria. "Introduction to the R package TDA." In: *arXiv preprint arXiv:1411.1830* (2014).

[Fei+17] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. "Detecting adversarial samples from artifacts." In: *arXiv preprint arXiv:1703.00410* (2017).

[Fel+23a] Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. "A holistic approach to unifying automatic concept extraction and concept importance estimation." In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 54805–54818.

[Fel+23b] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. "Craft: Concept recursive activation factorization for explainability." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2711–2721.

[FL22] Andrea Ferrario and Michele Loi. "How explainability contributes to trust in AI." In: *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 2022, pp. 1457–1466.

[Feu+24] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. "Generative ai." In: *Business & Information Systems Engineering* 66.1 (2024), pp. 111–126.

[Fin+19] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. "Adversarial attacks on medical machine learning." In: *Science* 363.6433 (2019), pp. 1287–1289.

[Flo+18] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. "AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations." In: *Minds and machines* 28 (2018), pp. 689–707.

[FV18] Ruth Fong and Andrea Vedaldi. "Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8730–8738.

[Fuk96] Francis Fukuyama. *Trust: The social virtues and the creation of prosperity*. Simon and Schuster, 1996.

[FMO14] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras." In: *Proceedings of the symposium on eye tracking research and applications*. 2014, pp. 255–258.

[GSC07] Ran Gal, Ariel Shamir, and Daniel Cohen-Or. "Pose-oblivious shape signature." In: *IEEE transactions on visualization and computer graphics* 13.2 (2007), pp. 261–271.

[Gar+18] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. "Word embeddings quantify 100 years of gender and ethnic stereotypes." In: *Proceedings of the National Academy of Sciences* 115.16 (2018), E3635–E3644.

[GFW03] Thomas Gärtner, Peter Flach, and Stefan Wrobel. "On graph kernels: Hardness results and efficient alternatives." In: *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*. Springer. 2003, pp. 129–143.

[Gefoo] David Gefen. "E-commerce: the role of familiarity and trust." In: *Omega* 28.6 (2000), pp. 725–737.

[GHO+17] Roger Ghanem, David Higdon, Houman Owhadi, et al. *Handbook of uncertainty quantification*. Vol. 6. Springer New York, 2017.

[Gho+19] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. "Towards automatic concept-based explanations." In: *Advances in neural information processing systems* 32 (2019).

[Gil+15] Syed Omer Gilani, Ramanathan Subramanian, Yan Yan, David Melcher, Nicu Sebe, and Stefan Winkler. "Pet: An eye-tracking dataset for animal-centric pascal object classes." In: *2015 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2015, pp. 1–6.

[GTM93] Daniel T Gilbert, Romin W Tafarodi, and Patrick S Malone. "You can't not believe everything you read." In: *Journal of personality and social psychology* 65.2 (1993), p. 221.

[Gil+21] Omri Gillath, Ting Ai, Michael S Branicky, Shawn Keshmiri, Robert B Davison, and Ryan Spaulding. "Attachment and trust in artificial intelligence." In: *Computers in Human Behavior* 115 (2021), p. 106607.

[Gil+18] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. "Explaining explanations: An overview of interpretability of machine learning." In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 2018, pp. 80–89.

[GW20] Ella Glikson and Anita Williams Woolley. "Human trust in artificial intelligence: Review of empirical research." In: *Academy of Management Annals* 14.2 (2020), pp. 627–660.

[Gol20] Sanford C Goldberg. "Trust and reliance 1." In: *The Routledge handbook of trust and philosophy* (2020), pp. 97–108.

[GH15] Carlos A Gomez-Uribe and Neil Hunt. "The netflix recommender system: Algorithms, business value, and innovation." In: *ACM Transactions on Management Information Systems (TMIS)* 6.4 (2015), pp. 1–19.

[Goo+13] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. "Challenges in representation learning: A report on three machine learning contests." In: *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*. Springer. 2013, pp. 117–124.

[GSS14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." In: *arXiv preprint arXiv:1412.6572* (2014).

[Goy+19] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. "Explaining classifiers with causal concept effect (cace)." In: *arXiv preprint arXiv:1907.07165* (2019).

[Gre18] Marvin J Greenberg. *Algebraic topology: a first course*. CRC Press, 2018.

[Gre97] Richard L Gregory. "Knowledge in perception and illusion." In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 352.1358 (1997), pp. 1121–1127.

[Gre15] Richard L Gregory. "Eye and brain: The psychology of seeing." In: (2015).

[Gre70] Richard Langton Gregory. "The intelligent eye." In: (1970).

[Gre+23] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. "Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection." In: *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. 2023, pp. 79–90.

[Gri+20] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. "A survey of deep learning techniques for autonomous driving." In: *Journal of field robotics* 37.3 (2020), pp. 362–386.

[Gro+17] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. "On the (statistical) detection of adversarial examples." In: *arXiv preprint arXiv:1702.06280* (2017).

[Gu+23] Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. "A survey on transferability of adversarial examples across deep neural networks." In: *arXiv preprint arXiv:2310.17626* (2023).

[GR14] Shixiang Gu and Luca Rigazio. "Towards deep neural network architectures robust to adversarial examples." In: *arXiv preprint arXiv:1412.5068* (2014).

[GP56] Hs H Günthard and Hans Primas. "Zusammenhang von Graphentheorie und MO-Theorie von Molekeln mit Systemen konjugierter Bindungen." In: *Helvetica Chimica Acta* 39.6 (1956), pp. 1645–1653.

[Guo+17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. "On calibration of modern neural networks." In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330.

[Guo+18] Wenbo Guo, Dongliang Mu, Jun Xu, Purui Su, Gang Wang, and Xinyu Xing. "Lemna: Explaining deep learning based security applications." In: *proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 2018, pp. 364–379.

[Har13] Robin Hartshorne. *Algebraic geometry*. Vol. 52. Springer Science & Business Media, 2013.

[Hat05] Allen Hatcher. *Algebraic topology*. 2005.

[Hau+13] Sascha Hauke, Sebastian Biedermann, Max Mühlhäuser, and Dominik Heider. "On the application of supervised machine learning to trustworthiness assessment." In: *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*. IEEE. 2013, pp. 525–534.

[Haw80] Douglas M Hawkins. *Identification of outliers*. Vol. 11. Springer, 1980.

[Haw14] Katherine Hawley. "Trust, distrust and commitment." In: *Noûs* 48.1 (2014), pp. 1–20.

[Hay17] Andrew F Hayes. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications, 2017.

[He+15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV]. URL: https://arxiv.org/abs/1512.03385.

[Hig+10] Dave M Higdon, Mark C Anderson, Salman Habib, Richard Klein, Mark Berliner, Curt Covey, Omar Ghattas, Carlo Graziani, Mark Seager, Joseph Sefcik, et al. *Uncertainty quantification and error analysis*. Tech. rep. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 2010.

[Hof+18] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. "Metrics for explainable AI: Challenges and prospects." In: *arXiv preprint arXiv:1812.04608* (2018).

[Hof06] Gert Jan Hofstede. "Intrinsic and enforceable trust: a research agenda." In: (2006).

[Hol+20] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. "Explainable AI methods-a brief overview." In: *International workshop on extending explainable AI beyond deep models and classifiers*. Springer. 2020, pp. 13–38.

