A.I. Robustness

a Human-Centered Perspective on Technological Challenges and Opportunities

Tocchetti, Andrea; Corti, Lorenzo; Balayn, Agathe; Yurrita, Mireia; Lippmann, Philip; Brambilla, Marco; Yang, Jie

DOI
10.1145/3665926

Publication date
2025

Document Version
Final published version

Published in
ACM Computing Surveys

Citation (APA)
Tocchetti, A., Corti, L., Balayn, A., Yurrita, M., Lippmann, P., Brambilla, M., & Yang, J. (2025). A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities. *ACM Computing Surveys*, *57*(6), Article 141. https://doi.org/10.1145/3665926

A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities

ANDREA TOCCHETTI, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy
LORENZO CORTI, TU Delft, Delft, Netherlands
AGATHE BALAYN, TU Delft, Delft, Netherlands
MIREIA YURRITA, TU Delft, Delft, Netherlands
PHILIP LIPPMANN, TU Delft, Delft, Netherlands
MARCO BRAMBILLA, Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milano, Italy
JIE YANG, TU Delft, Delft, Netherlands

Despite the impressive performance of Artificial Intelligence (AI) systems, their robustness remains elusive and constitutes a key issue that impedes large-scale adoption. Besides, robustness is interpreted differently across domains and contexts of AI. In this work, we systematically survey recent progress to provide a reconciled terminology of concepts around AI robustness. We introduce three taxonomies to organize and describe the literature both from a fundamental and applied point of view: (1) methods and approaches that address robustness in different phases of the machine learning pipeline; (2) methods improving robustness in specific model architectures, tasks, and systems; and in addition, (3) methodologies and insights around evaluating the robustness of AI systems, particularly the tradeoffs with other trustworthiness properties. Finally, we identify and discuss research gaps and opportunities and give an outlook on the field. We highlight the central role of humans in evaluating and enhancing AI robustness, considering the necessary knowledge they can provide, and discuss the need for better understanding practices and developing supportive tools in the future.

CCS Concepts: • **Human-centered computing**; • **General and reference → Surveys and overviews**; • **Computing methodologies → Artificial intelligence**; **Machine learning**;

Additional Key Words and Phrases: Artificial intelligence, robustness, human-centered AI, trustworthy AI

Andrea Tocchetti, Lorenzo Corti, and Agathe Balayn contributed equally to this research.

This research has been partially supported by the TU Delft *Design@Scale* AI Lab, the *HyperEdge Sensing* project funded by Cognizant, the EU's H2020 project 101016233 PERISCOPE, the EU's H2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955990, and the Ph.D. Scholarship on Explainable AI funded by Cefriel. The work is also part of the ICAI GENIUS lab of the research program ROBUST (project number KICH3.LTP.20.006), partly funded by the Dutch Research Council (NWO).

Authors' Contact Information: Andrea Tocchetti, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Milano, Italy; e-mail: andrea.tocchetti@polimi.it; Lorenzo Corti, TU Delft, Delft, Netherlands; e-mail: L.Corti@tudelft.nl; Agathe Balayn, TU Delft, Delft, Zuid-Holland, Netherlands; e-mail: a.m.a.balayn@tudelft.nl; Mireia Yurrita, TU Delft, Delft, Zuid-Holland, Netherlands; e-mail: m.yurritasemperena@tudelft.nl; Philip Lippmann, TU Delft, Delft, Zuid-Holland, Netherlands; e-mail: p.lippmann@tudelft.nl; Marco Brambilla, Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milano, Italy; e-mail: marco.brambilla@polimi.it; Jie Yang (Corresponding author), TU Delft, Delft, Zuid-Holland, Netherlands; e-mail: j.yang-3@tudelft.nl.

ACM 0360-0300/2025/02-ART141
https://doi.org/10.1145/3665926

ACM Reference Format:
Andrea Tocchetti, Lorenzo Corti, Agathe Balayn, Mireia Yurrita, Philip Lippmann, Marco Brambilla, and Jie Yang. 2025. A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities. *ACM Comput. Surv.* 57, 6, Article 141 (February 2025), 38 pages. https://doi.org/10.1145/3665926

1 Introduction

AI systems show potential and are expected to revolutionize existing workflows by combining human- and non-human skills [21]. Yet, there is still little insight into how we should deal with the tradeoffs of combining human and artificial agency, or the way in which these systems should be assessed and held accountable [70]. Furthermore, concerns about bias [32], inscrutability [12], and vulnerability [101] have also been raised. Consequently, several social actors, like the European High-Level Expert Group, have highlighted the need for socio-political deliberation around the design and governance of AI systems, and have defined principles for Trustworthy AI, i.e., the *Ethics Guidelines for Trustworthy AI* [190].

One of the core principles of Trustworthy AI is robustness [70], defined in **Machine Learning** (**ML**) as *the insensitivity of a model's performance to miscalculations of its parameters* [158, 273]. Examples like Tesla's Full Self-Driving mechanism erroneously identifying the moon as a yellow traffic light,[1] or Autopilot being fooled by stickers placed on the ground,[2] show that AI systems might be susceptible to errors and vulnerable to external attacks. This may result in undesired behavior and decreased performance [255]. Given the application of AI systems in safety-critical areas (e.g., medical diagnosis [23]), it is paramount to design reliable systems, so that they can be properly and safely integrated in the context of use. In response to this need, a growing body of literature focuses on developing and testing robust AI systems. Methodologies toward robust AI have addressed every phase of the ML pipeline, going from data collection and feature extraction, to model training and prediction [255]. Such methodologies have also been applied to a wide range of tasks and application areas, including (but not limited to) image classification [216] and object detection [44] in Computer Vision, or text classification in **Natural Language Processing** (**NLP**) [116].

Considering the increasing efforts devoted to this field within Trustworthy AI, in this article, we seek to analyze the progress made so far and give a *structured* overview of the suggested solutions. Furthermore, we also aim to identify the areas that have received less attention, highlighting research gaps, and projecting into future research directions. Our work differs from similar efforts in three main ways. (1) As opposed to some previous work [37, 81, 255], we do not limit the scope of our analysis to adversarial attacks. We argue that, as suggested by Drenkow et al. [64] or Liu et al. [199], natural (i.e., non-adversarial) perturbations constitute a common real-world menace that needs further attention. (2) As far as the application area is concerned, and contrary to surveys solely focusing on tasks like Computer Vision [64] or architectures like **Graph Neural Networks** (**GNNs**) [199], we do not limit our survey to any technology in particular. We rather conduct our search in a task-agnostic way. Such an approach helps us identify the most prominent trends within the field and compare the differences in effort and interest across applications as part of our survey. (3) Most importantly, as opposed to previous work, which has predominantly focused on surveying algorithm-centric solutions to AI robustness, we adopt a human-centered perspective by *additionally* including terms like *human computation* or *human knowledge* (see Section 2.1) to our search. We find that, even if there are already a few studies that implicitly involve humans

[1] https://www.autoweek.com/news/green-cars/a37114603/tesla-fsd-mistakes-moon-for-traffic-light/ (access 13 October 2022)
[2] https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf (access 13 October 2022)

(e.g., crowdworkers, ML practitioners) in their pipelines (e.g., for ML diagnosis), these represent a minority compared to algorithm-centric approaches. Furthermore, the studies that *do* include humans in the loop tend to disregard the challenges that human-led approaches face *in practice*. We, therefore, emphasize the potential that **human-in-the-loop** (**HIL**) approaches have for improving AI robustness, while we highlight the need to understand human-led practices in order to integrate robustness into existing workflows and tools. To inform such research opportunities, we advocate for a multidisciplinary approach and bring insights from human-centered fields, such as **explainable AI** (**XAI**), crowd computing, or HIL ML. We, therefore, make the following contributions:

(1) We summarize the main concepts around robust AI (Section 3). We consolidate the terminology used in this context, disentangling the meaning and scope of different constructs. We pay special attention to identifying the commonalities and differentiating aspects of the used terms.
(2) We systematically summarize around 370 papers on robust AI and related concepts (Section 2) and arrange them in three different taxonomies. First, we group papers that improve robustness by addressing different aspects of the ML pipeline (Section 4). We identified three main aspects that the selected studies work on input data, in-model attributes, and model post-processing aspects. Second, we discuss prior work that made progress in improving robustness for specific architectures (e.g., GNNs), specific tasks (i.e., NLP and Cybersecurity), and systems conceived within other fields of Trustworthy AI (i.e., explainable and fairness-aware systems) (Section 5). We focus on these particular architectures, systems, and fields as they have comparatively received little attention in previous surveys despite the importance of robustness as a desired property. Third, we create a taxonomy related to the assessment (e.g., through benchmarking or empirical studies) of robust AI systems (Section 6).
(3) We identify and discuss disparate research efforts in each of the established fields and identify research gaps. Specifically, we make a special in-depth analysis of the opportunities brought by one of the identified research gaps: the absence of human-centered work in existing methodologies (Section 7). We highlight the multidisciplinary nature of the robust AI field and provide an outlook for future research directions, bringing insights from human-centered fields (Section 8).

2 Survey Methodology: Paper Collection

In this chapter, we detail the process applied to collecting the final list of articles considered in this literature review. This includes keyword collection and curation, querying multiple databases, de-duplication, manual filtering, tagging, and analysis.

2.1 Collecting Papers

Defining Keywords. First, we curated the list of keywords to be used for querying articles. We inspected key definitions of robustness and robust AI [39, 85, 174] in the context of Computer Science and organized a preliminary list. We further enriched this list such that it covers aspects related to the trustworthiness of AI systems and to human-centeredness (including human knowledge) given the lack of a common viewpoint on robustness. Table 1 shows the complete list of keywords used.

Querying Publication Databases. Secondly, we queried multiple bibliographical databases by generating all possible triples of keywords based on the groups we defined, e.g., "Robustness" AND "Artificial Intelligence" AND "Explainability", finally leading to 156 unique search queries. Articles

Table 1. The Groups of Keywords Considered in the Data Collection Process and the Corresponding Keywords

Group Name	Keywords
Fundamental	Robustness, Robust
Scope	Artificial Intelligence, Machine Learning, Neural Network
Context	Trustworthy, Stability, Resilience, Reliability, Accountability, Transparency, Reproducibility Accuracy, Confidence, Performance, Design, Adversarial, Unknowns, Noise Human Computation, Human Knowledge, Human-In-the-Loop, Human Interpretation, Knowledge Base, Knowledge, Knowledge Elicitation, Reasoning Explainability, Explanation, Interpretability, Interpretable

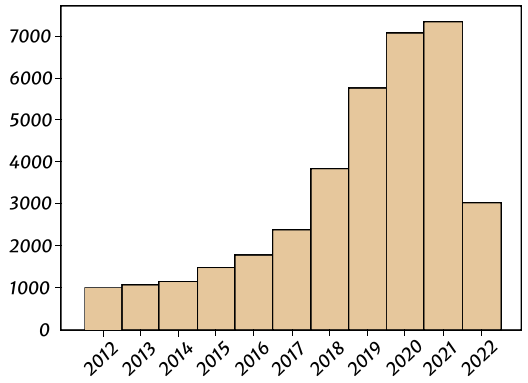

Fig. 1. Temporal distribution of the 35,800 unique papers published in the last 10 years. A growing trend of published papers about Robust AI over the years was observed. The amount of papers collected in 2022 is not to be considered relevant to this trend as the data was collected in July 2022.

have been collected in July 2022 through *Publish or Perish*[3] by querying the following supported bibliographical databases: Google Scholar, Scopus, Semantic Scholar, and Web of Science. Moreover, given the breadth of the literature on trustworthy and robust AI, we inspect literature from the last 10 years, i.e., articles published between January 2012 and July 2022.

2.2 Filtering Papers

Pre-filtering. We collected about 100,000 papers distributed as follows: 31,000 from Google Scholar, 18,450 from Scopus, 30,800 from Semantic Scholar, and 19,400 from Web of Science. Considering the breadth of the data collection, we sought to remove any duplicate entries in our results. Papers that had the same title and authors were filtered out, resulting in 45,400 papers. Duplicates that were undetected at this stage were discarded in the later ones. Then, papers published before the period of interest (January 2012 to July 2022) were filtered out, leading to 35,800 articles. Figure 1 displays the time distribution of the collected papers. We observe a growing interest in the considered topics over the years, which (partially) motivates the time constraints applied.

Further Inspecting Papers. At this stage, we manually inspected the abstracts of the collected papers to exclude the ones whose context or content require domain-specific expertise (e.g., healthcare), or deal with a notion of "robustness" that is not related to ML (e.g., signal processing). We ended up with 1,800 papers. While inspecting papers, we marked them with specific keywords (e.g., "Computer Vision" or "Loss Function"), to differentiate them in terms of content and type

[3]Harzing, A.W. (2007) Publish or Perish, available from https://harzing.com/resources/publish-or-perish

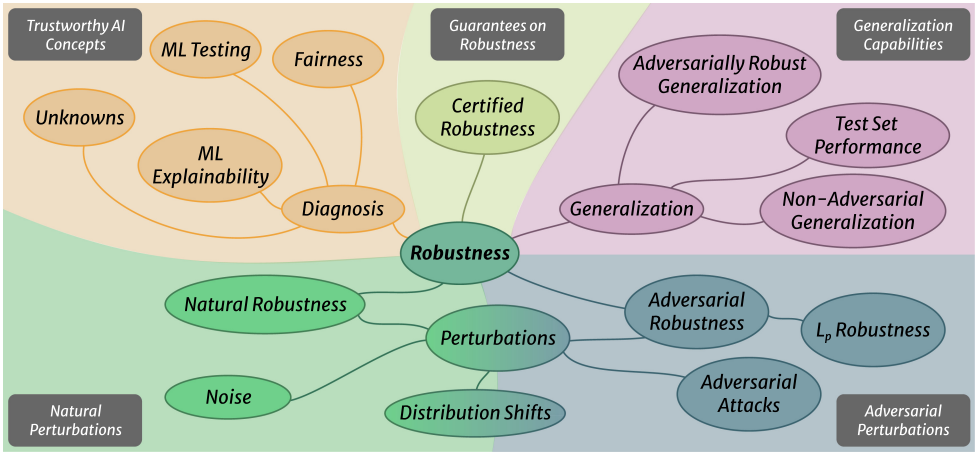

Fig. 2. Main concepts found through our analysis of the literature on Robust AI.

of publication (e.g., "Literature Review"). Consequently, we used those keywords to perform a final filtering step in which the papers tagged with the least frequent keywords (i.e., appearing only once) were excluded. Omitted keywords include: "audio signal" and "event detection". Throughout the entire process, we carefully analyzed the papers such that they contained significant or late progress in the area. These include 94.1 % papers published in peer-reviewed venues, 1.9 % non-archived peer-reviewed papers (i.e., accepted in workshops with no proceedings and published on arXiv), and 4 % non-peer-reviewed papers (i.e., only published on arXiv). The non-peer-reviewed papers in this survey have at least 50 citations if they were written before or in 2019 or at least 15 citations if they were written after 2019. In the end, this thorough inspection led to 560 papers that were systematically analyzed, out of which around 370 papers were systematically summarized and discussed.[4] The list of collected, filtered, and summarized papers can be found on GitHub.[5] We applied the same criteria when selecting and filtering additional papers for the discussion sections (Sections 7 and 8).

3 Overview of the Main Concepts Surrounding Robustness

From our collection of papers, we evinced that the notion of *Robustness* is ill-defined. A number of ML sub-domains refer to robustness from different viewpoints. We clarify the relations between these domains in Section 3.1. We also identify that a number of concepts directly related to robustness are used in different ways across research papers (Figure 2). We disambiguate the interpretation of related terms in Section 3.3. Finally, our analysis of the papers surfaced a few recurring themes, introduced in Section 3.4, and used to organize our survey for which the structure and primary references are summarized in Table 2.

