Specification Overfitting in Artificial Intelligence

Benjamin Roth[1,2†], Pedro Henrique Luz de Araujo[1,3*†],
Yuxi Xia[1,3], Saskia Kaltenbrunner[4], Christoph Korab[4]

[1]Faculty of Computer Science, University of Vienna, Vienna, Austria.
[2]Faculty of Philological and Cultural Studies, University of Vienna, Vienna, Austria.
[3]UniVie Doctoral School Computer Science, University of Vienna,Vienna, Austria.
[4]Department of Innovation and Digitalisation in Law, University of Vienna, Vienna, Austria.

*Corresponding author(s). E-mail(s): pedro.henrique.luz.de.araujo@univie.ac.at;
Contributing authors: benjamin.roth@univie.ac.at; yuxi.xia@univie.ac.at; saskia.kaltenbrunner@univie.ac.at; christoph.korab@univie.ac.at;
†These authors contributed equally to this work.

Abstract

Machine learning (ML) and artificial intelligence (AI) approaches are often criticized for their inherent bias and for their lack of control, accountability, and transparency. Consequently, regulatory bodies struggle with containing this technology's potential negative side effects. High-level requirements such as fairness and robustness need to be formalized into concrete specification metrics, imperfect proxies that capture isolated aspects of the underlying requirements. Given possible trade-offs between different metrics and their vulnerability to over-optimization, integrating specification metrics in system development processes is not trivial. This paper defines *specification overfitting*, a scenario where systems focus excessively on specified metrics to the detriment of high-level requirements and task performance. We present an extensive literature survey to categorize how researchers propose, measure, and optimize specification metrics in several AI fields (e.g., natural language processing, computer vision, reinforcement learning). Using a keyword-based search on papers from major AI conferences and journals between 2018 and mid-2023, we identify and analyze 74 papers that propose or optimize specification metrics. We find that although most papers implicitly address specification overfitting (e.g., by reporting more than one specification metric), they rarely discuss which role specification metrics should play in system development or explicitly define the scope and assumptions behind metric formulations.

Keywords: specification, overfitting, fairness, robustness, regulation, artificial intelligence

1 Introduction

The classical way (Shalev-Shwartz and Ben-David, 2014) of measuring the performance of predictive systems only on held-out data has been identified as inadequate to fully reflect the complexities of real-world use cases (Ribeiro et al, 2020), where it may be required that an algorithm fulfills additional properties that are not sufficiently reflected by reporting average performance metrics such as accuracy on the held-out data.

States, companies, and non-profit organizations have formulated ethical principles and high-level guidelines for AI (Fjeld et al, 2020; Hagendorff, 2020; Jobin et al, 2019). Laws and regulations are being formulated to make adherence to such principles legally binding (Wachter et al, 2017; Barocas and Selbst, 2016). The EU legal framework for AI (Veale and Borgesius, 2021) requires certification of AI systems on the basis of such laws. However, the details of such regulations are often (implicitly) relegated to standardization organizations (e.g., DIN, ETSI, ISO, NIST[1]), and since laws on regulating AI are very recent (or still in the making) there is little experience in how to translate the higher-level principles into low-level evaluation scenarios.

High-level guidelines may be formalized narrowly into concrete specifications and metrics, a process that requires making assumptions—*what* aspects of the underlying goal should be measured and *how* should they be measured—that can introduce mismatches between high-level principles and their measurements (Jacobs and Wallach, 2021). Moreover, an aspect that has not received enough attention is the question of what role the outcome of an evaluation w.r.t. additional specifications should play in the larger development cycle of AI systems. Most scientific publications (as we will show in our analysis) discussing the use of additional specification metrics for AI systems do not address the question of whether those additional feedback metrics can and should be used during system development and little research has been undertaken to study the effects of considering specification metrics in AI system development. As a step towards raising awareness of those questions, we provide a comprehensive first overview of papers that consider additional specification metrics, and we catalog the training and evaluation setups that are common when measurements of additional requirements are included in AI and ML scenarios.