[Hou17] Michael E Houle. "Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications." In: *Similarity Search and Applications: 10th International Conference, SISAP 2017, Munich, Germany, October 4-6, 2017, Proceedings 10*. Springer. 2017, pp. 64–79.

[Hsu+07] Meng-Hsiang Hsu, Teresa L Ju, Chia-Hui Yen, and Chun-Ming Chang. "Knowledge sharing behavior in virtual communities: The relationship between trust, self-efficacy, and outcome expectations." In: *International journal of human-computer studies* 65.2 (2007), pp. 153–169.

[Hu+19] Shi Hu, Daniel Worrall, Stefan Knegt, Bas Veeling, Henkjan Huisman, and Max Welling. "Supervised uncertainty quantification for segmentation with multiple annotations." In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer. 2019, pp. 137–145.

[Hua+17] Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. "Safety verification of deep neural networks." In: *International conference on computer aided verification*. Springer. 2017, pp. 3–29.

[HW21] Eyke Hüllermeier and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods." In: *Machine learning* 110.3 (2021), pp. 457–506.

[Ily+19] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. "Adversarial examples are not bugs, they are features." In: *Advances in neural information processing systems* 32 (2019).

[Ing99] Ronald Inglehart. "Trust, well-being and democracy." In: *Democracy and trust* 88 (1999), pp. 88–120.

[Ize12] Alan Julian Izenman. "Introduction to manifold learning." In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.5 (2012), pp. 439–446.

[Jac+21] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI." In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 624–635.

[Jay+18] Upul Jayasinghe, Gyu Myoung Lee, Tai-Won Um, and Qi Shi. "Machine learning based trust computational model for IoT services." In: *IEEE Transactions on Sustainable Computing* 4.1 (2018), pp. 39–52.

[JL17] Robin Jia and Percy Liang. "Adversarial examples for evaluating reading comprehension systems." In: *arXiv preprint arXiv:1707.07328* (2017).

[Jia+18] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. "To trust or not to trust a classifier." In: *Advances in neural information processing systems* 31 (2018).

[Jia+22] Minqi Jiang, Michael Dennis, Jack Parker-Holder, Andrei Lupu, Heinrich Küttler, Edward Grefenstette, Tim Rocktäschel, and Jakob Foerster. "Grounding aleatoric uncertainty for unsupervised environment design." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 32868–32881.

[Joh67] Stephen C Johnson. "Hierarchical clustering schemes." In: *Psychometrika* 32.3 (1967), pp. 241–254.

[Kab+19] HM Dipu Kabir, Abbas Khosravi, Saeid Nahavandi, and Abdollah Kavousi-Fard. "Partial adversarial training for neural network-based uncertainty quantification." In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 5.4 (2019), pp. 595–606.

[Kac66] Mark Kac. "Can one hear the shape of a drum?" In: *The american mathematical monthly* 73.4P2 (1966), pp. 1–23.

[Kap+23] Alexandra D Kaplan, Theresa T Kessler, J Christopher Brill, and Peter A Hancock. "Trust in artificial intelligence: Meta-analytic findings." In: *Human factors* 65.2 (2023), pp. 337–359.

[Kat+17] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. "Reluplex: An efficient SMT solver for verifying deep neural networks." In: *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*. Springer. 2017, pp. 97–117.

[KTW23] Ehsan Kazemi, Fariborz Taherkhani, and Liqiang Wang. "On complementing unsupervised learning with uncertainty quantification." In: *Pattern Recognition Letters* 176 (2023), pp. 69–75.

[Kel+19] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. "Gaze360: Physically unconstrained gaze estimation in the wild." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6912–6921.

[Kem18] Richard Kemp. "Legal Aspects of Artificial Intelligence (v. 2.0)." In: *Kemp IT Law.–2016.–URL: https://www. kempitlaw. com/wp-content/uploads/2016/11/Legal-Aspects-of-AI-Kemp-IT-Law-v2. 0-Nov-2016-. pdf* (2018).

[KO01] Marc C Kennedy and Anthony O'Hagan. "Bayesian calibration of computer models." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3 (2001), pp. 425–464.

[KK22] Majd Khalil and Benny Kimelfeld. "The Complexity of the Shapley Value for Regular Path Queries." In: *arXiv preprint arXiv:2212.07720* (2022).

[Kha+20] Muhammad Haris Khan, John McDonagh, Salman Khan, Muhammad Shahabuddin, Aditya Arora, Fahad Shahbaz Khan, Ling Shao, and Georgios Tzimiropoulos. "Animalweb: A large-scale hierarchical dataset of annotated animal faces." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 6939–6948.

[KM16] Firas A Khasawneh and Elizabeth Munch. "Chatter detection in turning using persistent homology." In: *Mechanical Systems and Signal Processing* 70 (2016), pp. 527–541.

[Kil25] Ido Kilovaty. "Hacking Generative AI." In: *Loyola of Los Angeles Law Review* 58 (2025).

[Kim+18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." In: *International conference on machine learning*. PMLR. 2018, pp. 2668–2677.

[KFR08] Dan J Kim, Donald L Ferrin, and H Raghav Rao. "A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents." In: *Decision support systems* 44.2 (2008), pp. 544–564.

[Kim+16] Dan J Kim, Myung-Seong Yim, Vijayan Sugumaran, and H Raghav Rao. "Web assurance seal services, trust and consumers' concerns: An investigation of e-commerce transaction intentions across two nations." In: *European Journal of Information Systems* 25.3 (2016), pp. 252–273.

[KB17] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: `1412.6980 [cs.LG]`. URL: `https://arxiv.org/abs/1412.6980`.

[KR00] Amna Kirmani and Akshay R Rao. "No pain, no gain: A critical review of the literature on signaling unobservable product quality." In: *Journal of marketing* 64.2 (2000), pp. 66–79.

[Kir17] Alexandra Kirsch. "Explain to whom? Putting the user in the center of explainable AI." In: *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2017)*. 2017.

[Kiz16] René F Kizilcec. "How much information? Effects of transparency on trust in an algorithmic interface." In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016, pp. 2390–2395.

[KR21] Bran Knowles and John T Richards. "The sanction of authority: Promoting public trust in AI." In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 262–271.

[Koc+23] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. "ChatGPT: Jack of all trades, master of none." In: *Information Fusion* (2023), p. 101861.

[Kof13] Kurt Koffka. *Principles of Gestalt psychology*. routledge, 2013.

[KP16] Risi Kondor and Horace Pan. "The multiscale laplacian graph kernel." In: *Advances in neural information processing systems* 29 (2016).

[Kon01] Igor Kononenko. "Machine learning for medical diagnosis: history, state of the art and perspective." In: *Artificial Intelligence in medicine* 23.1 (2001), pp. 89–109.

[Kon+22] Katiana Kontolati, Dimitrios Loukrezis, Dimitrios G Giovanis, Lohit Vandanapu, and Michael D Shields. "A survey of unsupervised learning methods for high-dimensional uncertainty quantification in black-box-type problems." In: *Journal of Computational Physics* 464 (2022), p. 111313.

[KVF13] Danai Koutra, Joshua T Vogelstein, and Christos Faloutsos. "Deltacon: A principled massive-graph similarity function." In: *Proceedings of the 2013 SIAM international conference on data mining*. SIAM. 2013, pp. 162–170.

[KH+09] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images." In: (2009).