3.1 The Various Definitions of Robustness

Given the broadness of the literature on robustness and the variety of contexts in which it is considered, addressed, and analyzed, we discuss and provide a common ground about the definitions of robustness and its associated concepts. Particularly, robustness is generally defined as *the insensitivity of a model's performance to miscalculations of its parameters* [158, 273], with Nobandegani

[4]Due to space limit, we leave the discussions of some papers (about 30%) in the supplementary material.
[5]https://github.com/AndreaTocchetti/ACMReviewPaperPolimiDelft.git

Table 2. Outline of the Article Structure and Corresponding References

Category	Classification	Sec.	Topic	References
Methods & Approaches for Improving Robustness	Processing the Training Data	4.1	Generating Adversarial Attacks	[1, 38, 38, 42, 45, 105, 105, 242, 252]
			Augmenting Data for Adversarial Robustness	[1, 4, 27, 41, 46, 56, 74, 118, 218, 226, 232, 240, 285]
			& for Non-Adversarial Robustness	[41, 74, 118, 157, 167, 173, 246, 277, 285]
	Designing In-Model Robustness Strategies	4.2	Improving Robustness through Training	[48, 49, 86, 88, 134, 135, 155, 186, 215, 225, 230, 245, 253, 268, 289] (adversarial training), [9, 14, 31, 40, 68, 92, 111, 113, 124, 126, 127, 137, 145, 151, 164, 184, 212, 222, 258, 260, 264, 287] (others)
			Improving Robustness through Architecture Design	[106, 108, 148, 256] (tweaking network layers), [50, 133, 197] (inherent robustness of networks), [57, 128, 135, 147] (searching network architectures)
	Leveraging Model Post-Processing Opportunities	4.3	Identifying Unnecessary or Unstable Model Attributes	[42, 73, 129]
			Fusing Models	[51, 172, 183, 218, 250, 266]
Robustness in Practical Fields	Robustness for Specific Architectures	5.1	Graph Neural Networks	[29, 47, 71, 77, 107, 132, 168, 239, 276]
			Bayesian Neural Networks	[36, 142, 231]
	Robustness for Specific Application Areas	5.2	Robustness for Natural Language Processing	[42, 66, 125, 172, 264, 284, 286]
			Robustness for Cybersecurity	[1, 2, 6, 8]
	Robustness for Specific Trustworthy AI Concepts	5.3	Robustness for Explainability	[5, 13, 15, 60, 117, 153, 166, 237, 246, 278]
			Robustness for Fairness	[3, 15, 178, 244, 269]
Robustness Assessments & Insights	Evaluation Strategies	6.1	Evaluation of Robustness	[20, 65, 67, 76, 83, 91, 102, 116, 121, 122, 136, 185, 203, 205, 227, 247, 263, 265, 274, 280, 281]
			Benchmarks	[54, 58, 63, 82, 90, 139, 159, 169, 223, 283]
			Metrics	[35, 115, 203, 219, 249, 254, 263, 267] (adv. robustness), [10, 24, 119, 162] (adv. attacks)
	Studies around Proposed Robustness Methods & Insights	6.2	Insights on Adversarial Robustness	[103, 192] (comparisons), [195, 210, 228] (inner model), [116, 150, 189, 243] (perturbations)
			Insights on Natural Robustness	[22, 288] (noise), [33, 59] (shifts)
	Trade-Offs Between Robustness and Other Trustworthy AI Concepts	6.3	Trade-Off with Accuracy	[144, 175, 216, 229]
			Trade-Off with Fairness	[26, 171, 257]
			Trade-Off with Explainability	[160, 252]

et al. [158] stating that *robust models should be insensitive to inaccuracies of their parameters, with little or no decline in their performance.* Two main robustness branches have been identified: robustness to adversarial attacks or perturbations, and robustness to natural perturbations.

3.1.1 Adversarial Robustness. Adversarial Robustness refers to the ability of models to maintain their performance under potential adversarial attacks and perturbations [283]. Adversarial perturbations are imperceptible, non-random modifications of the input to change a model's prediction, maximizing its error [221]. The result of such a process is called an adversarial example, i.e., an input x' close to a valid input x according to some distance metric (i.e., similarity), whose outputs are different [38]. Such data is employed to perform adversarial attacks, whose objective is to find any x' according to a given maximum attack distance [44]. The literature presents different classifications of adversarial attacks: targeted and untargeted [43], and white-, grey-, or black-box [149]. Targeted attacks generate adversarial examples misclassified as specific classes, while untargeted attacks generate misclassified samples in general. The main difference between white-, grey-, and black-box attacks is the attacker's knowledge about the model or the defense mechanism.

A similarity metric is often defined when generating attacks or evaluating robustness. Depending on the input domain, different metrics can be applied. These metrics are built as a function of a parameter (usually denoted with the letter p) whose value influences its computation. For example, Carlini et al. [38] define a generic p norm from which different metrics with different meanings are derived. In their case, when $p = 0$ (L_0 distance), the number of coordinates for which the valid and perturbed input are different is measured; when $p = 2$ (L_2 distance), the standard Euclidean distance between the valid and perturbed input is computed; when p = infinite (L_∞ distance), the maximum change to any coordinate is measured. A particular type of robustness is Certified Robustness that guarantees a stable classification for any input within a certain range [52].

3.1.2 Natural Robustness. Natural Robustness (a.k.a. Robustness against natural perturbations) is the capability of a model to preserve its performance under naturally-induced image corruptions or alterations [64]. Natural Perturbations (a.k.a. Common Corruptions [90] or Degradations [79]) are introduced through different types of commonly witnessed natural noise [242] (e.g., Gaussian noise in low lighting conditions [90]), and represent conditions more likely to occur in the real world compared to adversarial perturbations [64]. Temporal Perturbations are natural perturbations that hinder the capability of a model to detect objects in perceptually similar, nearby frames in videos [194]. All these perturbations result in a condition where the distribution of the test set differs from the one of the training set [112]. This condition is typically referred to in the literature with overlapping concepts, namely distribution shift [59, 224], **Out-of-Distribution** (**OOD**) data [80, 199], and data outside the training set [167].

3.2 Other Robustness-Related Terms

3.2.1 Generalization. Generalization is another widely used term in the robustness literature. In general, it is defined as the model's performance on unseen test scenarios [165] or as the closeness between the population (or test error) to the training error, even when minimizing the training error [156]. Two other types of generalization are also reported: adversarially robust [271] and non-adversarial generalizations [80, 167, 251, 282]. While the first one refers to the capability of a model to achieve high performance on novel adversarial samples, the second one is evaluated on non-adversarial samples (e.g., natural perturbations [251, 282], distribution shifts [80, 167]).

3.2.2 Performance. Across the inspected literature, the term performance is employed with a broad variety of meanings. Depending on the aspect of interest, it may refer to accuracy [64], robustness [118], runtime [203], or precision [263]. Given such variety, the actual meaning of performance will be addressed only when relevant to understand the concepts explained in the core survey.

3.3 Domains Adjacent to Robustness

ML explainability, fairness, trustworthiness, and testing are four research domains recurring across robustness literature. While there is no agreed upon definition of each of these fields and their goals, and we acknowledge it is not possible and desirable in the scope of this survey to provide a complete overview of these fields, we provide here explanations that are sufficient to understand the relation these fields bear to robustness.

3.3.1 Explainability. ML explainability is the field interested in developing post-hoc (explainability) methods and (inherently explainable) models that allow the internal functioning of ML systems to be understandable to humans [39]. We identify three types of relations between the explainability and robustness fields. A number of papers investigate how explainability methods can be used in order to *enhance the robustness* of models (see Section 7.1.2). Another set of papers

investigates *how robust existing explanability methods are* to various types of perturbations (see Section 5.3.1). A last set of papers instead studies how existing methods for enhancing robustness *trade off* with the explainability of the models, and especially with the alignment between the model features, and the features a human would expect the model to learn (see Section 6.3.3).

We also consider the field of *(un)known unknowns* [138] close to robustness, as they are typically caused by OOD samples. In this field, methods to identify and mitigate the presence of such unknowns are developed and, while these methods typically fall within explainability [196, 233], they are directly applicable to increase the robustness of a model.

3.3.2 Fairness. ML fairness in the broad sense is the field interested in making the outputs of an ML model non-harmful to the humans who are subject to the decisions made based on these outputs. Researchers in this field have developed a number of fairness metrics [236] and methods for mitigating unfairness [140]. We identify two types of relations between this field and robustness, similar to the relations between explainability and robustness: *robustness of fairness metrics and methods* to different types of natural and adversarial perturbations (see Section 5.3.2) and *tradeoffs* caused by the application of robustness methods (see Section 6.3.2).

3.3.3 Testing. ML testing [275] is a field emanated from software testing. It consists in developing methods and tools to identify and characterize any discrepancy between the expected and actual behavior of an ML model. While this field bears a broader scope, since brittleness to different perturbations represents one of the many types of unexpected behavior of a model, it is also narrow as it is solely interested in detecting the issue, but not its mitigation. Naturally, methods developed in this field could potentially be adapted in the future to better detect robustness-related issues.

3.4 Themes in Relation to These Robustness Definitions and Related Domains

Analyzing the collected publications through a thematic analysis approach [30], we iteratively and collaboratively identified three primary themes and three recurring categories within each of these themes (nine categories in total) that were deemed worth emphasizing (summarized in Figure 3).

3.4.1 Methods and Approaches for Improving Robustness. The most studied methods to achieve robustness are described in Section 4. They are categorized according to the stage of the ML pipeline to which they apply, that is either the processing of the training dataset, the model creation stage, or the post-processing of the trained model. Within each of these stages, the approaches vary across publications, and were further clustered into groups based on types of robustness (e.g., adversarial or natural perturbations), and specific ML components (e.g., training procedure or model architecture) they apply to. For each of the groups, we further delve into sub-groups based on the types of transformation applied to the component (e.g., different loss functions or regularizers), and describe the main similarities and differences across transformations (e.g., in terms of technical approach and performance).

3.4.2 Robustness in Practical Fields. While a majority of papers concentrate their studies and the evaluation of their robustness methods around computer vision or do not mention a specific field, we also identify a consequent number of papers that bear different focuses. We separated these papers from the ones discussed above, because they present particularities that are worth investigating. We categorize these papers broadly based on their research fields, and discuss them in Section 5. Within each of the categories, we describe the most researched sub-types for which we retrieved the most literature. Particularly, we identified focuses relating to specific model types (GNNs and Bayesian Learning), specific application areas (NLP, and Cybersecurity), and specific concepts within the trustworthy AI domain (explainability and fairness). The latter is particularly interesting because it differs from other works in its objectives. Contrary to all other papers which

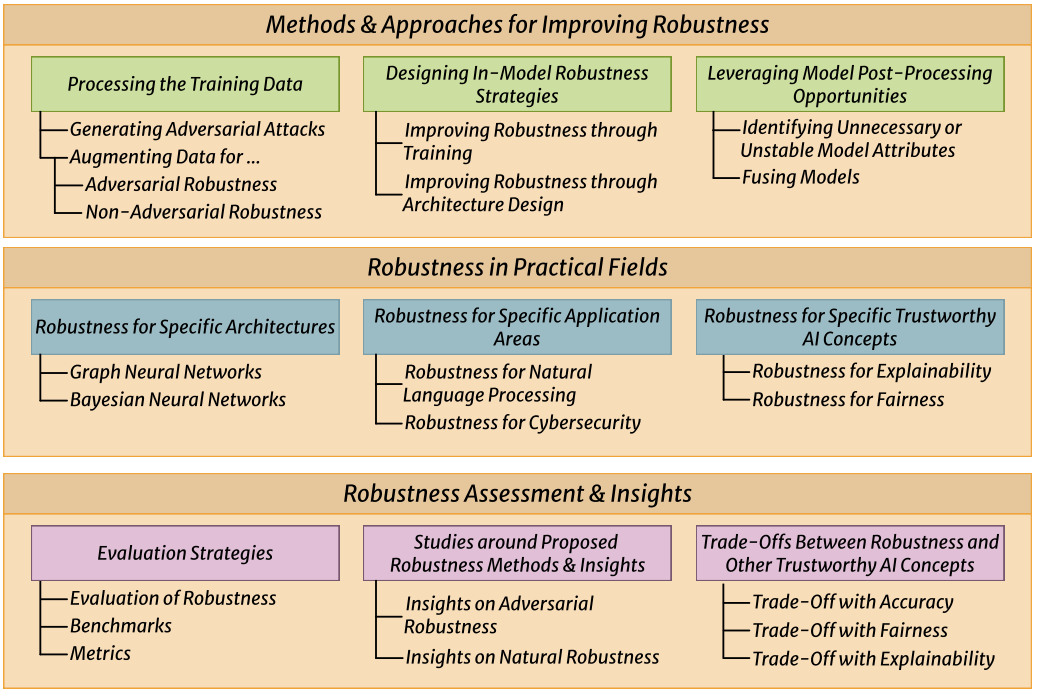

Fig. 3. The three themes and their sub-categories that shape our survey.

investigate model performance under perturbations, it instead investigates the evolution of fairness and explanations of a model under the effect of perturbations.

3.4.3 Robustness Assessment and Insights. The last theme we identified, described in Section 6, revolves around the assessment of robustness of a system. Particularly, the importance of developing procedures (methodologies, benchmarks, and metrics) to evaluate robustness emerged from the analysis and these procedures revealed to vary greatly across publications (be it publications whose primary contribution is an evaluation procedure, or a robustness method that requires to be evaluated through a defined procedure). We also identified a set of publications whose primary objective is to perform studies to evaluate existing robustness methods and collect insights to further characterize in which conditions each type of method performs best. Finally, the last recurring theme was tradeoffs, as many papers that propose or evaluate robustness methods tackle tradeoffs while striving to achieve other objectives, be it the model performance or the other trustworthy AI concepts identified earlier. The publications in this section of the survey are typically falling under the umbrella of computer vision publications, or of the different fields highlighted above.

4 Methods and Approaches for Improving Robustness

A large fraction of the literature is devoted to fundamental methods to improve the robustness of AI models across their lifecycle: training data augmentation with malicious samples, ad-hoc training procedures and architectures, and post-training pruning and model fusion.

4.1 Processing the Training Data

With the final aim of improving model robustness against adversarial attacks, noise, or common perturbations, several approaches focus on generating perturbations to perform data augmentation.

4.1.1 Generating Adversarial Attacks. A number of papers tackle the challenge of developing methods to generate adversarial attacks that prove deep learning models brittle. The proposed methods vary with regard to three main objectives. (a) The type of task targeted (e.g., NLP model [42, 105], image classification [38], or object detection models [45]). (b) The type of constraints imposed on the attack: attack on the physical space before capturing the digital data sample (e.g., by sticking images patches on the physical object to be recognized [242], or by processing this digital input sample [45]); general attack or attack that targets a particular component of the model (e.g., rationalizers of rationale models [42]); attacks that preserve certain properties of the input sample such as human consistency (e.g., Jin et al. [105] talk about human prediction consistency, semantic similarity, and text fluency with regard to the generated adversarial text samples), additionally to satisfy the constraint on similarity to the original sample [105]. (c) The type of brittleness targeted: the model makes a different (wrong) prediction when the transformed sample is inputted, or the explanations of the prediction also becomes flawed (i.e., the identified important features are not the correct ones) [252]. The works then differ by the approach taken to generate the attacks, be it through different optimization instances (objective functions) they use to find adversarial instances that fit the problem [38, 45], by leveraging **Generative Adversarial Networks (GANs)** [1], or through a rule-based algorithmic approach [42, 105].

4.1.2 Augmenting Data for Adversarial Robustness. Most of the identified literature focuses on transforming [27, 41, 226, 285], generating [4, 46, 118, 218, 240], or employing ready-to-use [56] data and/or adversarial samples to extend or create datasets to train more robust models. Such a data augmentation process can successfully improve adversarial robustness [41, 56, 118, 218, 226, 240, 285] and adversarial accuracy [4], while sometimes reducing time costs [41], and adversarial attack success rate [27]. When defending against adversarial attacks, GAN-based solutions are proven useful in achieving such an objective [1, 74, 218, 240]. In particular, they are employed to generate adversarial samples [1], perturbations [240], and boundary samples [218] to defend the networks against adversarial attacks. While most methods apply complex transformations to improve robustness, simple transformations, like rotation [226] and image background removal [232], are still proven effective. However, extending the training set is not always enough by itself. Hence, ad-hoc training procedures [41, 46] must be set in place to select [46] and adapt [41] the optimal training data to achieve adversarial robustness.

4.1.3 Augmenting Data for Non-Adversarial Robustness. Not all researchers aim to enhance models' defense against adversarial robustness. Noise [173], non-adversarial perturbations [74, 118, 157, 285], spurious correlations [41, 246], and distribution shift [167, 277] hinder the performance and resilience of models. In tackling such impairments, human rationale collection allows the generation of new datasets [167], counterfactual-augmented data [41] and the definition of proper perturbation levels [157], consequently improving performance [41], and model [157], and distributional shift [167] robustness. Custom [173] and pre-existing approaches are applied to perform data augmentation, consequently improving noise robustness [173] and performance [173]. On the other hand, data transformation [74, 285] and training [118] approaches are applied to improve model robustness [74, 118, 285] and reduce training time [74].

4.2 Designing In-Model Robustness Strategies

4.2.1 Improving Robustness Through Training. Training plays an integral part in creating ML models. Concerning robustness, Adversarial Training is the de-facto standard for building robust models. The core intuition behind it is to complement natural data with perturbed one such that models incorporate information about data that better represent real-world scenarios' variability. In this section, we discuss adversarial training approaches that adaptively change

the perturbation magnitude, allow for the learning of robust features, or include novel loss or regularization functions. Finally, we discuss approaches alternative to adversarial training.

Adversarial Training. Adversarial Training has proven to be a fundamental tool to build robust models and that is reflected in the amount of literature available for it: researchers have focused on improving the whole process and proposed a plethora of algorithms [86, 135, 215, 225, 245], borrowing different ML paradigms like self-supervised and unsupervised learning [155, 230], that are applicable to a variety of tasks (e.g., content recommendation [253, 268]). In this context, **Projected Gradient Descent (PGD)** [135] is a common white-box (i.e., the attacker knows everything about the model) algorithm. On the same note, Terzi et al. [225] and Gupta et al. [86] propose extensions of PGD by using Wasserstein distance in the adversarial search space, by replacing the initial adversarial training stages with *natural* training, or by encouraging the logits from clean examples to be similar to their adversarial counterparts, respectively. While training models with *natural* perturbations was proven effective in improving robustness, researchers demonstrated that generating and employing *dynamic* perturbations is another effective way of building robust models. Madaan et al. [134] and Cheng et al. [49] propose methods to generate *dynamic* perturbations at the level of single data instances that are then controlled by enforcing label consistency in the former case, and smooth labels in the latter. Differently, Rusak et al. [186] devise a neural network-based adversarial noise generator to tackle the online generation of perturbations. On the other hand, several works focus on leveraging other types of information. For example, Zoran et al. [289] adversarially train and analyze a neural model incorporating a human-inspired, visual attention component guided by a recurrent top-down sequential process. Shifting to model outputs, works from Wang et al. [245] and Stutz et al. [215] focus on differently treating misclassifications and rejecting low-confidence predictions. Similarly, Haase-Schütz et al. [88] and Cheng et al. [48] deal with progressively tuning labels starting from unlabelled data and through smoothing, respectively.