There is a long tradition of discussing the potential of unintended consequences and misalignment of goals for AI systems (Wiener, 1960). Recently, Malik (2020) discusses the sacrifices and pitfalls of translating open-ended and qualitative questions into quantitative machine learning settings. However, those works do not discuss how

[1] https://www.din.de/, https://www.etsi.org/, https://www.iso.org/, https://www.nist.gov/

current AI research deals with *competing* objectives and possible feedback loops that include measurements of *additional* specifications. Similarly, work on testing in ML and AI, e.g., Zhang et al (2022a), studies different properties that ML and AI systems can be evaluated for and different ways to do it, but there is no guidance on how to integrate *different* metrics in the AI development process. Work on underspecification in AI has shown how equivalent predictors —with the same test set task accuracy—can exhibit widely different behaviors on single instances or for properties not reflected in the held-out data. D'Amour et al (2022) remark that one should use specifications to enforce desired behaviors and use them to select systems in such cases, but they do not discuss how to resolve disagreements between different specifications and how to prevent overfitting to specifications (see below).

Our study is the first to define *specification overfitting*, the case of overfitting to desirable outcomes specified additionally besides the task metric. Whereas misalignment and underspecification concern the specification of a faulty main objective and the failure to specify additional desirable properties, specification overfitting describes a scenario where specification metrics improve to the detriment of the main task metric or other specifications. For this, we are the first to comprehensively categorize common practices in scenarios with specification metrics, i.e., different, possibly competing measurements of additional properties besides the main task metric (e.g., accuracy on identically and independently distributed held-out data). We cover papers from several fields, including NLP, computer vision, and reinforcement learning, and we provide a quantitative and qualitative analysis of the methods and recommendations in those papers.

We build our analysis on the results of a key-word-based search for papers from DBLP[2] (Ley, 2002) covering the main conferences in NLP, CV, and AI, from 1/2018 – 7/2023, retrieving those papers that deal with scenarios where additional requirements are measured in addition to a task metric. We stratified those papers to have equal coverage in 3 groups of application domains (NLP, Vision, Other) by ranking and filtering, keeping the most cited papers. We keep the resulting set of 74 papers for our in-depth analysis. We analyze how the additional requirements and specifications are reflected in the training and evaluation procedure described, encoding it in a categorical schema. We report this fine-grained categorical analysis, together with an aggregate overview, and discuss representative and interesting findings.

Of all 74 papers that measure an additional specification, 62 papers also attempt to improve on that metric. Forty-eight papers study the effect of the attempt to improve this metric on other metrics (including the task metric).

We find that most papers (59) do not recommend how to use the specifications' feedback in the development process. Of the ones that do, four recommend delegating the decision of how to use specification metrics to an expert, and three recommend using the feedback of specifications for debugging. Only one (Pfohl et al, 2022a) provides a concrete recommendation on how a specification should be employed during development to obtain an overall improved system.

Our survey reveals that despite a large body of research on specifications, additional requirements, and their optimization, there is currently no clear recommendation, let

[2] https://dblp.org/

alone consensus, on how to use them. Many works do not even address the concern of over-optimizing specifications. With increased formalization of regulatory requirements, incentives rise to narrowly follow specifications in artificial intelligence. Therefore, it is paramount to develop analysis schemes, recommendations, and best practices for developing AI systems with multiple, potentially competing quality metrics and specifications.

The remainder of this survey is structured in the following way: Section 2 contextualizes how legal frameworks define and regulate specifications. Section 3 gives an overview of the types of tasks and data included in our study. The common criterion is that additional specifications have been defined for those tasks (in addition to “success” on held-out data). Section 4 categorizes the types of additional specifications that have been proposed for measurement. Section 5 outlines approaches for optimizing the different types of specifications. In Section 6, we describe the analysis of papers we reviewed both quantitatively on an aggregate level, and we also discuss a few selected papers to illustrate the variety of work covered in our survey. We summarize our findings in Section 7.