[KK16] Philipp Kulms and Stefan Kopp. "The effect of embodiment and competence on trust and cooperation in human–agent interaction." In: *International Conference on Intelligent Virtual Agents*. Springer. 2016, pp. 75–84.

[Kun+19] Alexander Kunze, Stephen J Summerskill, Russell Marshall, and Ashleigh J Filtness. "Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces." In: *Ergonomics* 62.3 (2019), pp. 345–360.

[KGB16] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. "Adversarial machine learning at scale." In: *arXiv preprint arXiv:1611.01236* (2016).

[KGB18] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." In: *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[LB20] Himabindu Lakkaraju and Osbert Bastani. "" How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations." In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 79–85.

[Lam+22] Benjamin Lambert, Florence Forbes, Alan Tucholka, Senan Doyle, Harmonie Dehaene, and Michel Dojat. "Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis." In: *arXiv preprint arXiv:2210.03736* (2022).

[Lan22] Jobst Landgrebe. "Certifiable ai." In: *Applied Sciences* 12.3 (2022), p. 1050.

[Lan+23] Markus Langer, Cornelius J König, Caroline Back, and Victoria Hemsing. "Trust in Artificial Intelligence: Comparing trust processes between human and automated trustees in light of unfair bias." In: *Journal of Business and Psychology* 38.3 (2023), pp. 493–508.

[Lan+19] Jens Lansing, Nils Siegfried, Ali Sunyaev, and Alexander Benlian. "Strategic signaling through cloud service certifications: Comparing the relative importance of certifications' assurances to companies and consumers." In: *The Journal of Strategic Information Systems* 28.4 (2019), p. 101579.

[Lap+19] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. "Unmasking Clever Hans predictors and assessing what machines really learn." In: *Nature communications* 10.1 (2019), p. 1096.

[LR23] Andrew Lapworth and Tom Roberts. "Habit, Artificial Intelligence and the Ontological Performance of Trust." In: *Performance Research* 28.6 (2023), pp. 73–81.

[LV07] John A Lee and Michel Verleysen. *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[LS04] John D Lee and Katrina A See. "Trust in automation: Designing for appropriate reliance." In: *Human factors* 46.1 (2004), pp. 50–80.

[LL20] JoonHo Lee and Gyemin Lee. "Model uncertainty for unsupervised domain adaptation." In: *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2020, pp. 1841–1845.

[Lee+06] Kwan Min Lee, Wei Peng, Seung-A Jin, and Chang Yan. "Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction." In: *Journal of communication* 56.4 (2006), pp. 754–772.

[LBW22] Dan Ley, Umang Bhatt, and Adrian Weller. "Diverse, global and amortised counterfactual explanations for uncertainty estimates." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 7. 2022, pp. 7390–7398.

[Li+21a] Jiahui Li, Kun Kuang, Lin Li, Long Chen, Songyang Zhang, Jian Shao, and Jun Xiao. "Instance-wise or class-wise? a tale of neighbor shapley for concept-based explanation." In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 3664–3672.

[Li+21b] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. "A survey of convolutional neural networks: analysis, applications, and prospects." In: *IEEE transactions on neural networks and learning systems* 33.12 (2021), pp. 6999–7019.

[LGM20] Q Vera Liao, Daniel Gruen, and Sarah Miller. "Questioning the AI: informing design practices for explainable AI user experiences." In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–15.

[LDA09] Brian Y Lim, Anind K Dey, and Daniel Avrahami. "Why and why not explanations improve the intelligibility of context-aware intelligent systems." In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2009, pp. 2119–2128.

[Lin+17] Yen-Chen Lin, Ming-Yu Liu, Min Sun, and Jia-Bin Huang. "Detecting adversarial attacks on neural network policies with visual foresight." In: *arXiv preprint arXiv:1710.00814* (2017).

[LTS23] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. "Generating with confidence: Uncertainty quantification for black-box large language models." In: *arXiv preprint arXiv:2305.19187* (2023).

[LPK20] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable ai: A review of machine learning interpretability methods." In: *Entropy* 23.1 (2020), p. 18.

[Liu+19] Gang Liu, Yu Yu, Kenneth A Funes Mora, and Jean-Marc Odobez. "A differential approach for gaze estimation." In: *IEEE transactions on pattern analysis and machine intelligence* 43.3 (2019), pp. 1092–1099.

[Liu+25] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. "Uncertainty quantification and confidence calibration in large language models: A survey." In: *arXiv preprint arXiv:2503.15850* (2025).

[Liu+10] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. "Understanding of internal clustering validation measures." In: *2010 IEEE international conference on data mining*. Ieee. 2010, pp. 911–916.

[Liu+16] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. "Delving into transferable adversarial examples and black-box attacks." In: *arXiv preprint arXiv:1611.02770* (2016).

[Loc+21] Steven Lockey, Nicole Gillespie, Daniel Holm, and Ida Asadi Someh. "A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions." In: (2021).

[LM05] Daniel Lowd and Christopher Meek. "Adversarial learning." In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. 2005, pp. 641–647.

[Low+12] Paul Benjamin Lowry, Greg Moody, Anthony Vance, Matthew Jensen, Jeff Jenkins, and Taylor Wells. "Using an elaboration likelihood approach to better understand the persuasiveness of website privacy assurance cues for online consumers." In: *Journal of the American Society for Information Science and technology* 63.4 (2012), pp. 755–776.

[LIF17] Jiajun Lu, Theerasit Issaranon, and David Forsyth. "Safetynet: Detecting and rejecting adversarial examples robustly." In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 446–454.

[LCY18] Pei-Hsuan Lu, Pin-Yu Chen, and Chia-Mu Yu. "On the limitation of local intrinsic dimensionality for characterizing the subspaces of adversarial examples." In: *arXiv preprint arXiv:1803.09638* (2018).

[LMS22] Roman Lukyanenko, Wolfgang Maass, and Veda C Storey. "Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities." In: *Electronic Markets* 32.4 (2022), pp. 1993–2020.

[LL17a] Scott Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI]. URL: https://arxiv.org/abs/1705.07874.

[Lun+20] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. "From local explanations to global understanding with explainable AI for trees." In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[LL17b] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions." In: *Advances in neural information processing systems* 30 (2017).

[LL17c] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions." In: *Advances in neural information processing systems* 30 (2017).

[LL17d] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions." In: *Advances in neural information processing systems* 30 (2017).

[Lyo+98] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. "Coding facial expressions with gabor wavelets." In: *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE. 1998, pp. 200–205.

[Ma+18] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawnsullivan2015introduction Song, Michael E Houle, and James Bailey. "Characterizing adversarial subspaces using local intrinsic dimensionality." In: *arXiv preprint arXiv:1801.02613* (2018).

[Mad+18] Ali Madani, Ramy Arnaout, Mohammad Mofrad, and Rima Arnaout. "Fast and accurate view classification of echocardiograms using deep learning." In: *NPJ digital medicine* 1.1 (2018), p. 6.

[Mad24] A. Maddox. *Institutional trust among refugees in Germany*. BAMF Brief Analysis 2. Retrieved from Bundesamt für Migration und Flüchtlinge. Bundesamt für Migration und Flüchtlinge (BAMF), 2024. URL: https://www.bamf.de/SharedDocs/Anlagen/EN/Forschung/Kurzanalysen/kurzanalyse2-2024-iab-bamf-soep-institutionsvertrauen.pdf.