Beyond Adversarially Training. Other than directly employing enriched data to perform adversarial training, researchers devised other methods to enhance model robustness. These include techniques such as learning more robust feature representations, and training models through adapted regularizers and loss functions. Scholars drove models to learn robust feature representations in multiple ways, from designing novel methods altogether [113] to employing additional classifiers [14]. For example, Yang et al. [264] to apply perturbations on textual embeddings such that the corresponding words would be drawn toward positive samples rather than adversarial ones. Bai et al. [14] take a modelling approach to obtain robust features through the addition of auxiliary models to identify which channels in **convolutional neural networks (CNNs)** are more robust. Regularization is another tool that ML engineers can use when building models and, as such, it has also been used to make them more robust. Li and Zhang [124] propose a PAC-Bayesian approach to tackle the memorization of training labels in fine-tuning. Chan et al. [40] suggest an approach that optimizes the saliency of classifiers' Jacobian by adversarially regularizing the model's Jacobian to resemble natural training images. Concerning the usage of adapted loss functions for robustness, various functions were used to incorporate specific objectives: triplet loss [137], minimizing distance between true and false classes [127], mutual information [258], consistency across data augmentation strategies [222], perturbation regularizers [260], adding maximal class separation constraints [151], combining multiple losses [111] (e.g., Softmax and Center Loss), or approximating existing losses (e.g., Categorical Crossentropy) [68]. It is worth noting that loss functions tailored for robustness are not exclusive to models trained in isolation, and robust and natural models (acting as regularizers) can be jointly trained [9]. Conversely to these methods, researchers have studied alternative training procedures to adversarial training. Staib and Jegelka [212] has analyzed the relationship between adversarial training and robust optimization, proposing a

generalization of the former, which leads to stronger adversaries. Attention is also directed to leveraging input and output spaces. Li et al. [126] consider training robust models by leveraging the adversarial space of another model. Differently, Mirman et al. [145] and Rozsa et al. [184] leverage abstract interpretation and evolution stalling, respectively. The former generates abstract transformers to train certifiably robust models. The latter progressively tempers the contributions of correct predictions toward the loss function. Finally, Mirman et al. [145], Zi et al. [287], and Papernot et al. [164] leverage Distillation (a knowledge transfer technique in which a smaller model is trained to mimic a larger one) [31, 92] to obtain robust models.

4.2.2 Improving Robustness Through Architecture Design. Researchers have also investigated ways to make neural models robust from an architectural perspective.

Tweaking Neural Network Layers. We identified that a considerable amount of effort is directed toward Computer Vision applications, with many solutions aimed at integrating additional mechanisms of Convolutional Networks to enhance their robustness. Many adversarial attackers create harmful data instances by injecting noise perturbations in the input of the model. In line with this, many researchers have attempted to introduce mechanisms that take advantage of this information or directly try to mitigate the repercussions of such perturbations. For example, Jin et al. [106] introduce additive stochastic noise in the input layer of a CNN and re-parametrize the subsequent layers to take advantage of this additional information. Alternatively, Momeny et al. [148] introduce a CNN variant that is robust to noise by adapting dynamically both striding of convolutions and the following pooling operations. Work by Xu et al. [256] operate on the classification layer by constraining its weights to be orthogonal. Operating on network layers is not exclusive to the aforementioned discriminative models, but it has also found applications for generative models. For example, Kaneko et al. [108] propose a method to obtain GANs that do not require a large amount of correctly-labeled instances but still maintain a consistent behavior. They do this by integrating a noise transition model that maps clean and noisy labels which leads to GANs that are resilient to different magnitudes of label noise.

Leveraging the Inherent Robustness of Spiking Neural Networks (SNN). In parallel to such enhancements at the architectural level, a growing trend is represented by SNN [133]. SNNs are a particular type of neural network that mimics the behavior of biological neurons by incorporating the notion of time and both operating with and producing sequences of discrete events (i.e., spikes). Concretely, a neuron in a SNN transmits information only when its value surpasses a certain threshold. This particular kind of neural network was found to be inherently robust to certain types of adversarial attacks. Sharmin et al. [197] test SNNs directly against gradient-based (black-box) attacks and find that such architectures perform better than non-spiking counterparts without any kind of adversarial training. Inspired by neuroscience, Cheng et al. [50] formulate Lateral Interactions (i.e., intra-layer connections) for SNNs which provide both better efficiency when processing a series of spikes as well as better resistance to injected Gaussian noise.

Searching Neural Architectures. Connected to handcrafting robust neural architectures, scholars have started applying **Neural Architecture Search** (**NAS**) to such a problem. In general, NAS is an automatic procedure aimed at discovering the best architecture (e.g., in terms of accuracy) for a neural network for a specific task. Devaguptapu et al. [57] analyze the effects that a varying amount of parameters have on adversarial robustness: while NAS can be an alternative to adversarial training, handcrafted models are more robust on large datasets and against stronger attacks like PGD [135]. Their insights motivate other works in this space, that focus on strengthening NAS approaches by including different forms of regularization on the smoothness of the loss landscape [147], or the sensitivity of the network [61, 97]. A different take on using NAS is the one

of Li et al. [128]: architecture search was blended with existing models (e.g., ResNet) to find the minimal increase in model capacity allowing it to withstand adversarial attacks.

4.3 Leveraging Model Post-Processing Opportunities

Robustness can also be improved through methodologies applied after training the model.

4.3.1 Identifying Unnecessary or Unstable Model Attributes. Pruning (i.e., the act of removing neurons and/or connections from a model) has become a popular compression approach that aims at reducing the computational cost of training models [129]. Recent literature in Robust AI has explored the use of pruning techniques or methodologies inspired by pruning to enhance model robustness [42]. Chen et al. [42], for instance, design a methodology for selectively replacing ReLU neurons that are identified as unstable (i.e., neurons that operate in the flat area of the function) and insignificant with linear activation functions that help improve robustness at a minimal performance cost. In a similar vein, additional mechanisms have been suggested for dealing with unnecessary and/or unstable system attributes. For instance, Gao et al. [73] introduce DeepCloak, a novel method to detect and remove unnecessary classification features in deep neural networks, consequently reducing the capabilities of attackers to generate such attacks.

4.3.2 Fusing Models. Another approach for achieving post-model-training robustness consists of plugging additional models into a trained model. These additional models can be used to identify and deal with problematic data instances (e.g., OOD, mistaken [172], noisy [183], or adversarially modified [266] occurrences). In the first case, in the context of NLP, Pruthi et al. [172] attach a task-agnostic word recognition model to a classification model as a means to defend the main classifier against spelling mistakes. In the context of Computer Vision, Ye et al. [266] use an additional classifier to determine real vs. adversarially manipulated data instances. This additional classifier would receive an overlap of the data instance and its saliency map. Furthermore, applying model fusion to infected models allows comparing the robustness of small models with respect to compression techniques [250]. A prominent line of work in this field consists in using GANs as auxiliary models. This strategy has been used for dealing with input data [218] and models [51]. For the former, Sun et al. [218] use a Boundary Conditional GAN to generate boundary samples. These samples have true labels and are near the decision boundary of a pre-trained classifier. For the latter, Choi et al. [51] propose **Adversarially Robust GAN** (**ARGAN**) that trains the generator model to reflect the vulnerability of the target neural network model against adversarial examples and hence optimizes its parameter values.

5 Robustness in Practical Fields

While in Section 4, we discussed literature that improved AI robustness by working on different phases of the ML pipeline, in the current section, we discuss prior work that made progress in improving robustness for specific model architectures, tasks, and systems. We found a number of methods being tailored for specific AI architectures, addressing domain-specific needs (e.g., word spelling for NLP), or bridging the gap with non-functional requirements of Fairness and Explainability. These less explored settings are later discussed in 7.1.

5.1 Robustness for Specific Architectures

5.1.1 Graph Neural Networks. A number of papers investigate how to increase the robustness of specific types of model architectures. One of the most prominent ones is GNNs, given their high susceptibility to small adversarial perturbations. For example, on the problem of link prediction on knowledge graphs, Pezeshkpour et al. [168] propose an attack strategy aimed at finding the minimal perturbation necessary to produce a label change. Differently, the approach by Lou

et al. [132] determines controllability and connectivity robustness (i.e., how well a system can keep its connectedness and controllability against node- or edge-removal attacks) by compressing the high-dimensional adjacency matrix before feeding it to a CNN to perform the robustness prediction. Fox and Rajamanickam [71] investigate the impact of structural noise on the robustness of GNN and find them to be weak to both local and global structural noise. Geisler et al. [77] focus on particularly large graphs and devise both high-efficacy attack and low-memory footprint defense strategies, enabling works on large networks at scale. Finally, attention has also been paid to formally certifying the robustness of GNN [29, 239].

Interestingly, there have also been several proposals for new GNN frameworks with better robustness characteristics by design. For example, Jin et al. [107] establish a framework to jointly learn clean graph structures from perturbed ones as well as the parameters for a GNN that is robust to adversarial attacks by preserving selected low-rank, sparsity, and feature smoothness properties. Instead, [47] devised a framework that leverages similarity metrics and adaptive regularization techniques to jointly learn graph structure and graph embeddings. Differently, Zhang and Lu [276] introduce a framework where robustness to noise is achieved by means of an auxiliary, (node-level) masked model for neighborhood aggregation.

5.1.2 Bayesian Neural Networks. Many adversarial attack strategies are based on identifying directions of high variability. Since such variability can be intuitively linked to uncertainty in the prediction, **Bayesian Neural Networks** (**BNNs**) are naturally of interest for robustness research. Similarly, Carbone et al. [36] analyze BNN to show that they are robust to gradient-based attacks. Vadera et al. [231] focus on different inference methods and attacks whose goal is leading the model misclassifications, finding that Markov Chain Monte Carlo inference has excellent robustness to a variety of attacks. Finally, Miller et al. [142] aim to evaluate robustness by extracting label uncertainty from the object detection system via dropout sampling and find that the estimated label uncertainty can be used to increase performance under open-set conditions.

5.2 Robustness for Specific Application Areas

5.2.1 Robustness for NLP. The robustness of NLP systems is paramount. Adversarial attacks and training both represent active areas of research in recent years and aim to make NLP models less susceptible to attacks (e.g., word-level perturbations). As such, a multitude of approaches have been proposed specifically for this domain.

Zheng et al. [284] present an approach to study both where and how parsers make mistakes by searching over perturbations to existing texts at the sentence and phrase levels. Furthermore, they design algorithms to create such examples for white-box and black-box models. Instead, Yang et al. [264] propose a method designed to tackle word-level adversarial attacks by pulling words closer to their positive samples while pushing away negative ones. They find that their method improves model robustness against a wide set of adversarial attacks while keeping classification accuracy constant. Similarly, Du et al. [66] study the weakness of many state-of-the-art NLP models against word-level adversarial attacks and propose Robust Adversarial Training to combine adversarial training and data perturbation during training. Pruthi et al. [172] look to combat adversarial misspellings by attaching a word recognition model to the classification model. They find that the adversary can degrade the performance of a text classifier to the point where it is equivalent to random guessing just by altering two characters per sentence. Concerning noisy text, Zhou et al. [286] employ multi-task learning, where a transformer-based translation model is augmented with two decoders with different learning objectives. Similarly, Li et al. [125] use adversarial, multi-modal embeddings and neural machine translation to denoise input samples, making it effective against adversarially obfuscated texts. Finally, Chen et al. [42] found promising (albeit highly variable) results for models capable of generating rationales for their predictions.

5.2.2 Robustness for Cybersecurity. As AI finds increased interest in the field of Cybersecurity, the robustness of the overall system is crucial to achieving a satisfactory resistance to intelligent attacks. A significant focus has been the robustness of malware detection. For instance, Abusnaina et al. [2] improve malware classifier accuracy by augmenting training data with altered behavioral Control Flow Graphs extracted from the attacked code. In this direction, more specific research has been conducted around selected operating systems and platforms Anupama et al. [8] first use the Fisher score to identify and select the most relevant attributes for a classifier and subsequently develop three different adversarial attack generation approaches.

Beyond this, defenses against **distributed denial-of-service** (**DDoS**) attacks have been studied through the lens of robustness as well. Abdelaty et al. [1] present an adversarial, GAN-based training framework to produce strong adversarial examples for the DDoS domain to exploit the weaknesses of Network Intrusion Detection Systems. Adversarial samples are produced by combining GAN-generated and benign DDoS samples. Instead, Amarasinghe et al. [6] apply Layer-wise Relevance Propagation to the trained anomaly detector, yielding relevance scores for each individual feature.

5.3 Robustness for Specific Trustworthy AI Concepts

5.3.1 Robustness for Explainability. Robustness has been widely discussed in the context of explainability methods as well [60, 117]. Explainability is regarded as a fundamental aspect to foster trust in AI systems. However, explainers have been found to be as fragile as the models they strive to describe [207]. Thus, research around robust explanations is critical for Trustworthy AI.

Zhang et al. [278] proposed an approach to explore the input space to compute the percentage of inputs on which the prediction can be consistently explained by the height of the decision tree used to explain a neural network's prediction. However, their result is inconclusive as it may seem tied to imbalances in the data used. In a similar vein, Nanda et al. [153] propose a scalable framework using machine-checkable concepts to assess the quality of generated explanations with respect to robustness, specifically their vulnerability to adversarial attacks. Instead, Alvarez-Melis and Jaakkola [5] define a novel notion of robustness based on the point-wise, neighborhood-based local Lipschitz continuity. Gradient- and perturbation-based interpretability methods are evaluated, revealing the non-robustness of such practices and the high instability of perturbation-based methods. Atmakuri et al. [13] focus on understanding the adversarial robustness of explanation methods in the context of text modality. In particular, they utilize saliency maps to generate adversarial examples to evaluate the robustness of the model of interest. They find the used Integrated Gradient explanation method is weak against misspelling and synonym substitution attacks.

Robustness for Counterfactual Explanations. Multiple works address the robustness of counterfactual explanations for adversarial inputs. Virgolin and Fracaros [237] explore how to improve robustness by giving a formal definition of what it means to be robust toward perturbations and implementing this definition into a loss function. To test this definition, they release five datasets in the area of fair ML with reasonable perturbations and plausibility constraints. They find that robust counterfactuals can be found systematically if we account for robustness in the search process. Furthermore, Pawelczyk et al. [166] explore counterfactual explanations by formalizing the similarities between popular counterfactual explainers and adversarial example generators, identifying conditions when they are equivalent. On the other hand, Bajaj et al. [15] generate robust counterfactual explanations on GNNs by explicitly modeling the common decision logic of GNNs on similar input graphs. The robustness of the explanations is due to the common decision boundaries being derived from several, similar input graphs. Finally, the generation of robust text-based counterfactual explanations has also been studied for NLP tasks [246].

5.3.2 Robustness for Fairness. A key attribute of any system to be put into production is fairness. The relationship between fairness and robustness, and how one contributes to the other, has been receiving increased attention. Rezaei et al. [178] aim to make classifications that have robust fairness without relying on previously labeled data, as these may carry some inherent biases. Wang et al. [244] study the effect of relying on noisy protected group labels, providing a bound on the fairness violation concerning the true group. Similarly, Yurochkin et al. [269] propose an adversarial approach to fairness, using a distributionally robust approach to enforcing individual fairness during training. Furthermore, there have also been efforts to improve the fairness of graph-based counterfactual explanations. For example, Agarwal et al. [3] aim to establish a connection between counterfactual fairness and graph stability by developing layer-wise weight normalization and enforcing fairness and stability in the objective function. They see increases in fairness and stability without a decrease in performance.

6 Robustness Assessment and Insights

In parallel to developing novel methods to enhance model robustness, prior work devised evaluation procedures, extensive benchmarks, and empirical studies to assess the robustness AI models. Given the diversity of the suggested methods, in this section, we cover such efforts and highlight the lessons learnt when Robustness intersects other Trustworthy AI concepts: Fairness and Explainability.