[MW07] Poornima Madhavan and Douglas A Wiegmann. "Similarities and differences between human–human and human–automation trust: an integrative review." In: *Theoretical Issues in Ergonomics Science* 8.4 (2007), pp. 277–301.

[Mad+17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards deep learning models resistant to adversarial attacks." In: *arXiv preprint arXiv:1706.06083* (2017).

[MS18] Aleksander Mądry and Ludwig Schmidt. *A Brief Introduction to Adversarial Examples*. July 2018. URL: `http://gradientscience.org/intro_adversarial/`.

[Mak+19] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. "Recurrent neural network transducer for audio-visual speech recognition." In: *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE. 2019, pp. 905–912.

[Mar+14] Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. "The gudhi library: Simplicial complexes and persistent homology." In: *Mathematical Software–ICMS 2014: 4th International Congress, Seoul, South Korea, August 5-9, 2014. Proceedings 4*. Springer. 2014, pp. 167–174.

[Mar92] Stephen Marsh. "Trust in distributed artificial intelligence." In: *European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. Springer. 1992, pp. 94–112.

[Mar+23] Charles Marx, Youngsuk Park, Hilaf Hasson, Yuyang Wang, Stefano Ermon, and Luke Huan. "But are you sure? an uncertainty-aware perspective on explainable ai." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 7375–7391.

[Mas+12] Mohammad M Masud, Clay Woolam, Jing Gao, Latifur Khan, Jiawei Han, Kevin W Hamlen, and Nikunj C Oza. "Facing the reality of data stream classification: coping with scarcity of labeled data." In: *Knowledge and information systems* 33 (2012), pp. 213–244.

[MBL16] Tamilla Mavlanova, Raquel Benbunan-Fich, and Guido Lang. "The role of external and internal signals in E-commerce." In: *Decision Support Systems* 87 (2016), pp. 59–68.

[MDS95] Roger C Mayer, James H Davis, and F David Schoorman. "An integrative model of organizational trust." In: *Academy of management review* 20.3 (1995), pp. 709–734.

[May+15] Michael Mayhew, Michael Atighetchi, Aaron Adler, and Rachel Greenstadt. "Use of machine learning in big data analytics for insider threat detection." In: *MILCOM 2015-2015 IEEE Military Communications Conference*. IEEE. 2015, pp. 915–922.

[May14] Philipp Mayring. "Qualitative content analysis: theoretical foundation, basic procedures and software solution." In: (2014).

[McA95] Daniel J McAllister. "Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations." In: *Academy of management journal* 38.1 (1995), pp. 24–59.

[MB19] Patricia L McDermott and Ronna N ten Brink. "Practical guidance for evaluating calibrated trust." In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 63. 1. SAGE Publications Sage CA: Los Angeles, CA. 2019, pp. 362–366.

[MHA+17] Leland McInnes, John Healy, Steve Astels, et al. "hdbscan: Hierarchical density based clustering." In: *J. Open Source Softw.* 2.11 (2017), p. 205.

[MHM18] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." In: *arXiv preprint arXiv:1802.03426* (2018).

[MKC04] D Harrison McKnight, Charles J Kacmar, and Vivek Choudhury. "Shifting Factors and the Ineffectiveness of Third Party Assurance Seals: A two-stage model of initial trust in a web business." In: *Electronic markets* 14.3 (2004), pp. 252–266.

[MZ06] Carolyn McLeod and Edward N Zalta. "Trust In Stanford encyclopedia of philosophy." In: *Metaphysics Research Lab, Stanford University* (2006).

[MJ+01] Larry R Medsker, Lakhmi Jain, et al. "Recurrent neural networks." In: *Design and Applications* 5.64-67 (2001), p. 2.

[Mém11] Facundo Mémoli. "A spectral notion of Gromov–Wasserstein distance and related methods." In: *Applied and Computational Harmonic Analysis* 30.3 (2011), pp. 363–401.

[Met+17] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. "On detecting adversarial perturbations." In: *arXiv preprint arXiv:1702.04267* (2017).

[MSK16] John Michael, Natalie Sebanz, and Günther Knoblich. "The sense of commitment: A minimal approach." In: *Frontiers in psychology* 6 (2016), p. 162497.

[Mic+20] Rhiannon Michelmore, Matthew Wicker, Luca Laurenti, Luca Cardelli, Yarin Gal, and Marta Kwiatkowska. "Uncertainty quantification with statistical guarantees in end-to-end autonomous driving control." In: *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2020, pp. 7344–7350.

[Mil19] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences." In: *Artificial intelligence* 267 (2019), pp. 1–38.

[Min+22] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. "Explainable artificial intelligence: a comprehensive review." In: *Artificial Intelligence Review* (2022), pp. 1–66.

[Mir+16] Alexander G Mirnig, Philipp Wintersberger, Christine Sutter, and Jürgen Ziegler. "A framework for analyzing and calibrating trust in automated vehicles." In: *Adjunct proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications*. 2016, pp. 33–38.

[MZR21] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. "A multidisciplinary survey and framework for design and evaluation of explainable AI systems." In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11.3-4 (2021), pp. 1–45.

[Mon+19] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. "Layer-wise relevance propagation: an overview." In: *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), pp. 193–209.

[Moo17] Richard Moore. "Gricean communication and cognitive development." In: *The Philosophical Quarterly* 67.267 (2017), pp. 303–326.

[Moo+17] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. "Universal adversarial perturbations." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1765–1773.

[MFF16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. "Deepfool: a simple and accurate method to fool deep neural networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2574–2582.

[MCM19] Andrea Morichetta, Pedro Casas, and Marco Mellia. "EXPLAIN-IT: Towards explainable AI for unsupervised network traffic analysis." In: *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*. 2019, pp. 22–28.

[Mos+22] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. "SHAP-based explanation methods: a review for NLP interpretability." In: *Proceedings of the 29th international conference on computational linguistics*. 2022, pp. 4593–4603.

[MST20] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. "Explaining machine learning classifiers through diverse counterfactual explanations." In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 607–617.

[MR15] Sankha S Mukherjee and Neil Martin Robertson. "Deep head pose: Gaze-direction estimation in multimodal video." In: *IEEE Transactions on Multimedia* 17.11 (2015), pp. 2094–2107.

[Mun18] James R Munkres. *Elements of algebraic topology*. CRC press, 2018.

[Nai+22] Nithesh Naik, BM Hameed, Dasharathraj K Shetty, Dishant Swain, Milap Shah, Rahul Paul, Kaivalya Aggarwal, Sufyan Ibrahim, Vathsala Patil, Komal Smriti, et al. "Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility?" In: *Frontiers in surgery* 9 (2022), p. 266.

[Nas+19] Muhammad Muzammal Naseer, Salman H Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. "Cross-domain transferability of adversarial perturbations." In: *Advances in Neural Information Processing Systems* 32 (2019).

[NM00] Clifford Nass and Youngme Moon. "Machines and mindlessness: Social responses to computers." In: *Journal of social issues* 56.1 (2000), pp. 81–103.

[Nas17] Vladimir Nasteski. "An overview of the supervised machine learning methods." In: *Horizons. b* 4.51-62 (2017), p. 56.

[Nau+23] Meike Nauta, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. "Pip-net: Patch-based intuitive prototypes for interpretable image classification." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2744–2753.

[Net+11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. "Reading digits in natural images with unsupervised feature learning." In: *NIPS workshop on deep learning and unsupervised feature learning*. Vol. 2011. 2. Granada. 2011, p. 4.

[NV17] Albert Newen and Petra Vetter. "Why cognitive penetration of our perceptual experience is still the most plausible account." In: *Consciousness and cognition* 47 (2017), pp. 26–37.

[New+25a] Carina Newen, Daniel Bodemer, Sonja Glantz, Emmanuel Müller, Magdalena Wischnewski, and Lenka Schnaubert. "Uncertainty Awareness and Trust in Explainable AI – On Trust Calibration using Local and Global Explanations." In: *2025 IEEE International Conference on Data Mining (ICDM)*. 2025.

[New+25b] Carina Newen, Luca Hinkamp, Maria Ntonti, and Emmanuel Müller. *Do you see what I see? An Ambiguous Optical Illusion Dataset exposing limitations of Explainable AI*. 2025. arXiv: 2505.21589 [cs.CV]. URL: https://arxiv.org/abs/2505.21589.

[NM22a] Carina Newen and Emmanuel Müller. "Unsupervised DeepView: Global Explainability of Uncertainties for High Dimensional Data." In: *2022 IEEE International Conference on Knowledge Graph (ICKG)*. IEEE. 2022, pp. 196–202.

[NM22b] Carina Newen and Emmanuel Müller. "Unsupervised DeepView: Global Uncertainty Visualization for High Dimensional Data." In: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2022, pp. 1–8.

[NM22c] Carina Newen and Emmanuel Müller. "Unsupervised DeepView: Global Uncertainty Visualization for High Dimensional Data." In: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2022, pp. 1–8.

[NM23] Carina Newen and Emmanuel Müller. "On the Independence of Adversarial Transferability to Topological Changes in the Dataset." In: *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. 2023, pp. 1–8.

[NMN25] Carina Newen, Emmanuel Müller, and Albert Newen. "Trust and Uncertainties: Characterizing Trustworthy AI Systems Within a Multidimensional Theory of Trust." In: *Topoi* (2025), pp. 1–22.

[NPM25] Carina Newen, Sofia Vergara Puccini, and Emmanuel Müller. "Certainty Attacks using Explainability Preprocessing." In: *DAWAK 2025*. Springer, 2025.

[Ng+22] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. "Animal kingdom: A large and diverse dataset for animal behavior understanding." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 19023–19034.

[NK22] Thao Ngo and Nicole Krämer. "Exploring folk theories of algorithmic news curation for explainable design." In: *Behaviour & Information Technology* 41.15 (2022), pp. 3346–3359.

[NLC11] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival." In: *Proceedings of the National Academy of Sciences* 108.17 (2011), pp. 7265–7270.

[Noa+21] Adam Noack, Isaac Ahern, Dejing Dou, and Boyang Li. "An empirical study on the relation between network interpretability and adversarial robustness." In: *SN Computer Science* 2.1 (2021), p. 32.

[Now21] Helga Nowotny. *In AI we trust: Power, illusion and control of predictive algorithms*. John Wiley & Sons, 2021.

[Nug+20] Selin E Nugent, Paul Jackson, Susan Scott-Parker, James Partridge, Rebecca Raper, Chara Bakalis, Alex Shepherd, Arijit Mitra, Jintao Long, Kevin Maynard, et al. "Recruitment AI has a Disability Problem: Questions employers should be asking to ensure fairness in recruitment." In: (2020).

[Onei02] Onora O'neill. *Autonomy and trust in bioethics*. Cambridge University Press, 2002.

[ON15] Keiron O'shea and Ryan Nash. "An introduction to convolutional neural networks." In: *arXiv preprint arXiv:1511.08458* (2015).

[Ois+16] Benjamin C Oistad, Catherine E Sembroski, Kathryn A Gates, Margaret M Krupp, Marlena R Fraune, and Selma Šabanović. "Colleague or tool? Interactivity increases positive perceptions of and willingness to interact with a robotic co-worker." In: *International conference on social robotics*. Springer. 2016, pp. 774–785.

[Ope24] OpenAI. *ChatGPT*. Retrieved from OpenAI. 2024. URL: `https://www.openai.com/chatgpt`.

[OV23] Alina Oprea and Apostol Vassilev. *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations*. Tech. rep. National Institute of Standards and Technology, 2023.

[Oqu+15] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. "Is object localization for free?-weakly-supervised learning with convolutional neural networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 685–694.

[Osa+02] Robert Osada, Thomas Funkhouser, Bernard Chazelle, and David Dobkin. "Shape distributions." In: *ACM Transactions on Graphics (TOG)* 21.4 (2002), pp. 807–832.

[OEK18] Adnan Ozyilmaz, Berrin Erdogan, and Aysegul Karaeminogullari. "Trust in organization as a moderator of the relationship between self-efficacy and workplace outcomes: A social cognitive theory-based examination." In: *Journal of Occupational and Organizational Psychology* 91.1 (2018), pp. 181–204.

[PDG10] Panagiotis Papadimitriou, Ali Dasdan, and Hector Garcia-Molina. "Web graph similarity for anomaly detection." In: *Journal of Internet Services and Applications* 1 (2010), pp. 19–30.

[PMG16] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. arXiv 2016." In: *arXiv preprint arXiv:1605.07277* (2016).

[Pap+17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. "Practical black-box attacks against machine learning." In: *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 2017, pp. 506–519.

[Pap+16a] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. "The limitations of deep learning in adversarial settings." In: *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE. 2016, pp. 372–387.

[Pap+16b] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. "Distillation as a defense to adversarial perturbations against deep neural networks." In: *2016 IEEE symposium on security and privacy (SP)*. IEEE. 2016, pp. 582–597.

[PR97] Raja Parasuraman and Victor Riley. "Humans and automation: Use, misuse, disuse, abuse." In: *Human factors* 39.2 (1997), pp. 230–253.

[Par+22] Darsh Parekh, Nishi Poddar, Aakash Rajpurkar, Manisha Chahal, Neeraj Kumar, Gyanendra Prasad Joshi, and Woong Cho. "A review on autonomous vehicles: Progress, methods and challenges." In: *Electronics* 11.14 (2022), p. 2162.

[Par13] Wendy S Parker. "Ensemble modeling, uncertainty and robust predictions." In: *Wiley interdisciplinary reviews: Climate change* 4.3 (2013), pp. 213–223.

[Pas+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. "Pytorch: An imperative style, high-performance deep learning library." In: *Advances in neural information processing systems* 32 (2019).

[Pie11] Wolter Pieters. "Explanation and trust: what to tell the user in security and AI?" In: *Ethics and information technology* 13 (2011), pp. 53–64.

[Poe+23] Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. "Concept-based explainable artificial intelligence: A survey." In: *arXiv preprint arXiv:2312.12936* (2023).

[PHD16a] Gloria Pöhler, Tobias Heine, and Barbara Deml. "Item analysis and factor structure of a questionnaire for trust in automated systems." In: *Zeitschrift für Arbeitswissenschaft* 70 (2016), pp. 151–160.

[PHD16b] Gloria Pöhler, Tobias Heine, and Barbara Deml. "Itemanalyse und Faktorstruktur eines Fragebogens zur Messung von Vertrauen im Umgang mit automatischen Systemen." In: *Zeitschrift für Arbeitswissenschaft* 3.70 (2016), pp. 151–160.