6.1 Evaluation Strategies

6.1.1 Evaluation of Robustness. We found most methodologies around evaluating robustness to either compute a safe radius [116, 185] or region [83] within which the model performs robustly, or they compute complementary, error region [280]. Abstract Interpretation, i.e., *a theory which dictates how to obtain sound, computable, and precise finite approximations of potentially infinite sets of behaviors* [76], enables robustness evaluation when combined with techniques like constraint solving [263] and importance sampling [136]. Other evaluation approaches reformulate the robustness assessment problem from different perspectives. Tjeng et al. [227] formulate the verification of the robustness against adversarial attacks as a mixed integer linear program by expressing properties like adversarial accuracy as a conjunction, or disjunction, of linear properties over some set of polyhedra. Webb et al. [247] statistically evaluate robustness by estimating the proportion of inputs for which a defined adversarial property (i.e., an adversarial condition associated to a function that evaluates its violation) is unsatisfied (i.e., there are no counterexamples violating such a property). This reframing is useful to widen the variety of solutions that can be applied to assess robustness, consequently improving their scalability [227, 247], computational speed [227, 265], and enabling the application of pre-existing tools [91].

Evaluation of Certified Robustness. Much attention has also been devoted to evaluating certified robustness [65, 102, 121, 122, 203, 205, 281]. To this end, researchers focus on the efficient computation of robustness bounds [65, 121, 281] while also improving the training procedure to achieve efficiently certifiable [281], or ready to certify [102], models. Deterministic [122] and Random [67] Smoothing approaches have also proven to be effective in evaluating L_1 [122] and L_2 robustness. Nevertheless, overapproximation [203], orthogonalization relaxation [205], and regularization [102] have also been successfully applied to improve the computation of certifiable bounds in adversarial settings. Moreover, Zhang et al. [274] strive to generalize certification techniques to non-piecewise linear activation functions. Finally, additional works have focused on certifying robustness against random input noise from samples and geometric robustness [20].

6.1.2 Benchmarks. In addition to novel evaluation methods, some works also propose comprehensive benchmarks—encompassing approaches, datasets, and pipelines—to evaluate model robustness against selected sets of attacks. In Computer Vision, robustness against various types of adversarial attacks [63, 82, 169] and common corruptions [90, 139], including noise [90, 283], has been evaluated through benchmarking on datasets [90, 139, 169], with custom measures [63, 90], or using comprehensive frameworks [223]. In the first case, pictures are altered through adversarial or common perturbations (e.g., noise, blur) [90, 139, 223], and either generalizability [139] (i.e., whether the model can adequately classify newly perturbed pictures) or its behavior by means of custom metrics [63, 90] are evaluated. A few benchmarks have also been applied in the context of graph networks. For instance, Zheng et al. [283] develop scalable datasets to standardize the process of attack and defence, covering graph modification and graph injection attacks.

While some benchmarks focus on evaluating the effectiveness of defence methods [63, 283], others focus on the intrinsic robustness of the architecture [54, 223]. Tang et al. [223] benchmark architecture design and training techniques against adversarial and natural perturbations, and system noise through a comprehensive platform including pre-trained models and materials dedicated designing robust DNNs. Instead, Croce et al. [54] focus on resource availability and organize evaluation methods and robust models for researchers to use. Note that most benchmarks use well-known datasets (e.g., MNIST or ImageNet). In this sense, some authors have argued that implicitly assuming the data is correct should not be lightly accepted as it may influence the benchmarking process and results [159]. From the broader perspective of Trustworthy AI, evaluating model robustness can be seen as a part of a process to evaluate fairness. Driven by such an objective, Ding et al. [58] create a series of datasets to benchmark their fairness with respect to noise and data distribution shifts.

6.1.3 Metrics. To evaluate model robustness, not only it is essential to choose the proper method, but it is also fundamental to have metrics that properly represent model robustness, attack efficacy, and computational costs. Most of the literature focuses on describing metrics to evaluate the robustness of networks against adversarial attacks [249, 267]. These metrics are generated by either treating the robustness analysis as a local Lipschitz constant estimation problem [249], or by qualitatively interpreting the adversarial attack and defence efficacies through loss visualization [267]. Particularly, the former [249] aims to disentangle the relationships between the evaluation process and the model or attack employed, leading to model-agnostic and attack-agnostic metrics. Besides, while most of the literature addresses robustness in Computer Vision, a small part of the literature discusses robustness in other contexts. In NLP, extending robustness through a metric aligned with linguistic fidelity has proven effective in improving performance on complex linguistic phenomena [115]. Recent research [35] has denoted the lack of proper robustness metrics for tree-based classifiers. Such scarce findings highlight the need for creating sound and robust metrics in less covered contexts. Another relatively unexplored research area revolves around practical, computational aspects like: enhancing methods' precision in computing robustness bounds [203, 263], reducing their computational complexity [219], or execution time [254].

Instead, other researchers focus on suggesting metrics for different aspects of adversarial attacks, devising approaches for evaluating the convergence stability of adversarial examples generation [119] and comparing adversarial attack algorithms [24]. Beyond the necessity for metrics to assess model robustness, other metrics have proven useful in elucidating the relationships between robustness and adversarial examples [10] and accuracy [162].

6.2 Studies Around Proposed Robustness Methods and Insights

6.2.1 Insights on Adversarial Robustness. Studying the adversarial robustness of different ML techniques has been a persistent research focus in recent years.

Based on Comparisons. Beyond formal methods and frameworks, there are several examples of papers empirically evaluating robustness through comparison [103, 192]. For instance, Jere et al. [103] compared the generalization capabilities of CNNs and their eigenvalues and further compared what features are exploited by naturally trained and adversarially trained models. They found that for the same dataset, naturally trained models exploit high-level human-imperceptible features and adversarially robust models exploit low-level human-perceptible features. Another example in this line is the work by Sehwag et al. [192] who inspected the transferability of the robustness of classifiers trained on proxy distributions from generative model to real data distribution, discovering that the difference between the robustness of classifiers trained on such datasets is upper bounded by the Wasserstein distance between them.

Based on the Investigation of Activation Function and Weights Perturbations. There have been several works studying the robustness of models under perturbation of weights or due to changes in activation functions. For example, Tsai et al. [228] studied the robustness of feed-forward neural networks in a pairwise class margin and their generalization behavior under different types of weight perturbation. Furthermore, they designed a novel loss function for training generalizable and robust neural networks against weight perturbation. Song et al. [210] showed that adversarial training is not directly applicable to quantized networks. They proposed a solution to minimize adversarial and quantization losses with better resistance to white- and black-box attacks. Another work that focused on such attacks is Shao et al. [195], who studied the robustness of vision transformers against adversarial perturbations under various black-box and white-box settings.

Based on Language Perturbations. Diverse strategies have been applied in the context of the robustness of NLP models. More commonly, these deal with synthetic character-level [150] or word-level [116, 150] perturbations of text samples. Beyond lexical changes, Sanchez et al. [189] explored the robustness of Natural Language Inference models on semantic perturbations. Regardless, these works found existing models to be fragile even for small perturbations. Finally, Wang et al. [243] resorted to human-generated annotations to compile a dataset for robust sentiment classification.

6.2.2 Insights on Natural Robustness. Substantial research has been devoted to model robustness to noise and OOD data, both prevalent in real-world settings.

Based on Robustness to Noise. A prominent line of work is evaluating robustness of AI systems against noise [22, 288]. Some examples in this area include the study conducted by Ziyadinov and Tereshonok [288], who evaluated whether training CNNs using noisy data increases their generalization capabilities and resilience against adversarial attacks. They found that the amount of uncertainty in the training dataset affects both the recognition accuracy and the dependence of the recognition accuracy on the uncertainty in the testing dataset. Furthermore, they showed that a dataset with such uncertainty can improve recognition accuracy, consequently enhancing its generalizability and resilience against adversarial attacks. Bar et al. [22] also evaluated the robustness of deep neural networks to label noise by applying spectral analysis. The authors demonstrated that regularizing the network Jacobian reduces the high frequency in the learned mapping and show the effectiveness of Spectral Normalization in increasing the robustness of the network, independently from the architecture and the dataset.

Based on Robustness to Differences in Distributions. Another area of interest is studying differences in data distributions. On the problem of object-centric learning, Dittadi et al. [59] discovered that the overall segmentation performance and downstream prediction of in-distribution objects is not affected by a single OOD object. On the other hand, Burns and Steinhardt [33] studied

adaptive batch normalization, which aligns mean and variance of each channel in CNNs across two distributions . They found that for distribution shifts that do not involve changes in local image statistics, accuracy can be degraded because of batch normalization.

6.3 Tradeoffs Between Robustness and Other Trustworthy AI Concepts

6.3.1 Tradeoff with Accuracy. A key question to be asked when analyzing the robustness of a system is what the impact of the changes is on the accuracy of the model. Multiple studies have found a significant tradeoff between robustness and accuracy, where an increase in one leads to a decrease in the other. Su et al. [216] evaluated the robustness of 18 existing deep image classification models, focusing on the tradeoff between robustness and performance. They found that model architecture is a more critical factor to robustness than model size and that networks of the same family share similar robustness properties. Raghunathan et al. [175] further discussed this and described in detail the effect of augmentation achieved through adversarial training on the standard error in linear regression models when the predictor has zero standard and robust error. Tsipras et al. [229] also studied how robustness and accuracy tradeoff, as well as the features that were learned. While Miller et al. [144] investigated the connection between accuracy in- and OOD and show that that OOD performance is strongly correlated with in-distribution performance for a wide range of models and distribution shifts.

6.3.2 Tradeoff with Fairness. Benz et al. [26] evaluated the impact of robustness on accuracy and fairness. They found inter-class discrepancies in accuracy and robustness, specifically in adversarially trained models and that adaptively adjusting class-wise loss weights negatively affects overall performance. Xu et al. [257] hypothesized that adversarial training algorithms tend to introduce severe disparity in accuracy and robustness between different groups of data, and showed this phenomenon can happen under adversarial training algorithms minimizing neural network models' robustness errors. They also propose a Fair-Robust-Learning framework to mitigate unfairness in adversarial defenses. On the other hand, Pruksachatkun et al. [171] studied if an increase in robustness can improve fairness. They investigated the utility of certified word substitution robustness methods to improve the *equality of odds* and *equality of opportunity* in text classification tasks. They found that certified robustness methods improve fairness, and using both robustness and bias mitigation methods in training results in an improvement for both.

6.3.3 Tradeoff with Explainability. Few works investigate the extent to which methods for increasing model robustness impact the features such models use to make predictions, and especially to what extent these features remain meaningful to human judgement. Especially, Woods et al. [252] showed that the fidelity of explanations is negatively impacted by adversarial attacks, and propose a regularization method for increasing robustness lead to better model explanations (termed *Adversarial Explanations*). Nourelahi et al. [160] investigated how methods dealing with OOD examples impact the alignment of the features the model has learned with features a human would expect to use. While this is an initial empirical exploration, their results illustrate the complexity of the relation between robustness and feature alignment, as there does not seem to be a model that performs consistently better over these criteria. They suggest to extend their benchmark effort to more types of models, and of robustness and explainability techniques.

7 Discussion: Disparate Research on the Various Facets of Robustness

The robustness of AI systems is a broad, open problem under the umbrella of Trustworthy AI and the copious amount of literature that can be found is a testament to that. Researchers from diverse domains have studied the impact of controlled data perturbations as well as naturally-occurring ones, how to strengthen neural architectures through additional mechanisms, and how

to efficiently and effectively train models underlying such systems. In this section, we summarize the gaps and trends we evinced from our inspection of the existing literature.

7.1 Addressing Gaps from the Literature

7.1.1 Gaps within Robustness.

Natural Brittleness. We found that little attention is put on defining natural perturbations and attacks. Instead, much work revolves around defining synthetic attacks and evaluating defense mechanisms against them. While this may make sense from the perspective of a malicious attacker, it does not necessarily translate to robustness in real-world operating conditions. Only a few works in Computer Vision focus on such a type of attacks. Another interesting research direction is signaled by the lack of model-agnostic adversaries. While both automatic and rule-based approaches to generating adversaries exist, these tend to be targeted toward certain types of AI systems. Obtaining model-agnostic attacks would be the dual case to such a scenario and could provide for a common baseline for evaluating the robustness of AI systems. Moreover, achieving model-agnostic and perturbation-agnostic evaluations approaches would allow to disentangle the relationship between these scenario-specific aspects and the actual robustness of the model, finally leading to an unbiased analysis of the robustness of a system [249].

The Computer Vision Hegemony. The immediate outcome of our survey is the extensive effort put into studying—and enhancing—the robustness of models targeted toward Computer Vision, especially CNNs. Papers from this sub-field of **Artificial Intelligence** (**AI**) greatly outnumber the ones from other areas, like NLP. We found this to be the case regardless of the aspect (attack generation, defense, etc.) scholars focus on. While important, such a focus being put on Computer Vision only begs the question of why other domains have received little contributions compared to the former. Possible explanations for this can be traced back to difficulties in defining perturbations and attacks within certain data manifolds (e.g., word embeddings), or to the lack of alignment between robustness in ML and robustness in specific application domains (e.g., signal processing). On the other hand, the intrinsic complexity of pictures compared to other types of data, in particular with respect to the features that can be perturbed and the diversity in the available approaches to evaluate distances between pictures, influence the broadness of the research field.

7.1.2 Gaps Stemming from the Intersection Between Robustness and Other Trustworthy AI Concepts.

Robustness and Explainability. Considering the brittleness of existing AI systems in conjunction with their opaqueness, their explainability is of paramount importance. XAI methods have been, and still are being, proposed [84, 85] to tackle such a challenge. However, on one hand, few works discuss the robustness of XAI methods and of the produced explanations, yet this is a crucial dimension that needs to be addressed to obtain explanations that are both faithful (i.e., correctly describing model behavior) and trustworthy.

On the other hand, explainability can better inform the ideation and implementation of approaches geared toward robustness. However, little work has been conducted in this direction. These works all rely on the idea that when the model features extracted via an explainability method are aligned with human reasoning (i.e., the features are meaningful for a human to make a prediction for a data sample), then the model should be more robust. In terms of evaluation, only Nanda et al. [153] have investigated how explainability can be used in order to evaluate the robustness of a model, with the assumption mentioned above. In terms of improvement, Kortylewski et al. [113] proposed *Compositional Neural Networks*, a unification of CNNs with part-based models (inherently interpretable models), and show that these new networks increase model robustness

to various partial occlusions of objects. Chen et al. [42] also demonstrated that inherently interpretable models such as rationale models in NLP are naturally more robust to certain adversarial attacks yet are still brittle to certain scenarios. Similarly, Li et al. [123] proposed a model training framework that combines adversarial training with constraints for ensuring the meaningfulness of the model features, reaching higher model robustness. Finally, Freitas et al. [72] tackle adversarial robustness with model features, by making the additional assumption that when the model features are not meaningful, the model might be under attack.

Tensions between Accuracy, Robustness, Fairness, and Explainability. Connected to the above points, it is worth noting how existing research is focused on enhancing robustness at the expense of accuracy, much like optimizing for accuracy led to a lack of explainability. Similarly, scholars have studied the interplay with fairness as well as the possible issues stemming from it. These dimensions are not exclusive and need to be addressed holistically and considered on equal terms when aiming to build trustworthy and fair AI systems. In this sense, sole data-driven approaches have shown their limitations. Discussions around these topics have pointed toward the need for integrating symbolic knowledge. However, few of them touched upon which kind of knowledge is needed and how to collect it. In Section 7.3 and Section 8, we provide a commentary on human-centered approaches and how these approaches can provide a path toward tackling the aforementioned challenges for robust AI.

7.2 Latest Breakthroughs: Generative Foundation Models

While writing this survey, considerable engineering advancements have been made in the space of generative foundation models, such as DALL-E 3 and GPT-4. However, such advancements are often due to an increase in model size and largely prioritize properties like textual fluency. As a result, several robustness-related challenges, e.g., hallucinations [104] and poor performance on OOD data [241], remain unanswered. Because of their lacklustre understanding capabilities [25], these shortcomings emerge even when more sophisticated prompting strategies are used, e.g., in-context learning [62] or chain-of-thought prompting [248]. Only retrieval-augmented generation [201] appears to mitigate hallucinations. Despite the central role of humans in the creation of these models (e.g., with Reinforcement Learning through Human Feedback [213]), research is still centered around benchmarking [87]. In the sections that will follow, we discuss opportunities around human involvement and human knowledge to improve and evaluate robustness of AI models.

7.3 Deepening the Research on Human Involvement for Existing Robustness Methods

A number of papers we surveyed implicitly involve humans to instantiate the methods they propose, either to assess or enhance a model's robustness. Yet, they do not delve deeper into the challenges for a human agent to perform their task, which constitutes an obstacle to the development of methods and frameworks for overcoming these challenges. This merits further investigation as such human involvement is essential to the success of the methods. Especially, we identify two main areas where human involvement is necessary but lacks research.

7.3.1 Increasing Robustness. Various methods that aim at increasing robustness implicitly employ humans, without extensive focus. Jin et al. [105], for instance, collect potential adversarial examples by executing a sequence of engineered steps that could be refined by the practitioner who would leverage existing tools for, e.g., identifying synonyms and antonyms, ranking word importance, and so on. Peterson et al. [167], Chang et al. [41], Nanda et al. [153], and Ning et al. [157], respectively, show that one can train more robust models by leveraging human uncertainty on sample labels instead of using reconciled binary labels, by integrating human rationales for the labeling process into the training process, or by actively querying the most relevant levels of

perturbations from an expert during training. While these are promising research directions, these works could further be improved by exploiting existing works on human computation assessing the quality of crowdsourced outputs [100], or designing crowdsourcing tasks that remove task ambiguity and lead to higher quality outputs [69], especially in the context of subjective tasks. This could serve to understand the nature of uncertainties and define rationales that are relevant to robustness.