[Pop+09] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. "Multilayer perceptron and neural networks." In: *WSEAS Transactions on Circuits and Systems* 8.7 (2009), pp. 579–588.

[Qai+19] Talha Qaiser, Yee-Wah Tsang, Daiki Taniyama, Naoya Sakamoto, Kazuaki Nakane, David Epstein, and Nasir Rajpoot. "Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features." In: *Medical image analysis* 55 (2019), pp. 1–14.

[RTM13] Irene Rae, Leila Takayama, and Bilge Mutlu. "In-body experiences: embodiment, control, and trust in robot-mediated communication." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2013, pp. 1921–1930.

[RWN17] MINA Rawat, Martin Wistuba, and Maria-Irina Nicolae. "Harnessing model uncertainty for detecting adversarial examples." In: *NIPS Workshop on Bayesian Deep Learning*. 2017.

[RY22] Amy Rechkemmer and Ming Yin. "When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models." In: *Proceedings of the 2022 chi conference on human factors in computing systems*. 2022, pp. 1–14.

[Rei+21] Annika Reinke, Minu D Tizabi, Carole H Sudre, Matthias Eisenmann, Tim Rädsch, Michael Baumgartner, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, et al. "Common limitations of image processing metrics: A picture story." In: *arXiv preprint arXiv:2104.05642* (2021).

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016. arXiv: `1602.04938 [cs.LG]`.

[RSG18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: High-precision model-agnostic explanations." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[RL19] Mireia Ribera and Àgata Lapedriza García. "Can we do better explanations? A proposal of user-centered explainable AI." In: CEUR Workshop Proceedings. 2019.

[Rob+23] David A Robb, José Lopes, Muneeb I Ahmad, Peter E McKenna, Xingkun Liu, Katrin Lohan, and Helen Hastie. "Seeing eye to eye: trustworthy embodiment for task-based conversational agents." In: *Frontiers in Robotics and AI* 10 (2023), p. 1234767.

[RWS11] Vanessa Robins, Peter John Wood, and Adrian P Sheppard. "Theory and algorithms for constructing discrete Morse complexes from grayscale digital images." In: *IEEE Transactions on pattern analysis and machine intelligence* 33.8 (2011), pp. 1646–1658.

[RBP18] Roberto Rocchetta, Matteo Broggi, and Edoardo Patelli. "Do we have enough data? Robust reliability via uncertainty quantification." In: *Applied Mathematical Modelling* 54 (2018), pp. 710–721.

[RR19] Avi Rosenfeld and Ariella Richardson. "Explainability in human–agent systems." In: *Autonomous Agents and Multi-Agent Systems* 33 (2019), pp. 673–705.

[RD18] Andrew Ross and Finale Doshi-Velez. "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[Ros18] Francesca Rossi. "Building trust in artificial intelligence." In: *Journal of international affairs* 72.1 (2018), pp. 127–134.

[RU18] Cynthia Rudin and Berk Ustun. "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice." In: *Interfaces* 48.5 (2018), pp. 449–466.

[Rus+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "Imagenet large scale visual recognition challenge." In: *International journal of computer vision* 115 (2015), pp. 211–252.

[RN21] Stuart J Russell and Peter Norvig. "Artificial Intelligence: A Modern Approach, Global Edition 4e." In: (2021).

[Rya20] Mark Ryan. "In AI we trust: ethics, artificial intelligence, and reliability." In: *Science and Engineering Ethics* 26.5 (2020), pp. 2749–2767.

[Saa+19] Khaled Saab, Jared Dunnmon, Alexander Ratner, Daniel Rubin, and Christopher Ré. "Improving sample complexity with observational supervision." In: (2019).

[Sal+15] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust." In: *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 2015, pp. 141–148.

[SKB22] Abhishek Singh Sambyal, Narayanan C Krishnan, and Deepti R Bathula. "Towards reducing aleatoric uncertainty for medical imaging tasks." In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–4.

[Sam+21] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. "Explaining deep neural networks and beyond: A review of methods and applications." In: *Proceedings of the IEEE* 109.3 (2021), pp. 247–278.

[San20] Martin Sand. *Sven Nyholm: Humans and Robots: Ethics, Agency, and Anthropomorphism*. 2020.

[SF12] Alberto Sanfeliu and King-Sun Fu. "A distance measure between attributed relational graphs for pattern recognition." In: *IEEE transactions on systems, man, and cybernetics* 3 (2012), pp. 353–362.

[SM21] Filippo Santoni de Sio and Giulio Mecacci. "Four responsibility gaps with artificial intelligence: Why they matter and how to address them." In: *Philosophy & Technology* 34.4 (2021), pp. 1057–1084.

[SL21] Nadine Schlicker and Markus Langer. "Towards warranted trust: A model on the relation between actual and perceived system trustworthiness." In: *Proceedings of Mensch und Computer 2021*. 2021, pp. 325–329.

[SB19] Philipp Schmidt and Felix Biessmann. "Quantifying interpretability and trust in machine learning systems." In: *arXiv preprint arXiv:1901.08558* (2019).

[Sch+19] Jonas Schuett et al. "A legal definition of AI." In: *arXiv preprint arXiv:1909.01095* (2019).

[SGH15] Alexander Schulz, Andrej Gisbrecht, and Barbara Hammer. "Using discriminative dimensionality reduction to visualize classifiers." In: *Neural Processing Letters* 42 (2015), pp. 27–54.

[SHH20] Alexander Schulz, Fabian Hinder, and Barbara Hammer. "DeepView: Visualizing classification boundaries of deep neural networks as scatter plots using discriminative dimensionality reduction." In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*. 2020, pp. 2305–2311. DOI: 10.24963/ijcai.2020/319.

[SK19] Patrick Schwab and Walter Karlen. "Cxplain: Causal explanations for model interpretation under uncertainty." In: *Advances in neural information processing systems* 32 (2019).

[Sch+23] Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, Günter Klambauer, and Sepp Hochreiter. "Quantification of uncertainty with adversarial models." In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 19446–19484.

[Sel+17] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization." In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[SXL19] Haeseung Seo, Aiping Xiong, and Dongwon Lee. "Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation." In: *Proceedings of the 10th ACM Conference on Web Science*. 2019, pp. 265–274.

[Set+21] Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. "Glocalx-from local to global explanations of black box ai models." In: *Artificial Intelligence* 294 (2021), p. 103457.

[Seu21] Dominik Seuß. "Bridging the gap between explainable AI and uncertainty quantification to enhance trustability." In: *arXiv preprint arXiv:2105.11828* (2021).

[SDB16] Lee M Seversky, Shelby Davis, and Matthew Berger. "On time-series topological data analysis: New data and opportunities." In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 59–67.

[SYN15] Uri Shaham, Yutaro Yamada, and Sahand Negahban. "Understanding adversarial training: Increasing local stability of neural nets through robust optimization." In: *arXiv preprint arXiv:1511.05432* (2015).

[Sha+24] Haz Sameen Shahgir, Khondker Salman Sayeed, Abhik Bhattacharjee, Wasi Uddin Ahmad, Yue Dong, and Rifat Shahriyar. "Illusionvqa: A challenging optical illusion dataset for vision language models." In: *arXiv preprint arXiv:2403.15952* (2024).

[SB06] Hamid R Sheikh and Alan C Bovik. "Image information and visual quality." In: *IEEE Transactions on image processing* 15.2 (2006), pp. 430–444.

[Shi21] Donghee Shin. "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI." In: *International Journal of Human-Computer Studies* 146 (2021), p. 102551.

[Sho+25] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. "A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions." In: *ACM Computing Surveys* (2025).

[SW18] Keng Siau and Weiyu Wang. "Building trust in artificial intelligence, machine learning, and robotics." In: *Cutter business technology journal* 31.2 (2018), pp. 47–53.

[SMC+07] Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. "Topological methods for the analysis of high dimensional data sets and 3d object recognition." In: *PBG@ Eurographics* 2 (2007), pp. 091–100.

[SL95] Jonas Sjöberg and Lennart Ljung. "Overtraining, regularization and searching for a minimum, with application to neural networks." In: *International Journal of Control* 62.6 (1995), pp. 1391–1407.

[Skr+10] Primoz Skraba, Maks Ovsjanikov, Frederic Chazal, and Leonidas Guibas. "Persistence-based segmentation of deformable shapes." In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE. 2010, pp. 45–52.

[Sla+21] Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. "Reliable post hoc explanations: Modeling uncertainty in explainability." In: *Advances in neural information processing systems* 34 (2021), pp. 9391–9404.

[SG18] Lewis Smith and Yarin Gal. "Understanding measures of uncertainty for adversarial example detection." In: *arXiv preprint arXiv:1803.08533* (2018).

[Smi24] Ralph C Smith. *Uncertainty quantification: theory, implementation, and applications*. SIAM, 2024.

[Soi17] Christian Soize. *Uncertainty quantification*. Springer, 2017.

[SY15] Yan-Yan Song and LU Ying. "Decision tree methods: applications for classification and prediction." In: *Shanghai archives of psychiatry* 27.2 (2015), p. 130.

[Spe+10] Dan Sperber, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. "Epistemic vigilance." In: *Mind & language* 25.4 (2010), pp. 359–393.

[SS97] Alessandro Sperduti and Antonina Starita. "Supervised neural networks for the classification of structures." In: *IEEE transactions on neural networks* 8.3 (1997), pp. 714–735.

[Sul15] Timothy John Sullivan. *Introduction to uncertainty quantification*. Vol. 63. Springer, 2015.

[Sun+23] Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. "Explain any concept: Segment anything meets concept-based explanation." In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 21826–21840.

[Sun+08] S Shyam Sundar et al. *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA, 2008.

[STY17] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[SLL20] Harini Suresh, Natalie Lao, and Ilaria Liccardi. "Misplaced trust: Measuring the interference of machine learning in human decision-making." In: *Proceedings of the 12th ACM Conference on Web Science*. 2020, pp. 315–324.

[Sut19] Margit Sutrop. "Should we trust artificial intelligence?" In: *Trames* 23.4 (2019), pp. 499–522.

[Sze+13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. "Intriguing properties of neural networks." In: *arXiv preprint arXiv:1312.6199* (2013).

[Tab+19] Elham Tabassi, Kevin J Burns, Michael Hadjimichael, Andres D Molina-Markham, and Julian T Sexton. "A taxonomy and terminology of adversarial machine learning." In: *NIST IR* 2019 (2019), pp. 1–29.

[Tah+19] Ahmed Taha, Yi-Ting Chen, Teruhisa Misu, Abhinav Shrivastava, and Larry Davis. "Unsupervised data uncertainty learning in visual retrieval systems." In: *arXiv preprint arXiv:1902.02586* (2019).

[Tan+19] Ryutaro Tanno, Daniel Worrall, Enrico Kaden, Aurobrata Ghosh, Francesco Grussu, Alberto Bizzi, Stamatios N Sotiropoulos, Antonio Criminisi, and Daniel C Alexander. "Uncertainty quantification in deep learning for safer neuroimage enhancement." In: *arXiv preprint arXiv:1907.13418* (2019).

[TSL00] Joshua B Tenenbaum, Vin de Silva, and John C Langford. "A global geometric framework for nonlinear dimensionality reduction." In: *science* 290.5500 (2000), pp. 2319–2323.

[Ten+19] Nathan L Tenhundfeld, Ewart J De Visser, Kerstin S Haring, Anthony J Ries, Victor S Finomore, and Chad C Tossell. "Calibrating trust in automation through familiarity with the autoparking feature of a Tesla Model X." In: *Journal of cognitive engineering and decision making* 13.4 (2019), pp. 279–294.

[TG09] Ellen FJ Ter Huurne and Jan M Gutteling. "How to trust? The importance of self-efficacy and social trust in public responses to industrial risks." In: *Journal of risk research* 12.6 (2009), pp. 809–824.

[Tha+17] Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim, and Aboul Ella Hassanien. "Linear discriminant analysis: A detailed tutorial." In: *AI communications* 30.2 (2017), pp. 169–190.

[TLS21] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. "Trustworthy artificial intelligence." In: *Electronic Markets* 31 (2021), pp. 447–464.

[Tor+20] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. "The relationship between trust in AI and trustworthy machine learning technologies." In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 272–283.

[Tos+24] Chad C Tossell, Nathan L Tenhundfeld, Ali Momen, Katrina Cooley, and Ewart J de Visser. "Student perceptions of ChatGPT use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence." In: *IEEE Transactions on Learning Technologies* (2024).

[Tra+17] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. "The space of transferable adversarial examples." In: *arXiv preprint arXiv:1704.03453* (2017).

[Tsi+18a] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. "Robustness may be at odds with accuracy." In: *arXiv preprint arXiv:1805.12152* (2018).

[Tsi+18b] Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alexander Bronstein, and Emmanuel Müller. "Netlsd: hearing the shape of a graph." In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2018, pp. 2347–2356.

[Tsu18] Ken Tsui. *Perhaps the Simplest Introduction of Adversarial Examples Ever*. Aug. 2018. URL: `https://towardsdatascience.com/perhaps-the-simplest-introduction-of-adversarial-examples-ever-c0839a759b8d`.

[TS+97] Janet M Twomey, Alice E Smith, et al. "Validation and verification." In: *Artificial neural networks for civil engineers: Fundamentals and applications* (1997), pp. 44–64.

[Ume17] Yuhei Umeda. "Time series classification via topological data analysis." In: *Information and Media Technologies* 12 (2017), pp. 228–239.

[UNE21] C UNESCO. *Recommendation on the ethics of artificial intelligence*. 2021.

[VH08] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[Van+17] Anouk Van Maris, Hagen Lehmann, Lorenzo Natale, and Beata Grzyb. "The influence of a robot's embodiment on trust: A longitudinal study." In: *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on human-robot interaction*. 2017, pp. 313–314.

[VZ21] Michael Veale and Frederik Zuiderveen Borgesius. "Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach." In: *Computer Law Review International* 22.4 (2021), pp. 97–112.

[VBC21] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. "How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies." In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021), pp. 1–39.

[WC21] Echo Wen Wan and Rocky Peng Chen. "Anthropomorphism and object attachment." In: *Current Opinion in Psychology* 39 (2021), pp. 88–93.

[Wan+20a] Jingwen Wang, Xuyang Jing, Zheng Yan, Yulong Fu, Witold Pedrycz, and Laurence T Yang. "A survey on trust evaluation based on machine learning." In: *ACM Computing Surveys (CSUR)* 53.5 (2020), pp. 1–36.