7.3.2 Evaluating Robustness. To design appropriate perturbations or attacks on which a model should be robust, one often needs human knowledge. For instance, Jin et al. [105] and La Malfa and Kwiatkowska [115] generate adversarial attacks on text samples that have to verify a number of human-defined constraints for them to be deemed realistic by humans. Yet, designing such constraints and empirically evaluating (through user studies) to what extent the samples transformed by the corresponding constrained attack align with the human idea of "realistic" sample, has not been investigated extensively, despite how crucial that is for engineering "good" attacks.

Similarly, works on robustness to natural perturbations should ideally define a comprehensive set of domain-specific perturbations relevant to the problem at hand and its context. However, to the best of our knowledge, existing works that develop benchmarks or robustness-enhancing methods [90, 112] with regard to such perturbations have not investigated ways to be more comprehensive. While we believe in the impossibility to reach comprehensiveness (previously unheard-of perturbations can always arise), one could develop tools to support the definition of relevant perturbations. For instance, we envision the usefulness of fine-grained, actionable taxonomies of perturbations (e.g., Koh et al. [112] talk about subpopulation shifts and domain generalization, but this might vary in different domains and types of tasks); collaborative documentation of domain-specific perturbations; libraries to generate such perturbations semi-automatically; and frameworks and metrics to uncover new types of perturbations in the wild, potentially involving humans at runtime.

8 A Conspicuous Absent From the Literature: The ML Practitioner

Last but not least, our systematic survey also reveals another prominent research gap: the absence of human-centered work in proposed approaches, and the lack of technologies and workflows to support ML practitioners in handling robustness. In this section, we discuss relevant research literature, and future research directions regarding this topic.

8.1 Robustness By Human-Knowledge Diagnosis

One notable absentee from the retrieved papers is robustness by human-based diagnosis. Existing works focus on generating OOD data to make a model fail, and later expose this model to this data during training to make it more robust [27, 41, 74, 167]. Especially for robustness to natural perturbations, this means that one should characterize the type of data the model might encounter before being able to generate such data [59, 64, 80]. This is not always possible in practice, due to the known challenges ML practitioners typically face when working with data and models. For instance, due to contractual and privacy reasons [95, 235], ML practitioners might not have access to deployment data, preventing them from reasoning about OOD data, or the goal and context of application of the ML model they develop might change over time, rendering what might be at present considered (or not) OOD (in-)valid [187]. ML practitioners might also face difficulties in collaborating with domain experts [152, 272], e.g., to reflect on what data should be considered within or out of the distribution, or to evaluate the meaningfulness of a model features (used to estimate the robustness of the model [72, 153]). Besides, it is well-known that ML practitioners

might not receive enough support from their organization, e.g., in terms of budget, time, training, to dive into questions of trustworthiness of their ML models in general [176].

To circumvent this issue, a major, promising research direction surfaces from comparing the surveyed robustness methods to existing works in other computer science fields. This direction revolves around developing complementary, hybrid human–machine approaches, that would leverage research progress in human-centered fields, essentially explainability, crowdsourcing and HIL ML, as well as knowledge-based systems, to estimate model performance on more realistic data distributions without requiring such distributions.

8.1.1 Existing Approaches. Only few related works leverage human capabilities to identify and mitigate potential failures of a model. In particular, explanations for datasets [188] have been proposed that could be leveraged by a practitioner to identify data skews that might impact the model performance. In this vein, Liu et al. [131] introduce a hybrid approach to identify unknown unknowns, where humans first identify and describe patterns in a small set of unknown unknowns, and then classifiers are learned to recognize these patterns automatically in new samples. Departing from datasets, Stacey et al. [211] and Arous et al. [11] have trained models whose features are better aligned with human reasoning (with the assumption that alignment leads to stronger robustness), by leveraging human explanations of the right answer to the inference task and controlling the features learned by the model during training to align with these human explanations.

8.1.2 Envisioned Research Opportunities. The above approaches reveal that instead of looking solely at the outputs of a model and its confidence in its predictions, one can leverage additional information such as the model features or training dataset, to estimate the model's robustness. Especially, even when a model prediction is correct, the model features might not be meaningful. Hence, assessing model features and their human-alignment can allow to shift from solely evaluating the correctness of the predictions on the available test, to indirectly assessing the robustness of the model to OOD data points. Moreover, understanding characteristics of the datasets that led to such learned features could later on serve to mitigate unaligned features.

Surfacing Model Features using Research on Explainability and Human Computation. To surface a model's features, one can rely on a plethora of explainability methods [188]. Certain models are built with the idea of being explainable by design [220, 279], while others are applied post-hoc interpretability methods [19, 179, 214], with different properties (e.g., different nature of explanations being correlation- or causation-based, different scopes be it local or global, different mediums be it visual or textual) [130, 204, 209]. It is now important to adapt such feature explanations to allow for checking their alignment with human-expected features.

In that regard, the push toward human-centered explanations for ML practitioners is highly relevant. Existing explanations often leave space for many different human interpretations, for which the practitioners do not always have domain expertise to disambiguate the highest-fidelity features. For instance, methods that output saliency maps [202] or image patches [78, 110] do not pinpoint the actual human-interpretable features the model has learned. Yet, one might need clear human concepts to reason over the alignment of the features [17]. Hence, further research on *semantic, concept-based explanations* acquired via human computation is needed [19, 93].

Leveraging Literature on Knowledge Acquisition for Identifying Expected Features. To reason over feature alignment, one also needs to develop an understanding of the model expected features. While very few works have looked into this problem [196], existing works on commonsense-knowledge acquisition [270] could be leveraged to that end. These works propose to harvest knowledge automatically from existing resources such as text libraries, or through the involvement of human agents (e.g., through efficient and low-cost interactions within Games with a Purpose

[16, 182, 238]), or other types of carefully designed crowdsourcing tasks [99, 191]. One would need to investigate how to adapt such approaches to collect relevant knowledge, and how to represent this knowledge into relevant feature-based information.

Comparing Features via Reasoning Frameworks and Interactive Tools. Finally, practitioners need tools to check the alignment between the model and expected features. Interactive frameworks and user interfaces [17], e.g., *Shared Interest* [28], take a step in that direction as they enable manual exploration of model features, with various degrees of automation for comparing to expected features. Inspired by the literature on AI diagnosis, such as abductive reasoning [53, 181], automated feature-reasoning methods could also fasten the process while making it more reliable.

8.2 Involving Humans in Other Phases of the ML Lifecycle

Broader ML literature has also proposed other approaches to involve humans and make "better" models. Yet, none of these approaches has considered making the models more robust. Instead, they focus on increasing the performance of the model on the test set. Hence, we suggest to investigate how to adapt such approaches to increase model robustness.

8.2.1 ML with a Reject Option. While ML models typically make predictions for all input samples, this might not be reasonable and turn dangerous in high-stake domains, when the predictions are likely to be incorrect. Accordingly, a number of research works have developed methods to learn when to appropriately reject a prediction, and defer the decision about the sample to a human agent [89]. Proposed rejectors can either be *separate rejectors* placed before the predictor, that select the input samples to input to this predictor; *dependent rejectors* placed after the predictor and re-using its information (e.g., confidence metrics) to decide which predictions not to account for; and *integrated rejectors* that are combined to the predictor, by treating the rejection option as an additional label to the ones to predict. Each type of rejector bears advantages and disadvantages based on the context of the decision, and would merit being adapted to robustness, as we only found few works toward that direction [163, 215].

8.2.2 HIL ML Pipelines. HIL ML [234] is traditionally concerned with developing learning frameworks that account for the noisy crowd labels [177], or "learning from crowds", through models of the annotation process (e.g., task difficulty, task subjectivity, annotator expertise). Such frameworks often rely on active learning to reduce annotation cost [259, 261]. More recent works around HIL ML also devise new approaches to build better model pipelines by involving the crowd, such as to identify weak components of a system [161], to identify noise and biases in the training data [98, 262], or to propose potential data-based explanations to wrong predictions [34]. While we could find a few works that investigate the intersection between active learning and adversarial training [141, 143, 200, 206], we could not find any work that looks more broadly at the different types of robustness, and the different ways of bringing humans in the ML pipeline. These intersections are yet promising as they constitute more realistic scenarios of the development of ML systems and they succeeded in making models more accurate in the past.

8.3 Supporting ML Practitioners in Handling Robustness

Beyond research, ML practitioners build ML systems in practice. Hence, it is not sufficient to develop methods that work in theory: research should understand the obstacles practitioners encounter in making their systems robust. While studying the gap between research and practice has revealed highly insightful for various ML contexts [94, 96, 109, 114, 130, 170, 217], to the best of our knowledge, it has not been studied for ML robustness. Possibly the closest work is that of Shankar et al. [193] that investigated MLOps practices toward monitoring of data shifts or attacks.

8.3.1 Understanding Practices Around Robustness. The human–computer interaction community has performed qualitative, empirical, studies, based on semi-structured interviews with ML practitioners, about e.g., stakeholder collaboration [114, 170], debugging practices [18], and the use of tools such as explainability methods [94, 96, 130] or fairness toolkits [120, 180]. These studies have resulted in frameworks modeling the practitioner's process and challenges, and discussions around the fit of existing research works to answer these challenges. We argue that adopting similar research questions would reveal useful to better direct robustness research. For instance, Liao et al. [130] have constituted a question bank that highlights the questions practitioners ask when building a model by exploiting explainability. A robustness question bank would similarly provide a structured understanding of research gaps. Moreover, Deng et al. [55] have shown a major gap in terms of guidance for practitioners to choose appropriate fairness metrics and mitigation methods. Acknowledging the plethora of robustness metrics and methods, user studies around robustness would reveal a similar gap that could be filled by taking inspiration from the fairness literature.

8.3.2 Integrating Robustness into Existing Workflows. To support practitioners in model building, researchers have developed workflows [208] and tools, e.g., user interfaces to investigate models, training datasets, and related failures [17, 154], documentation or checklists [7, 75, 146] to support making and documenting relevant choices, and so on. Similarly, we argue that robustness research should not only focus on algorithmic evaluation and improvement, but also aim at developing new supportive tools and integrating them into existing solutions. Closest to supporting practitioners in handling robustness, Shen [198] propose the idea of establishing trust contracts, i.e., contract data distributions and tasks that define the type of task and data that is in- and out-of-distribution. Yet, this remains challenging as there is no appropriate way to formalize such contracts.

9 Conclusion

In this survey, we collected, structured, and discussed literature related to robustness in AI systems. To this end, we performed a rigorous data collection process where we collected, filtered, summarized, and organized literature related to AI robustness generated in the last 10 years. As part of our review, and as opposed to prior surveys, we searched for robustness solutions to *both* adversarial and natural perturbations in a task-agnostic way. Furthermore, we sought to cover both algorithmic-centric and human-lead approaches. Based on this literature, we first discussed the main concepts, definitions, and domains associated with robustness, disambiguating the terminology used in this field. We then generated a taxonomy to structure the reviewed papers and to spot recurring themes. We identified three main themes and thoroughly discussed them. In particular, we focused on (1) fundamental approaches to improve model robustness against adversarial and non-adversarial perturbations, (2) applied approaches to enhance robustness in different application areas, and (3) evaluation approaches and insights. We finalized our article by describing the research gaps identified in the literature and by highlighting the scarcity of solutions that include humans as central actors for improved robustness. We argue that humans could play a fundamental role in improving, evaluating, and validating AI robustness. Consequently, we suggest future research directions that could benefit from including humans in the loop and point to the challenges (and concomitant research opportunities) that arise when advocating for human-led practices for AI robustness. In conclusion, we contributed to the existing literature with an informative review that summarizes and organizes recent work in the field of AI robustness while also suggesting novel human-centered approaches for the research community to explore, discuss, and further develop.

References

[1] Maged Abdelaty, Sandra Scott-Hayward, Roberto Doriguzzi-Corin, and Domenico Siracusa. 2021. GADoT: GAN-based adversarial training for robust DDoS attack detection. In *CNS*. IEEE, 119–127.

[2] Ahmed Abusnaina, Mohammed Abuhamad, Hisham Alasmary, Afsah Anwar, Rhongho Jang, Saeed Salem, DaeHun Nyang, and David Mohaisen. 2022. DL-FHMC: Deep learning-based fine-grained hierarchical learning approach for robust malware classification. *IEEE Transactions on Dependable and Secure Computing* 19, 5 (2022), 3432–3447. DOI: https://doi.org/10.1109/TDSC.2021.3097296

[3] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. 2021. Towards a unified framework for fair and stable graph representation learning. In *UAI*. PMLR, 2114–2124.

[4] Sheikh Waqas Akhtar, Saad Rehman, Mahmood Akhtar, Muazzam A. Khan, Farhan Riaz, Qaiser Chaudry, and Rupert Young. 2016. Improving the robustness of neural networks using k-support norm based adversarial training. *IEEE Access* 4 (2016), 9501–9511. DOI: https://doi.org/10.1109/ACCESS.2016.2643678

[5] David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the Robustness of Interpretability Methods. arXiv:1806.08049. [cs.LG]. DOI: https://doi.org/10.48550/ARXIV.1806.08049

[6] Kasun Amarasinghe, Kevin Kenney, and Milos Manic. 2018. Toward explainable deep neural network based anomaly detection. In *HSI*. 311–317. DOI: https://doi.org/10.1109/HSI.2018.8430788

[7] Ariful Islam Anik and Andrea Bunt. 2021. Data-centric explanations: Explaining training data of machine learning systems to promote transparency. In *CHI*. 1–13.

[8] M. L. Anupama, P. Vinod, Corrado Aaron Visaggio, M. A. Arya, Josna Philomina, Rincy Raphael, Anson Pinhero, K. S. Ajith, and P. Mathiyalagan. 2021. Detection and robustness evaluation of android malware classifiers. *Journal of Computer Virology and Hacking Techniques* 18, 3 (2021), 1–24.

[9] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. 2020. Adversarial concurrent training: Optimizing robustness and accuracy trade-off of deep neural networks. In *31st British Machine Vision Conference (BMVC'20)*, BMVA Press. Retrieved from https://www.bmvc2020-conference.com/assets/papers/0859.pdf

[10] Paolo Arcaini, Andrea Bombarda, Silvia Bonfanti, and Angelo Gargantini. 2020. Dealing with robustness of convolutional neural networks for image classification. In *AITest*. 7–14. DOI: https://doi.org/10.1109/AITEST49225.2020.00009

[11] Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. Marta: Leveraging human rationales for explainable text classification. In *AAAI*. Vol. 35, 5868–5876.

[12] Alejandro Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[13] Shriya Atmakuri, Tejas Chheda, Dinesh Kandula, Nishant Yadav, Taesung Lee, and Hessel Tuinhof. 2022. Robustness of Explanation Methods for NLP Models. arXiv:2206.12284. [cs.CL]. DOI: https://doi.org/10.48550/ARXIV.2206.12284

[14] Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. 2022. Improving Adversarial Robustness via Channel-wise Activation Suppressing. arXiv:2103.08307. [cs.LG]. DOI: https://doi.org/10.48550/ARXIV.2103.08307

[15] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. 2021. Robust counterfactual explanations on graph neural networks. In *NeurIPS*. Vol. 34, Curran Associates, Inc., 5644–5655. Retrieved from https://proceedings.neurips.cc/paper/2021/file/2c8c3a57383c63caef6724343eb62257-Paper.pdf

[16] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. 2022. Ready player one! Eliciting diverse knowledge using a configurable game. In *WWW*. 1709–1719.

[17] Agathe Balayn, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2022. How can explainability methods be used to support bug identification in computer vision models?. In *CHI'22*. 1–16.

[18] Agathe Balayn, Natasa Rikalo, Jie Yang, and Alessandro Bozzon. 2023. Faulty or ready? Handling failures in deep-learning computer vision models until deployment: A study of practices, challenges, and needs. In *CHI'23*.

[19] Agathe Balayn, Panagiotis Soilis, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2021. What do you mean? Interpreting image classification with crowdsourced concept extraction and analysis. In *WWW*. 1937–1948.

[20] Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. 2019. Certifying geometric robustness of neural networks. In *NeurIPS*. Vol. 32, Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2019/file/f7fa6aca028e7ff4ef62d75ed025fe76-Paper.pdf

[21] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *AAAI*. Vol. 33, 2429–2437.

[22] Oshrat Bar, Amnon Drory, and Raja Giryes. 2022. A spectral perspective of DNN robustness to label noise. In *AIStats (PMLR, Vol. 151)*. PMLR, 3732–3752. Retrieved from https://proceedings.mlr.press/v151/bar22a.html

[23] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. 2021. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence* 3, 12 (2021), 1061–1070.

[24] Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya V. Nori, and Antonio Criminisi. 2016. Measuring neural net robustness with constraints. In *NeurIPS*. Curran Associates, 2621–2629. DOI: https://doi.org/10.5555/3157382.3157391

[25] Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.), Association for Computational Linguistics, Online, 5185–5198. DOI: https://doi.org/10.18653/v1/2020.acl-main.463

[26] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. 2021. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS*. PMLR, 325–342.