[Wan+20b] Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, and Quanshi Zhang. "A unified approach to interpreting and boosting adversarial transferability." In: *arXiv preprint arXiv:2010.04055* (2020).

[WTC17] Xin Wang, Nicolas Thome, and Matthieu Cord. "Gaze latent support vector machine for image classification improved by weakly supervised region selection." In: *Pattern Recognition* 72 (2017), pp. 59–71.

[Wan+21a] Xinming Wang, Jianhua Zhang, Hanlin Zhang, Shuwen Zhao, and Honghai Liu. "Vision-based gaze estimation: A review." In: *IEEE Transactions on Cognitive and Developmental Systems* 14.2 (2021), pp. 316–332.

[Wan+21b] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. "Data-uncertainty guided multi-phase learning for semi-supervised object detection." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 4568–4577.

[Wan+21c] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. "Feature importance-aware transferable adversarial attacks." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 7639–7648.

[Wan+04] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. "Image quality assessment: from error visibility to structural similarity." In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[Was+23] Futa Waseda, Sosuke Nishikawa, Trung-Nghia Le, Huy H Nguyen, and Isao Echizen. "Closer look at the transferability of adversarial examples: How they fool different models differently." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 1360–1368.

[Wat23] David S Watson. "On the philosophy of unsupervised learning." In: *Philosophy & Technology* 36.2 (2023), p. 28.

[Wil16] Sven Wille. *Prototype for a Fair and Efficient Democratic Government*. Scotts Valley: CreateSpace Independent Publishing Platform, 2016.

[Win+21] Philip Matthias Winter, Sebastian Eder, Johannes Weissenböck, Christoph Schwald, Thomas Doms, Tom Vogt, Sepp Hochreiter, and Bernhard Nessler. "Trusted artificial intelligence: Towards certification of machine learning applications." In: *arXiv preprint arXiv:2103.16910* (2021).

[Wis+24a] Magdalena Wischnewski, Nicole Kramer, Christian Janiesch, Emmanuel Muller, Theodor Schnitzler, and Carina Newen. "In seal we trust?: Investigating the effect of certifications on perceived trustworthiness of AI systems." In: *Human-Machine Communication* 8 (2024), pp. 141–162.

[Wis+24b] Magdalena Wischnewski, Nicole Krämer, Christian Janiesch, Emmanuel Müller, Theodor Schnitzler, and Carina Newen. "In Seal We Trust? Investigating the Effect of Certifications on Perceived Trustworthiness of AI Systems." In: *Human-Machine Communication* 8.1 (2024), p. 7.

[WKM23] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. "Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions." In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–16.

[WX18] Rey Wiyatno and Anqi Xu. "Maximal jacobian-based saliency map attack." In: *arXiv preprint arXiv:1808.07945* (2018).

[WZ20] Lei Wu and Zhanxing Zhu. "Towards understanding and improving the transferability of adversarial examples in deep neural networks." In: *Asian Conference on Machine Learning*. PMLR. 2020, pp. 837–850.

[WZT+18] Lei Wu, Zhanxing Zhu, Cheng Tai, et al. "Understanding and enhancing the transferability of adversarial examples." In: *arXiv preprint arXiv:1802.09707* (2018).

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms." In: *arXiv preprint arXiv:1708.07747* (2017).

[Xie+19] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. "Improving transferability of adversarial examples with input diversity." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2730–2739.

[Xu+19] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. "Explainable AI: A brief survey on history, research areas, approaches and challenges." In: *Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8*. Springer. 2019, pp. 563–574.

[Yan+20] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin 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*. 2020, pp. 189–201.

[YW22] Rongbin Yang and Santoso Wibowo. "User trust in artificial intelligence: A comprehensive conceptual framework." In: *Electronic Markets* 32.4 (2022), pp. 2053–2077.

[Yao+19] Zhewei Yao, Amir Gholami, Peng Xu, Kurt Keutzer, and Michael W Mahoney. "Trust region based adversarial attack on neural networks." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 11350–11359.

[Yeh+20] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. "On completeness-aware concept-based explanations in deep neural networks." In: *Advances in neural information processing systems* 33 (2020), pp. 20554–20565.

[YW01] Michelle Yeh and Christopher D Wickens. "Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration." In: *Human Factors* 43.3 (2001), pp. 355–365.

[YWW19] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. "Understanding the effect of accuracy on trust in machine learning models." In: *Proceedings of the 2019 chi conference on human factors in computing systems*. 2019, pp. 1–12.

[YB22] Bang Xiang Yong and Alexandra Brintrup. "Bayesian autoencoders with uncertainty quantification: Towards trustworthy anomaly detection." In: *Expert Systems with Applications* 209 (2022), p. 118196.

[Yu+21] Hang Yu, Laurence T Yang, Qingchen Zhang, David Armstrong, and M Jamal Deen. "Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives." In: *Neurocomputing* 444 (2021), pp. 92–110.

[ZBW22] John Zerilli, Umang Bhatt, and Adrian Weller. "How transparency modulates trust in artificial intelligence." In: *Patterns* 3.4 (2022).

[ZBZ23] Minxing Zhang, Michael Backes, and Xiao Zhang. "Generating Less Certain Adversarial Examples Improves Robust Generalization." In: *arXiv preprint arXiv:2310.04539* (2023).

[Zha+21] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. "Invertible concept-based explanations for cnn models with non-negative concept activation vectors." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 13. 2021, pp. 11682–11690.

[Zha+25] Xiaoge Zhang, Tao Wang, Lei Ma, and Sankaran Mahadevan. "Reliability engineering, risk management, and trustworthiness assurance for AI systems." In: *Journal of Reliability Science and Engineering* 1.2 (2025), p. 022001.

[Zha+20] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. "Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation." In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer. 2020, pp. 365–381.

[Zha+15] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. "Appearance-based gaze estimation in the wild." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4511–4520.

[Zha+19] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. "" Why should you trust my explanation?" understanding uncertainty in LIME explanations." In: *arXiv preprint arXiv:1904.12991* (2019).

[ZLB20] Yunfeng Zhang, Q Vera Liao, and Rachel KE 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*. 2020, pp. 295–305.

[ZP14] Kang Zhao and Li Pan. "A machine learning based trust evaluation framework for online social networks." In: *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*. IEEE. 2014, pp. 69–74.

[Zho+22] Xinlei Zhou, Han Liu, Farhad Pourpanah, Tieyong Zeng, and Xizhao Wang. "A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications." In: *Neurocomputing* 489 (2022), pp. 449–465.

[ZC04] Afra Zomorodian and Gunnar Carlsson. "Computing persistent homology." In: *Proceedings of the twentieth annual symposium on Computational geometry*. 2004, pp. 347–356.

[Zou+20] Junhua Zou, Zhisong Pan, Junyang Qiu, Xin Liu, Ting Rui, and Wei Li. "Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting." In: *European Conference on Computer Vision*. Springer. 2020, pp. 563–579.

DECLARATION

I, Carina Newen, hereby confirm that this work is my own and is expressed in my own words. Any uses made within it of the works of other authors in any form (ideas, equations, figures, text, tables) are properly acknowledged at the point of their use. A full list of the references employed has been included.

Dortmund, Juli 2025

Carina Newen

This work has been financially supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Aliliance centers within the UA Ruhr (https://uaruhr.de).