[27] Arjun Nitin Bhagoji, Daniel Cullina, Chawin Sitawarin, and Prateek Mittal. 2018. Enhancing robustness of machine learning systems via data transformations. In *CISS*. IEEE, 1–5.

[28] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. 2022. Shared interest: Measuring human-AI alignment to identify recurring patterns in model behavior. In *CHI*. 1–17.

[29] Aleksandar Bojchevski and S. Günnemann. 2019. Certifiable robustness to graph perturbations. In *NeurIPS* .

[30] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[31] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *KDD '06*. ACM, 535–541. DOI: https://doi.org/10.1145/1150402.1150464

[32] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*. PMLR, 77–91.

[33] Collin Burns and Jacob Steinhardt. 2021. Limitations of post-hoc feature alignment for robustness. In *CVPR*. 2525–2533.

[34] Ángel Alexander Cabrera, Abraham J. Druck, Jason I. Hong, and Adam Perer. 2021. Discovering and validating AI errors with crowdsourced failure reports. *ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–22.

[35] Stefano Calzavara, Lorenzo Cazzaro, Claudio Lucchese, Federico Marcuzzi, and Salvatore Orlando. 2022. Beyond robustness: Resilience verification of tree-based classifiers. *Computers & Security* 121, 10 (2022), 102843. DOI: https://doi.org/10.1016/j.cose.2022.102843

[36] Ginevra Carbone, Matthew Wicker, Luca Laurenti, A. Patane, L. Bortolussi, and Guido Sanguinetti. 2020. Robustness of Bayesian neural networks to gradient-based attacks. In *NeurIPS*. Vol. 33, Curran Associates, 15602–15613. Retrieved from https://proceedings.neurips.cc/paper/2020/file/b3f61131b6eceeb2b14835fa648a48ff-Paper.pdf

[37] Nicholas Carlini, A. Athalye, N. Papernot, W. Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. 2019. On evaluating adversarial robustness. arXiv:1902.06705. [cs.LG]. Retrieved from https://arxiv.org/abs/1902.06705

[38] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *SP*. IEEE, 39–57.

[39] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[40] Alvin Chan, Yi Tay, Yew-Soon Ong, and Jie Fu. 2020. Jacobian adversarially regularized networks for robustness. In *8th International Conference on Learning Representations (ICLR'20)*, OpenReview.net. Retrieved from https://openreview.net/forum?id=Hke0V1rKPS

[41] C. Chang, G. Adam, and A. Goldenberg. 2021. Towards robust classification model by counterfactual and invariant data generation. In *2021 CVPR*. IEEE Computer Society, Los Alamitos, CA, USA, 15207–15216. DOI: https://doi.org/10.1109/CVPR46437.2021.01496

[42] Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3792–3805. DOI: https://doi.org/10.48550/ARXIV.2204.11790

[43] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. 2018. EAD: Elastic-net attacks to deep neural networks via adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (2018). DOI: https://doi.org/10.1609/aaai.v32i1.11302

[44] Shang-Tse Chen, C. Cornelius, J. Martin, and D. Horng Chau. 2019. ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector. In *Machine Learning and Knowledge Discovery in Databases*. M. Berlingerio, F. Bonchi, T. Gärtner, N. Hurley, and G. Ifrim (Eds.), Springer, 52–68. DOI: https://doi.org/10.1007/978-3-030-10925-7_4

[45] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. 2018. ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector. In *ECML/KDD*. Springer, 52–68.

[46] Xiangning Chen, Cihang Xie, Mingxing Tan, Li Zhang, Cho-Jui Hsieh, and Boqing Gong. 2021. Robust and accurate object detection via adversarial learning. In *CVPR*. 16622–16631.
[47] Yu Chen, Lingfei Wu, and Mohammed Zaki. 2020. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In *NeurIPS*. Vol. 33, Curran Associates, Inc., 19314–19326. Retrieved from https://proceedings.neurips.cc/paper/2020/file/e05c7ba4e087beea9410929698dc41a6-Paper.pdf
[48] Minhao Cheng, Pin-Yu Chen, Sijia Liu, Shiyu Chang, Cho-Jui Hsieh, and Payel Das. 2021. Self-progressing robust training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7107–7115. DOI: https://doi.org/10.48550/ARXIV.2012.11769
[49] Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, and Cho-Jui Hsieh. 2022. CAT: Customized adversarial training for improved robustness. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI'22)*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 673–679. DOI: https://doi.org/10.24963/ijcai.2022/95
[50] Xiang Cheng, Yunzhe Hao, Jiaming Xu, and Bo Xu. 2020. LISNN: Improving spiking neural networks with lateral interactions for robust object recognition. In *IJCAI*. 1519–1525. DOI: https://doi.org/10.24963/ijcai.2020/211
[51] Seok-Hwan Choi, Jin-Myeong Shin, Peng Liu, and Yoon-Ho Choi. 2022. ARGAN: Adversarially robust generative adversarial networks for deep neural networks against adversarial examples. *IEEE Access* 10 (2022), 33602–33615. DOI: https://doi.org/10.1109/ACCESS.2022.3160283
[52] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). Proceedings of Machine Learning Research, Vol. 97. PMLR, 1310–1320. https://proceedings.mlr.press/v97/cohen19c.html
[53] Luca Console, Daniele Theseider Dupre, and Pietro Torasso. 1989. A theory of diagnosis for incomplete causal models.. In *IJCAI*. 1311–1317.
[54] Francesco Croce, M. Andriushchenko, V. Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2021. RobustBench: A standardized adversarial robustness benchmark. In *NeurIPS*. Vol. 1. Retrieved from https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/a3c65c2974270fd093ee8a9bf8ae7d0b-Paper-round2.pdf
[55] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *FAccT '22*. DOI: https://doi.org/10.1145/3531146.3533113
[56] Zhun Deng, Linjun Zhang, Amirata Ghorbani, and James Zou. 2021. Improving adversarial robustness via unlabeled out-of-domain data. In *International Conference on Artificial Intelligence and Statistics*, PMLR, 2845–2853. DOI: https://doi.org/10.48550/ARXIV.2006.08476
[57] Chaitanya Devaguptapu, Devansh Agarwal, Gaurav Mittal, Pulkit Gopalani, and Vineeth N. Balasubramanian. 2021. On adversarial robustness: A neural architecture search perspective. In *ICCV*. 152–161.
[58] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. In *NeurIPS*. Vol. 34, Curran Associates, Inc., 6478–6490. Retrieved from https://proceedings.neurips.cc/paper/2021/file/32e54441e6382a7fbacbbbaf3c450059-Paper.pdf
[59] Andrea Dittadi, Samuele Papa, Michele De Vita, Bernhard Schölkopf, Ole Winther, and Francesco Locatello. 2022. Generalization and robustness implications in object-centric learning. In *International Conference on Machine Learning*, PMLR, 5221–5285. DOI: https://doi.org/10.48550/ARXIV.2107.00637
[60] Ann-Kathrin Dombrowski, Christopher J. Anders, Klaus-Robert Müller, and Pan Kessel. 2022. Towards robust explanations for deep neural networks. *Pattern Recognition* 121, 1 (2022), 108194.
[61] Minjing Dong, Yanxi Li, Yunhe Wang, and Chang Xu. 2020. Adversarially robust neural architectures. *arXiv preprint arXiv:2009.00902* (2020). DOI: https://doi.org/10.48550/ARXIV.2009.00902
[62] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. arXiv:2301.00234 [cs.CL]. Retrieved from https://arxiv.org/abs/2301.00234
[63] Yinpeng Dong, Qi-An Fu, X. Yang, T. Pang, H. Su, Zihao Xiao, and Jun Zhu. 2020. Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 321–331. DOI: https://doi.org/10.48550/ARXIV.1912.11852
[64] Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. 2022. A Systematic Review of Robustness in Deep Learning for Computer Vision: Mind the gap? arXiv:2112.00639 [cs.CV]. Retrieved from https://arxiv.org/abs/2112.00639
[65] Tianyu Du, Shouling Ji, Lujia Shen, Yao Zhang, Jinfeng Li, Jie Shi, Chengfang Fang, Jianwei Yin, Raheem Beyah, and Ting Wang. 2021. Cert-RNN: Towards certifying the robustness of recurrent neural networks. In *CCS*. 516–534.
[66] Xiaohu Du, Jie Yu, Shasha Li, Zibo Yi, Hai Liu, and Jun Ma. 2021. Combating word-level adversarial text with robust adversarial training. In *IJCNN*. 1–8. DOI: https://doi.org/10.1109/IJCNN52387.2021.9533725

[67] Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. 2020. A framework for robustness certification of smoothed classifiers using f-divergences. In *ICLR*. Retrieved from https://openreview.net/forum?id=SJlKrkSFPH

[68] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. 2020. Can cross entropy loss be robust to label noise?. In *IJCAI-20*. 2206–2212. DOI: https://doi.org/10.24963/ijcai.2020/305 Main track.

[69] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. 2013. Keep it simple: Reward and task design in crowdsourcing. In *CHItaly '13*. 1–4.

[70] Luciano Floridi. 2019. Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 6 (2019), 261–262.

[71] James Fox and Sivasankaran Rajamanickam. 2019. How robust are graph neural networks to structural noise? *arXiv preprint arXiv:1912.10206* (2019). DOI: https://doi.org/10.48550/ARXIV.1912.10206

[72] Scott Freitas, Shang-Tse Chen, Zijie J. Wang, and Duen Horng Chau. 2020. Unmask: Adversarial detection and defense through robust feature alignment. In *Big Data*. IEEE, 1081–1088.

[73] Ji Gao, Beilun Wang, Zeming Lin, Weilin Xu, and Yanjun Qi. 2017. Masking deep neural network models for robustness against adversarial samples. *arXiv preprint arXiv:1702.06763* (2017). DOI: https://doi.org/10.48550/ARXIV.1702.06763

[74] Xiang Gao, Ripon K. Saha, Mukul R. Prasad, and Abhik Roychoudhury. 2020. Fuzz testing based data augmentation to improve robustness of deep neural networks. In *ICSE (ICSE '20)*. ACM, New York, NY, USA, 1147–1158. DOI: https://doi.org/10.1145/3377811.3380415

[75] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM* 64, 12 (2021), 86–92.

[76] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. 2018. AI2: Safety and robustness certification of neural networks with abstract interpretation. In *SP*. IEEE, 3–18.

[77] Simon Geisler, Tobias Schmidt, Hakan Şirin, Daniel Zügner, Aleksandar Bojchevski, and Stephan Günnemann. 2021. Robustness of graph neural networks at scale. In *NeurIPS*. Vol. 34, Curran Associates, Inc., 7637–7649. Retrieved from https://proceedings.neurips.cc/paper/2021/file/3ea2db50e62ceefceaf70a9d9a56a6f4-Paper.pdf

[78] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. 2019. Towards automatic concept-based explanations. In *NeurIPS*.

[79] Sanjukta Ghosh, Rohan Shet, Peter Amon, Andreas Hutter, and André Kaup. 2018. Robustness of deep convolutional neural networks for image degradations. In *ICASSP*. 2916–2920. DOI: https://doi.org/10.1109/ICASSP.2018.8461907

[80] Tejas Gokhale, Swaroop Mishra, Man Luo, Bhavdeep Sachdeva, and Chitta Baral. 2022. Generalized but not robust? Comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. In *Findings of the Association for Computational Linguistics (ACL'22)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2705–2718. DOI: https://doi.org/10.18653/v1/2022.findings-acl.213

[81] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. 2018. Making machine learning robust against adversarial inputs. *Communications of the ACM* 61, 7 (Jun 2018), 56–66. DOI: https://doi.org/10.1145/3134599

[82] Dou Goodman, Hao Xin, Wang Yang, Wu Yuesheng, Xiong Junfeng, and Zhang Huan. 2020. Advbox: A toolbox to generate adversarial examples that fool neural networks. arXiv:2001.05574. [cs.LG]. Retrieved from https://arxiv.org/abs/2001.05574

[83] Divya Gopinath, G. Katz, C. S. Păsăreanu, and Clark Barrett. 2018. Deepsafe: A data-driven approach for assessing robustness of neural networks. In *ATVA*. Springer, 3–19.

[84] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery* (28 Apr 2022), 1–55. DOI: https://doi.org/10.1007/s10618-022-00831-6

[85] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys* 51, 5, Article 93 (Aug 2018), 42 pages. DOI: https://doi.org/10.1145/3236009

[86] Sidharth Gupta, P. Dube, and Ashish Verma. 2020. Improving the affordability of robustness training for DNNs. In *CVPR*.

[87] Sireesh Gururaja, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. 2023. To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Houda Bouamor, Juan Pino, and Kalika Bali (Eds.), Association for Computational Linguistics, Singapore, 13310–13325. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.822

[88] C. Haase-Schutz, R. Stal, H. Hertlein, and B. Sick. 2021. Iterative label improvement: Robust training by confidence based filtering and dataset partitioning. In *25th International Conference on Pattern Recognition (ICPR'21)*. IEEE Computer Society, Los Alamitos, CA, USA, 9483–9490. DOI: https://doi.org/10.1109/ICPR48806.2021.9411918

[89] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. 2024. Machine learning with a reject option: A survey. *Machine Learning* 113, 5 (2024), 3073–3110. DOI: https://doi.org/10.1007/s10994-024-06534-x
[90] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. arXiv:1903.12261. [cs.LG]. DOI: https://doi.org/10.48550/ARXIV.1903.12261
[91] Patrick Henriksen, Kerstin Hammernik, Daniel Rueckert, and Alessio Lomuscio. 2021. Bias field robustness verification of large neural image classifiers. In *32nd British Machine Vision Conference (BMVC'21)*. BMVA Press, 202. Retrieved 22-June-2022 from https://www.bmvc2021-virtualconference.com/assets/papers/1291.pdf
[92] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531. [stat.ML]. DOI: https://doi.org/10.48550/ARXIV.1503.02531
[93] P. Hitzler and M. K. Sarker. 2022. Human-centered concept explanations for neural networks. *Neuro-Symbolic Artificial Intelligence: The State of the Art* 342, 337 (2022), 2.
[94] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *CHI*. 1–13.
[95] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.
[96] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26.
[97] Ramtin Hosseini, Xingyi Yang, and Pengtao Xie. 2021. DSRNA: Differentiable search of robust neural architectures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'21)*. Computer Vision Foundation/IEEE, 6196–6205. DOI: https://doi.org/10.1109/CVPR46437.2021.00613
[98] Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K. Thiruvathukal, and Ming Yin. 2020. Crowdsourcing detection of sampling biases in image datasets. In *WWW*. 2955–2961.
[99] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *2019 EMNLP-IJCNLP*. 2391–2401.
[100] Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *ISWC*. Springer, 486–504.
[101] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. 2018. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *SP*. IEEE, 19–35.
[102] Jongheon Jeong and Jinwoo Shin. 2020. Consistency regularization for certified robustness of smoothed classifiers. In *NeurIPS*. Vol. 33, Curran Associates, Inc., 10558–10570. Retrieved from https://proceedings.neurips.cc/paper/2020/file/77330e1330ae2b086e5bfcae50d9ffae-Paper.pdf
[103] Malhar Jere, Maghav Kumar, and Farinaz Koushanfar. 2020. A singular value perspective on model robustness. arXiv:2012.03516. [cs.CV]. Retrieved from https://arxiv.org/abs/2012.03516
[104] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys* 55, 12, Article 248 (Mar 2023), 38 pages. DOI: https://doi.org/10.1145/3571730
[105] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *AAAI*. Vol. 34, 8018–8025.
[106] Jonghoon Jin, Aysegul Dundar, and Eugenio Culurciello. 2015. Robust Convolutional Neural Networks under Adversarial Noise. arXiv:1511.06306. [cs.LG]. DOI: https://doi.org/10.48550/ARXIV.1511.06306
[107] Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. 2020. Graph structure learning for robust graph neural networks. In *SIGKDD*. ACM, New York, NY, USA, 66–74. DOI: https://doi.org/10.1145/3394486.3403049
[108] Takuhiro Kaneko, Yoshitaka Ushiku, and Tatsuya Harada. 2019. Label-noise robust generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*. Computer Vision Foundation/IEEE, Long Beach, CA, USA, 2467–2476. DOI: https://doi.org/10.1109/CVPR.2019.00257
[109] Arpan Kar and Amit Kumar Kushwaha. 2021. Facilitators and barriers of artificial intelligence adoption in business - insights from opinions using big data analytics. *Information Systems Frontiers* 25, 2 (2021), 16–21. DOI: https://doi.org/10.1007/s10796-021-10219-4
[110] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors. In *ICML*.
[111] Marvin Klingner, Andreas Bar, and Tim Fingscheidt. 2020. Improved noise and attack robustness for semantic segmentation by using multi-task training with self-supervised depth estimation. In *CVPR Workshops*.

[112] Pang Wei Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, Weihua Hu, Michihiro Yasunaga, R. L. Phillips, Irena Gao, Tony Lee, Etiene David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*. PMLR, 5637–5664.

[113] A. Kortylewski, Q. Liu, A. Wang, Y. Sun, and A. Yuille. 2021. Compositional convolutional neural networks: A robust and interpretable model for object recognition under occlusion. *I. Journal of Computer Vision* 129, 3 (2021), 736–760.

[114] Sean Kross and Philip Guo. 2021. Orienting, framing, bridging, magic, and counseling: How data scientists navigate the outer loop of client collaborations in industry and academia. *ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[115] Emanuele La Malfa and Marta Kwiatkowska. 2022. The king is naked: On the notion of robustness for natural language processing. In *AAAI*. Vol. 36, 11047–11057.

[116] Emanuele La Malfa, Min Wu, L. Laurenti, B. Wang, A. Hartshorn, and Marta Kwiatkowska. 2020. Assessing robustness of text classification through maximal safe radius computation. In *EMNLP*. ACL, 2949–2968. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.266

[117] Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. 2020. Robust and stable black box explanations. In *ICML (ICML'20)*. JMLR.org, Article 522, 11 pages. DOI: https://doi.org/10.5555/3524938.3525460

[118] Alfred Laugros, Alice Caplier, and Matthieu Ospici. 2020. Addressing neural network robustness with mixup and targeted labeling adversarial training. In *Computer Vision - ECCV 2020 Workshops - Glasgow*, Adrien Bartoli and Andrea Fusiello (Eds.). Lecture Notes in Computer Science, Springer, 178–195. DOI: https://doi.org/10.1007/978-3-030-68238-5_14

[119] Hyungyu Lee, Ho Bae, and Sungroh Yoon. 2021. Gradient masking of label smoothing in adversarial robustness. *IEEE Access* 9 (2021), 6453–6464. DOI: https://doi.org/10.1109/ACCESS.2020.3048120

[120] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *CHI*. 1–13.

[121] Klas Leino, Z. Wang, and M. Fredrikson. 2021. Globally-robust neural networks. In *ICML*. Vol. 139, PMLR, 6212–6222. Retrieved from https://proceedings.mlr.press/v139/leino21a.html

[122] Alexander Levine and Soheil Feizi. 2021. Improved, deterministic smoothing for L_1 certified robustness. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, Marina Meila and Tong Zhang (Eds.). Vol. 139, PMLR, 6254–6264. http://proceedings.mlr.press/v139/levine21a.html

[123] Dongfang Li, Baotian Hu, Qingcai Chen, Tujie Xu, Jingcong Tao, and Yunan Zhang. 2022. Unifying model explainability and robustness for joint text classification and rationale extraction. In *AAAI*. Vol. 36, 10947–10955.

[124] Dongyue Li and Hongyang Zhang. 2021. Improved regularization and robustness for fine-tuning in neural networks. In *NeurIPS*. Vol. 34, Curran Associates, Inc., 27249–27262. Retrieved from https://proceedings.neurips.cc/paper/2021/file/e4a93f0332b2519177ed55741ea4e5e7-Paper.pdf

[125] Jinfeng Li, Tianyu Du, Shouling Ji, Rong Zhang, Quan Lu, Min Yang, and Ting Wang. 2020. {TextShield}: Robust text classification based on multimodal embedding and neural machine translation. In *USENIX*. 1381–1398.

[126] Linyi Li, Zexuan Zhong, Bo Li, and Tao Xie. 2019. Robustra: Training provable robust neural networks over reference adversarial space.. In *IJCAI*. 4711–4717.

[127] Xin Li, Xiangrui Li, Deng Pan, and Dongxiao Zhu. 2021. Improving adversarial robustness via probabilistically compact loss with logit Cconstraints. In *Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI'21), Thirty-Third Conference on Innovative Applications of Artificial Intelligence (IAAI'21), The Eleventh Symposium on Educational Advances in Artificial Intelligence (EAAI'21), Virtual Event, February 2-9, 2021*, AAAI Press, 8482–8490. DOI: https://doi.org/10.1609/AAAI.V35I10.17030

[128] Yanxi Li, Zhaohui Yang, Yunhe Wang, and Chang Xu. 2021. Neural architecture dilation for adversarial robustness. In *NeurIPS*. Vol. 34, Curran Associates, Inc., 29578–29589. Retrieved from https://proceedings.neurips.cc/paper/2021/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf

[129] Zhimin Li, Shusen Liu, Xin Yu, Kailkhura Bhavya, Jie Cao, Diffenderfer James Daniel, Peer-Timo Bremer, and Valerio Pascucci. 2023. *"Understanding Robustness Lottery": A Geometric Visual Comparative Analysis of Neural Network Pruning Approaches*.

[130] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *CHI*. 1–15.

[131] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. 2020. Towards hybrid human-AI workflows for unknown unknown detection. In *WWW*. 2432–2442.

[132] Yang Lou, Ruizi Wu, Junli Li, Lin Wang, Xiang Li, and Guanrong Chen. 2023. A learning convolutional neural network approach for network robustness prediction. *IEEE Transactions on Cybernetics* 53, 7 (2023), 4531–4544. DOI: https://doi.org/10.1109/TCYB.2022.3207878

[133] Wolfgang Maass. 1997. Networks of spiking neurons: The third generation of neural network models. *Neural Networks* 10, 9 (1997), 1659–1671. DOI: https://doi.org/10.1016/S0893-6080(97)00011-7

[134] Divyam Madaan, Jinwoo Shin, and Sung Ju Hwang. 2021. Learning to generate noise for multi-attack robustness. In *Proceedings of the 38th International Conference on Machine Learning (ICML'21)*, Vol. 139, PMLR, 7279–7289. Retrieved from http://proceedings.mlr.press/v139/madaan21a.html

[135] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *ICLR*. OpenReview.net. Retrieved from https://openreview.net/forum?id=rJzIBfZAb

[136] Ravi Mangal, Aditya V. Nori, and Alessandro Orso. 2019. Robustness of neural networks: a probabilistic and practical approach. In *Proceedings of the 41st International Conference on Software Engineering: New Ideas and Emerging Results, ICSE (NIER) 2019, Montreal, QC, Canada, May 29-31, 2019*, IEEE/ACM, 93–96. DOI: https://doi.org/10.1109/ICSE-NIER.2019.00032

[137] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. 2019. Metric learning for adversarial robustness. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 478–489. Retrieved from https://proceedings.neurips.cc/paper/2019/hash/c24cd76e1ce41366a4bbe8a49b02a028-Abstract.html

[138] Gary Marcus. 2020. *The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence*.

[139] Alexander Mathis, Thomas Biasi, Steffen Schneider, Mert Yuksekgonul, Byron Rogers, Matthias Bethge, and Mackenzie W. Mathis. 2021. Pretraining boosts out-of-domain robustness for pose estimation. In *WACV*. 1859–1868.

[140] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6 (2021), 1–35.

[141] Brad Miller, Alex Kantchelian, Sadia Afroz, Rekha Bachwani, E. Dauber, L. Huang, M. C. Tschantz, A. D. Joseph, and J. Doug Tygar. 2014. Adversarial active learning. In *AISec*. 3–14.

[142] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. 2018. Dropout sampling for robust object detection in open-set conditions. In *ICRA*. 3243–3249. DOI: https://doi.org/10.1109/ICRA.2018.8460700

[143] David J. Miller, Xinyi Hu, Zhicong Qiu, and George Kesidis. 2017. Adversarial learning: A critical review and active learning study. In *MLSP*. IEEE, 1–6.

[144] John P. Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *ICML*. PMLR, 7721–7735.

[145] Matthew Mirman, Timon Gehr, and Martin Vechev. 2018. Differentiable abstract interpretation for provably robust neural networks. In *ICML*. Vol. 80, PMLR, 3578–3586. Retrieved from https://proceedings.mlr.press/v80/mirman18b.html

[146] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *FAccT*. 220–229.

[147] Jisoo Mok, Byunggook Na, Hyeokjun Choe, and Sungroh Yoon. 2021. AdvRush: Searching for adversarially robust neural architectures. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, IEEE, 12302–12312. DOI: https://doi.org/10.1109/ICCV48922.2021.01210

[148] Mohammad Momeny, Ali Mohammad Latif, Mehdi Agha Sarram, Razieh Sheikhpour, and Yu Dong Zhang. 2021. A noise robust convolutional neural network for image classification. *Results in Engineering* 10, 2 (2021), 100225. DOI: https://doi.org/10.1016/j.rineng.2021.100225

[149] Seyed-Mohsen Moosavi-Dezfooli, Ashish Shrivastava, and Oncel Tuzel. 2019. *Divide, Denoise, and Defend against Adversarial Attacks*.

[150] Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event/Punta Cana, Dominican Republic, 7-11 November, 2021*, Association for Computational Linguistics, 1558–1570. DOI: https://doi.org/10.18653/V1/2021.EMNLP-MAIN.117

[151] Aamir Mustafa, S. H. Khan, M. Hayat, R. Goecke, Jianbing Shen, and Ling Shao. 2021. Deeply supervised discriminative learning for adversarial defense. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 9 (2021), 3154–3166. DOI: https://doi.org/10.1109/TPAMI.2020.2978474

[152] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. Collaboration challenges in building ML-enabled systems: Communication, documentation, engineering, and process. In *ICSE*. 413–425.

[153] Vedant Nanda, Till Speicher, John P. Dickerson, Krishna P. Gummadi, and Muhammad Bilal Zafar. 2022. Unifying model explainability and robustness for joint text classification and rationale extraction. In *AAAI*. Vol. 36, 10947–10955.

[154] Shweta Narkar, Yunfeng Zhang, Q. Vera Liao, Dakuo Wang, and Justin D. Weisz. 2021. Model LineUpper: Supporting interactive model comparison at multiple levels for AutoML. In *IUI*. 170–174.

[155] Muzammal Naseer, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. 2020. A Self-supervised Approach for Adversarial Robustness. In *2020 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation/IEEE, 259–268. DOI: https://doi.org/10.1109/CVPR42600.2020.00034

[156] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. 2017. Exploring Generalization in deep learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017*, Long Beach, CA, USA, 5947–5956. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/10ce03a1ed01077e3e289f3e53c72813-Abstract.html

[157] Kun-Peng Ning, Lue Tao, Songcan Chen, and Sheng-Jun Huang. 2021. Improving model robustness by adaptively correcting perturbation levels with active queries. In *EAAI*. AAAI Press, 9161–9169. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/17106

[158] Ardavan Salehi Nobandegani, Kevin da Silva Castanheira, Timothy O'Donnell, and Thomas R. Shultz. 2019. On robustness: An undervalued dimension of human rationality. In *CogSci*. 3327.

[159] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS*.

[160] Mehdi Nourelahi, Lars Kotthoff, Peijie Chen, and Anh Nguyen. 2023. How explainable are adversarially-robust CNNs?

[161] Besmira Nushi, Ece Kamar, Eric Horvitz, and Donald Kossmann. 2017. On human intellect and machine failures: Troubleshooting integrative machine learning systems. In *AAAI*.

[162] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. 2022. Robustness and accuracy could be reconcilable by (Proper) definition. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research)*, PMLR, 17258–17277. Retrieved from https://proceedings.mlr.press/v162/pang22a.html

[163] Tianyu Pang, Huishuai Zhang, Di He, Yinpeng Dong, Hang Su, Wei Chen, Jun Zhu, and Tie-Yan Liu. 2022. Two coupled rejection metrics can tell adversarial examples apart. In *CVPR*. 15223–15233.

[164] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE SP*. 582–597. DOI: https://doi.org/10.1109/SP.2016.41

[165] Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. 2018. *Generalizability vs. Robustness: Adversarial Examples for Medical Imaging*.

[166] Martin Pawelczyk, Chirag Agarwal, Shalmali Joshi, Sohini Upadhyay, and Himabindu Lakkaraju. 2022. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *AISTATS*. Vol. 151, PMLR, 4574–4594. Retrieved from https://proceedings.mlr.press/v151/pawelczyk22a.html

[167] Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, IEEE, 9616–9625. DOI: https://doi.org/10.1109/ICCV.2019.00971

[168] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In *ACL*. ACL, Minneapolis, Minnesota, 3336–3347. DOI: https://doi.org/10.18653/v1/N19-1337

[169] Maura Pintor, Daniele Angioni, Angelo Sotgiu, Luca Demetrio, Ambra Demontis, Battista Biggio, and Fabio Roli. 2023. ImageNet-Patch: A dataset for benchmarking machine learning robustness against adversarial patches. *Pattern Recognit.* 134 (2023), 109064. DOI: https://doi.org/10.1016/J.PATCOG.2022.109064

[170] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How AI developers overcome communication challenges in a multidisciplinary team: A case study. *ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[171] Yada Pruksachatkun, S. Krishna, J. Dhamala, R. Gupta, and Kai-Wei Chang. 2021. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. In *ACL-IJCNLP*. 3320–3331. DOI: https://doi.org/10.18653/v1/2021.findings-acl.294

[172] Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *ACL*. 5582–5591. DOI: https://doi.org/10.18653/v1/P19-1561

[173] Yanmin Qian, Hu Hu, and Tian Tan. 2019. Data augmentation using generative adversarial networks for robust speech recognition. *Speech Communication* 114, 9 (2019), 1–9. DOI: https://doi.org/10.1016/j.specom.2019.08.006

[174] R. Hamon, H. Junklewitz, and J. I. Sanchez Martin. 2020. Robustness and explainability of artificial intelligence. KJ-1407 NA-30040-EN-N (online) (2020). Publications Office. DOI: https://doi.org/10.2760/57493. (online).

[175] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. 2020. Understanding and Mitigating the Tradeoff between Robustness and Accuracy. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event (Proceedings of Machine Learning Research)*, PMLR, 7909–7919. Retrieved from http://proceedings.mlr.press/v119/raghunathan20a.html

[176] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[177] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *J. Mach. Learn. Res.* 11 (2010), 1297–1322. DOI: https://doi.org/10.5555/1756006.1859894

[178] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. 2021. Robust fairness under covariate shift. *AAAI* 35, 11 (May 2021), 9419–9427. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/17135

[179] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *SIGKDD*. 1135–1144.

[180] Brianna Richardson, Jean Garcia-Gathright, Samuel F. Way, Jennifer Thom, and Henriette Cramer. 2021. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ML toolkits. In *CHI*. 1–13.

[181] Matthew Richardson and Pedro Domingos. 2006. Markov logic networks. *Machine Learning* 62, 1 (2006), 107–136.

[182] Christos Rodosthenous and Loizos Michael. 2016. A hybrid approach to commonsense knowledge acquisition. In *STAIRS 2016*. IOS Press, 111–122.

[183] Sudipta Singha Roy, Sk. Imran Hossain, M. A. H. Akhand, and Kazuyuki Murase. 2018. A robust system for noisy image classification combining denoising autoencoder and convolutional neural network. *International Journal of Advanced Computer Science and Applications* 9, 1 (2018), 224–235. DOI: https://doi.org/10.14569/IJACSA.2018.090131

[184] Andras Rozsa, Manuel Günther, and Terrance E. Boult. 2018. Towards robust deep neural networks with bANG. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV'18), Lake Tahoe, NV, USA, March 12-15, 2018*, IEEE Computer Society, 803–811. DOI: https://doi.org/10.1109/WACV.2018.00093

[185] Wenjie Ruan, Min Wu, Youcheng Sun, Xiaowei Huang, Daniel Kroening, and Marta Kwiatkowska. 2019. Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI'19), Macao, China, August 10-16, 2019*, ijcai.org, 5944–5952. DOI: https://doi.org/10.24963/IJCAI.2019/824

[186] Evgenia Rusak, Lukas Schott, Roland S. Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. 2020. A Simple way to make neural networks robust against diverse image corruptions. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III* (Lecture Notes in Computer Science), Springer, 53–69. DOI: https://doi.org/10.1007/978-3-030-58580-8_4

[187] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *2021 CHI*. 1–15.

[188] Wojciech Samek and Klaus-Robert Müller. 2019. Towards explainable artificial intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, Klaus-Robert, and Müller (Eds.). Springer International Publishing, 5–22.

[189] Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *ACL*. ACL, New Orleans, Louisiana, 1975–1985. DOI: https://doi.org/10.18653/v1/N18-1179

[190] Filippo Santoni de Sio. 2021. The European Commission report on ethics of connected and automated vehicles and the future of ethics of transportation. *Ethics and Information Technology* 23, 4 (2021), 713–726.

[191] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. SocialIQA: Commonsense reasoning about social interactions. In *EMNLP*.

[192] Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. 2022. Robust learning meets generative models: Can proxy distributions improve adversarial robustness? In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net. Retrieved from https://openreview.net/forum?id=WVX0NNVBBkV

[193] Shreya Shankar, Rolando Garcia, Joseph M. Hellerstein, and Aditya G. Parameswaran. 2022. Operationalizing machine learning: An interview study. arXiv:2209.09125. Retrieved from https://arxiv.org/abs/2209.09125

[194] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. 2021. Do image classifiers generalize across time? In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, IEEE, 9641–9649. DOI: https://doi.org/10.1109/ICCV48922.2021.00952

[195] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. 2022. On the adversarial robustness of vision transformers. *Trans. Mach. Learn. Res.* 2022 (2022). Retrieved from https://openreview.net/forum?id=lE7K4n1Esk

[196] Shahin Sharifi Noorian, S. Qiu, U. Gadiraju, J. Yang, and Alessandro Bozzon. 2022. What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition. In *WWW*. 882–892.

[197] Saima Sharmin, Nitin Rathi, Priyadarshini Panda, and Kaushik Roy. 2020. Inherent adversarial robustness of deep spiking neural networks: Effects of discrete input encoding and non-linear activations. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIX* (Lecture Notes in Computer Science), Springer, 399–414. DOI: https://doi.org/10.1007/978-3-030-58526-6_24

[198] Max W. Shen. 2022. Trust in AI: Interpretability is not necessary or sufficient, while black-box interaction is necessary and sufficient.

[199] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2023. Towards Out-Of-distribution generalization: A survey.
[200] Dule Shu, Nandi O. Leslie, Charles A. Kamhoua, and Conrad S. Tucker. 2020. Generative adversarial attacks against intrusion detection systems using active learning. In *WiseML*. 1–6.
[201] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the ACL: EMNLP 2021*. ACL, 3784–3803. DOI: https://doi.org/10.18653/v1/2021.findings-emnlp.320
[202] K. Simonyan, A. Vedaldi, and A. Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*.
[203] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin T. Vechev. 2019. Boosting robustness certification of neural networks. In *ICLR*.
[204] Vinay Singh, Iuliia Konovalova, and Arpan Kumar Kar. 2023. When to choose ranked area integrals versus integrated gradient for explainable artificial intelligence – a comparison of algorithms. *Benchmarking: An International Journal* 30, 9 (01 Jan 2023), 3067–3089. DOI: https://doi.org/10.1108/BIJ-02-2022-0112
[205] Sahil Singla, Surbhi Singla, and Soheil Feizi. 2022. Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100. In *ICLR*. Retrieved from https://openreview.net/forum?id=tD7eCtaSkR
[206] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational adversarial active learning. In *ICCV*. 5972–5981.
[207] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *AIES'20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, ACM, 180–186. DOI: https://doi.org/10.1145/3375627.3375830
[208] Carol J. Smith. 2019. Designing trustworthy AI: A human-machine teaming framework to guide development. arXiv:1910.03515. Retrieved from https://arxiv.org/abs/1910.03515
[209] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *2020 FAccT*. 56–67.
[210] Chang Song, Elias Fallon, and Hai Li. 2021. *Improving Adversarial Robustness in Weight-quantized Neural Networks*.
[211] Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *AAAI*. Vol. 36, 11349–11357.
[212] Matthew Staib and Stefanie Jegelka. 2017. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, 4.
[213] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *NIPS*. Vol. 33. Curran Associates, Inc., 3008–3021. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf
[214] E. Štrumbelj and I. Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* (2014).
[215] David Stutz, Matthias Hein, and Bernt Schiele. 2020. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *ICML* 119 (2020), 9155–9166.
[216] Dong Su, H. Zhang, H. Chen, J. Yi, Pin-Yu Chen, and Yupeng Gao. 2018. Is robustness the cost of accuracy? A comprehensive study on the robustness of 18 deep image classification models. In *ECCV*. Springer, Cham, 644–661.
[217] Sayma Suha and Tahsina Sanam. 2023. Exploring dominant factors for ensuring the sustainability of utilizing artificial intelligence in healthcare decision making: An emerging country context. *International Journal of Information Management Data Insights* 3, 1 (04 2023), 100170. DOI: https://doi.org/10.1016/j.jjimei.2023.100170
[218] Ke Sun, Zhanxing Zhu, and Zhouchen Lin. 2019. *Enhancing the Robustness of Deep Neural Networks by Boundary Conditional GAN*.
[219] Weidi Sun, Yuteng Lu, Xiyue Zhang, and Meng Sun. 2022. DeepGlobal: A framework for global robustness verification of feedforward neural networks. *J. Syst. Archit.* 128 (2022), 102582. DOI: https://doi.org/10.1016/J.SYSARC.2022.102582
[220] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *ICML*.
[221] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Retrieved from http://arxiv.org/abs/1312.6199
[222] Jihoon Tack, Sihyun Yu, Jongheon Jeong, Minseon Kim, Sung Ju Hwang, and Jinwoo Shin. 2022. Consistency regularization for adversarial robustness. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 8 (Jun. 2022), 8414–8422. DOI: https://doi.org/10.1609/aaai.v36i8.20817
[223] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, Philip H. S. Torr, and Dacheng Tao. 2022. *RobustART: Benchmarking Robustness on Architecture Design and Training Techniques*.

[224] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*. Vol. 33, Curran Associates, 18583–18599. Retrieved from https://proceedings.neurips.cc/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf

[225] Matteo Terzi, Gian Antonio Susto, and Pratik Chaudhari. 2020. Directional adversarial training for cost sensitive deep learning classification applications. *Engineering Applications of Artificial Intelligence* 91, 5 (2020), 103550.

[226] Dang Duy Thang and Toshihiro Matsui. 2019. Image transformation can make neural networks more robust against adversarial examples.

[227] Vincent Tjeng, Kai Yuanqing Xiao, and Russ Tedrake. 2019. Evaluating robustness of neural networks with mixed integer programming. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net. Retrieved from https://openreview.net/forum?id=HyGIdiRqtm

[228] Yu-Lin Tsai, Chia-Yi Hsu, Chia-Mu Yu, and Pin-Yu Chen. 2021. Formalizing generalization and adversarial robustness of neural networks to weight perturbations. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 19692–19704. Retrieved from https://proceedings.neurips.cc/paper/2021/hash/a3ab4ff8fa4deed2e3bae3a5077675f0-Abstract.html

[229] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2018. Robustness may be at odds with accuracy. In *ICLR*.

[230] Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and Pushmeet Kohli. 2019. *Are Labels Required for Improving Adversarial Robustness?* Curran Associates Inc., Red Hook, NY, USA. DOI: https://doi.org/10.5555/3454287.3455381

[231] Meet P. Vadera, Satya Narayan Shukla, Brian Jalaian, and Benjamin M. Marlin. 2020. Assessing the adversarial robustness of monte carlo and distillation methods for deep bayesian neural network classification.

[232] Pratik Vaishnavi, Tianji Cong, Kevin Eykholt, Atul Prakash, and Amir Rahmati. 2019. *Can Attention Masks Improve Adversarial Robustness?*

[233] Colin Vandenhof. 2019. A hybrid approach to identifying unknown unknowns of predictive models. In *HCOMP*. Vol. 7, 180–187.

[234] Jennifer Wortman Vaughan. 2018. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research* 18, 193 (2018), 1–46.

[235] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *CHI*. 1–14.

[236] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Fairware*. IEEE, 1–7.

[237] Marco Virgolin and Saverio Fracaros. 2023. On the robustness of sparse counterfactual explanations to adverse perturbations. *Artificial Intelligence* 316, 3 (2023), 103840. DOI: https://doi.org/10.1016/j.artint.2022.103840

[238] Luis Von Ahn, M. Kedia, and M. Blum. 2006. Verbosity: A game for collecting common-sense facts. In *SIGCHI*. 75–78.

[239] Binghui Wang, Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. 2021. Certified robustness of graph neural networks against adversarial structural perturbation. In *KDD'21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, ACM, 1645–1653. DOI: https://doi.org/10.1145/3447548.3467295

[240] Huaxia Wang and Chun-Nam Yu. 2019. A direct approach to robust deep learning using adversarial networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net. Retrieved from https://openreview.net/forum?id=S1lIMn05F7

[241] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Weirong Ye, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xingxu Xie. 2023. On the robustness of ChatGPT: An adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

[242] Jiakai Wang, Zixin Yin, Pengfei Hu, Aishan Liu, Renshuai Tao, Haotong Qin, Xianglong Liu, and Dacheng Tao. 2022. Defensive patches for robust recognition in the physical world. In *CVPR*. 2456–2465.

[243] Lijie Wang, Hao Liu, Shuyuan Peng, Hongxuan Tang, Xinyan Xiao, Ying Chen, Hua Wu, and Haifeng Wang. 2021. DuTrust: A Sentiment Analysis Dataset for Trustworthiness Evaluation.

[244] Serena Lutong Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya R. Gupta, and Michael I. Jordan. 2020. Robust optimization for fairness with noisy protected groups. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Retrieved from https://proceedings.neurips.cc/paper/2020/hash/37d097caf1299d9aa79c2c2b843d2d78-Abstract.html

[245] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. 2020. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*.

[246] Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *EAAI*. AAAI Press, 14024–14031. Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/17651

[247] Stefan Webb, Tom Rainforth, Yee Whye Teh, and M. Pawan Kumar. 2019. A statistical approach to assessing neural network robustness. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net. Retrieved from https://openreview.net/forum?id=S1xcx3C5FX

[248] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*. Vol. 35, 24824–24837. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[249] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. 2018. Evaluating the robustness of neural networks: An extreme value theory approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net. Retrieved from https://openreview.net/forum?id=BkUHlMZ0b

[250] Arie Wahyu Wijayanto, Jun Jin Choong, Kaushalya Madhawa, and Tsuyoshi Murata. 2019. Towards robust compressed convolutional neural networks. In *BigComp*. 1–8. DOI: https://doi.org/10.1109/BIGCOMP.2019.8679132

[251] Eric Wong and J. Zico Kolter. 2021. Learning perturbation sets for robust machine learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net. Retrieved from https://openreview.net/forum?id=MIDckA56aD

[252] Walt Woods, Jack Chen, and Christof Teuscher. 2019. Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nature Machine Intelligence* 1, 11 (2019), 508–516.

[253] Chenwang Wu, Defu Lian, Yong Ge, Zhihao Zhu, Enhong Chen, and Senchao Yuan. 2021. Fight fire with fire: Towards robust recommender systems via adversarial poisoning training. In *SIGIR '21*. ACM, 1074–1083. DOI: https://doi.org/10.1145/3404835.3462914

[254] Yiting Wu and Min Zhang. 2021. Tightening robustness verification of convolutional neural networks with fine-grained linear approximation. In *AAAI*. Vol. 35, 11674–11681.

[255] Pulei Xiong, Scott Buffett, Shahrear Iqbal, Philippe Lamontagne, Mohammad Mamun, and Heather Molyneaux. 2022. Towards a robust and trustworthy machine learning system development: An engineering perspective. *Journal of Information Security and Applications* 65, 2 (2022), 103121.

[256] Cong Xu, Xiang Li, and Min Yang. 2022. An orthogonal classifier for improving the adversarial robustness of neural networks. *Information Sciences* 591, 10 (2022), 251–262. DOI: https://doi.org/10.1016/j.ins.2022.01.039

[257] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. 2021. To be robust or to be fair: Towards fairness in adversarial training. In *ICML*. Vol. 139, PMLR, 11492–11501. Retrieved from https://proceedings.mlr.press/v139/xu21b.html

[258] Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. 2019. L_DMI: A novel information-theoretic loss function for training deep nets robust to label noise. In *NeurIPS*. Vol. 32, Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2019/file/8a1ee9f2b7abe6e88d1a479ab6a42c5e-Paper.pdf

[259] Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G. Dy. 2011. Active learning from crowds. In *ICML*. 1161–1168.

[260] Ziang Yan, Yiwen Guo, and Changshui Zhang. 2018. Deep Defense: Training DNNs with improved adversarial robustness. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 417–426. Retrieved from https://proceedings.neurips.cc/paper/2018/hash/8f121ce07d74717e0b1f21d122e04521-Abstract.html

[261] J. Yang, T. Drake, A Damianou, and Y. Maarek. 2018. Leveraging crowdsourcing data for deep active learning. An application: Learning intents in Alexa. In *WWW*.

[262] Jie Yang, Alisa Smirnova, Dingqi Yang, Gianluca Demartini, Yuan Lu, and Philippe Cudre-Mauroux. 2019. Scalpel-CD: Leveraging crowdsourcing and deep probabilistic modeling for debugging noisy training data. In *WWW*. 2158–2168.

[263] Pengfei Yang, J. Li, J. Liu, C.-C. Huang, R. Li, L. Chen, X. Huang, and Lijun Zhang. 2021. Enhancing robustness verification for deep neural networks via symbolic propagation. *Formal Aspects of Computing* 33, 3 (06 2021), 407–435. DOI: https://doi.org/10.1007/s00165-021-00548-1

[264] Yichen Yang, Xiaosen Wang, and Kun He. 2022. Robust textual embedding against word-level adversarial attacks. In *Uncertainty in Artificial Intelligence, Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, UAI 2022, 1-5 August 2022, Eindhoven, The Netherlands (Proceedings of Machine Learning Research)*, PMLR, 2214–2224. Retrieved from https://proceedings.mlr.press/v180/yang22c.html

[265] Muneki Yasuda, Hironori Sakata, Seung-Il Cho, Tomochika Harada, Atushi Tanaka, and Michio Yokoyama. 2019. An efficient test method for noise robustness of deep neural networks. *Nonlinear Theory and Its Applications, IEICE* 10, 2 (01 2019), 221–235. DOI: https://doi.org/10.1587/nolta.10.221

[266] Dengpan Ye, Chuanxi Chen, Changrui Liu, Hao Wang, and Shunzhi Jiang. 2021. Detection defense against adversarial attacks with saliency map. *International Journal of Intelligent Systems* 37, 12 (2021), 10193–10210.

[267] Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen. 2019. Interpreting and evaluating neural network robustness. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, ijcai.org, 4199–4205. DOI: https://doi.org/10.24963/IJCAI.2019/583

[268] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. Adversarial collaborative neural network for robust recommendation. In *SIGIR'19*. ACM, New York, NY, USA, 1065–1068. DOI: https://doi.org/10.1145/3331184.3331321
[269] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. 2020. Training individually fair ML models with Sensitive Subspace Robustness.
[270] Liang-Jun Zang, Cong Cao, Ya-Nan Cao, Yu-Ming Wu, and Cun-Gen Cao. 2013. A survey of commonsense knowledge acquisition. *Journal of Computer Science and Technology* 28, 4 (2013), 689–719.
[271] Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. 2019. *Adversarially Robust Generalization Just Requires More Unlabeled Data.*
[272] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? Roles, workflows, and tools. *ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.
[273] Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. 2021. Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Transactions on Image Processing* 30, 30–121 (2021), 1291–1304. DOI: https://doi.org/10.1109/tip.2020.3042083
[274] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. 2018. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 4944–4953. Retrieved from https://proceedings.neurips.cc/paper/2018/hash/d04863f100d59b3eb688a11f95b0ae60-Abstract.html
[275] Jie M. Zhang, Mark Harman, Lei Ma, and Yang Liu. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* 48, 1 (2020), 1–36.
[276] Li Zhang and Haiping Lu. 2020. A feature-importance-aware and robust aggregator for GCN. In *CIKM*. ACM, 1813–1822. DOI: https://doi.org/10.1145/3340531.3411983
[277] Marvin Zhang, Sergey Levine, and Chelsea Finn. 2022. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35, Curran Associates, Inc., 38629–38642. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2022/file/fc28053a08f59fccb48b11f2e31e81c7-Paper-Conference.pdf
[278] Mengdi Zhang, Jun Sun, and Jingyi Wang. 2022. Which neural network makes more explainable decisions? An approach towards measuring explainability. *Automated Software Engineering* 29, 2 (09 Apr 2022), 39. DOI: https://doi.org/10.1007/s10515-022-00338-w
[279] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. 2018. Interpretable convolutional neural networks. In *CVPR*.
[280] Xiao Zhang and David E. Evans. 2022. Understanding intrinsic robustness using label uncertainty. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net. Retrieved from https://openreview.net/forum?id=6ET9SzlgNX
[281] Yuhao Zhang, Aws Albarghouthi, and Loris D'Antoni. 2021. Certified robustness to programmable transformations in LSTMs. In *EMNLP*. ACL, 1068–1083. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.82
[282] Long Zhao, Ting Liu, Xi Peng, and Dimitris N. Metaxas. 2020. Maximum-entropy adversarial data augmentation for improved generalization and robustness. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Retrieved from https://proceedings.neurips.cc/paper/2020/hash/a5bfc9e07964f8dddeb95fc584cd965d-Abstract.html
[283] Qinkai Zheng, Xu Zou, Yuxiao Dong, Yukuo Cen, Da Yin, Jiarong Xu, Yang Yang, and Jie Tang. 2021. Graph robustness benchmark: Benchmarking the adversarial robustness of graph machine learning. In *NeurIPS*. Retrieved from https://openreview.net/forum?id=NxWUnvwFV4
[284] Xiaoqing Zheng, J. Zeng, Y. Zhou, C.-J. Hsieh, Minhao Cheng, and Xuanjing Huang. 2020. Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples. In *ACL*. 6600–6610. DOI: https://doi.org/10.18653/v1/2020.acl-main.590
[285] Yiqi Zhong, Lei Wu, Xianming Liu, and Junjun Jiang. 2022. *Exploiting the Potential of Datasets: A Data-Centric Approach for Model Robustness.*
[286] Shuyan Zhou, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. 2019. Improving robustness of neural machine translation with multi-task learning. In *WMT*. 565–571.
[287] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. 2021. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *ICCV*. 16443–16452.
[288] Vadim Ziyadinov and Maxim Tereshonok. 2022. Noise immunity and robustness study of image recognition using a convolutional neural network. *Sensors* 22, 3 (2022), 1241. DOI: https://doi.org/10.3390/s22031241
[289] Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, and Pushmeet Kohli. 2020. Towards robust image classification using sequential attention models. In *CVPR*.

Received 15 October 2022; revised 21 December 2023; accepted 14 May 2024