2 Specifications and the legal framework

Without proper ethical guidelines and safeguards, AI systems may exacerbate inequalities, further marginalize vulnerable communities, and cause physical or psychological harm. Responsible AI requires that the design and development of AI systems are aligned with universal values, principles, and international norms (Kiden et al, 2024). Alignment to ethical values gives rise to a double challenge: the normative challenge of deciding *which* principles should be considered, and the technical challenge of deciding *how* these principles should be encoded in AI systems (Iniesta, 2023). That is, putting AI ethics into practice requires AI regulation and the translation of principles into concrete standards and requirements (Bleher and Braun, 2023).

Legislators around the world seek to regulate AI technology as the production and deployment of such systems increase rapidly. Such regulation often entails broad terms that leave room for interpretation and need translation into a specification to help developers achieve compliance. The interaction between legal terms and their technical interpretation becomes essential for lawmakers to ensure that systems adhere to the desired principles, and often takes place through the development of recognized standards. While this phenomenon is inherent to technical legislation worldwide, we will highlight the European Artificial Intelligence Act (AI Act) (European Parliament and Council of the European Union, 2024) as an example because it is the first comprehensive regulation of AI.

The AI Act is the central European legislation project to establish harmonized requirements for Artificial Intelligence in the Union and attempt a conciliation of the technology with its fundamental values. It establishes several requirements that systems must observe to be authorized for the internal market. The AI Act establishes the category of high-risk AI systems in Article 6 and subjects mainly these systems to its requirements. In the final version of the AI Act, General Purpose AI systems are also subject to certain requirements before they can be deployed in the market. It builds—in the spirit of prior European product safety regulation (The European Comission, 2008)—on the established mechanism of European harmonized standards, making visible the interplay between regulatory approaches and technical specifications.

Enforcing requirements through standardization

The lawmaker operates—similar to other technical legislation—with rather broad terms that leave room for interpretation to ensure sufficient flexibility of the acts. In this manner, the AI Act establishes that providers of AI systems have to demonstrate the conformity of their systems to harmonized standards or that the systems have to be tested against “appropriate standards”. As a result, the legislative act alone is not enough for providers to obtain information on the procedure they have to observe. Rather – even though only marginally mentioned – the harmonized standards become central sources for establishing conformity.

Having reliable metrics for measuring certain qualities of AI systems that the legislator (and by extension, society) demands of them becomes essential because the logic of the law transforms these metrics into the determining factor for severe liability questions. Liability claims are facilitated by an assumption of causality in the case of non-compliance according to Article 4 (2) of the proposed AI Liability Directive (European Parliament and Council of the European Union, 2022), and regulatory fines can be administered according to the AI Act. On the other hand, for systems compliant with harmonized standards, establishing a liability according to the new product liability regime will become much more difficult, if not impossible, for claimants, because in order to establish the defectiveness of a product, interventions by regulatory authorities and product safety requirements (such as the ones in the AI Act) have to be taken into account. Furthermore, Article 10 of the proposed new product liability directive even establishes that liability is excluded if the defectiveness is due to the product’s compliance with mandatory regulations (such as the AI Act).

Harmonized standards and the AI Act

The Act treats favorably the harmonization via the mechanism of Regulation 1025/2012 (European Parliament and Council of the European Union, 2012), in which the Commission issues a standardization request to the standardization organizations. A reference to the harmonized standard is then published in the Official Journal of the European Union. Systems complying with such standards will be (rebuttably) assumed to also comply with the related requirements of the AI Act. Should this strategy fail, i.e., no harmonized standards be passed, the Act attributes a backup role to the Commission in Article 41. In this case, the Commission can become active and adopt “common specifications”.

The AI Act treats technical specifications as the central way of determining compliance with the requirements. In doing so, it remains vague and delegates the responsibility (and decision-making authority) entirely to executive bodies, namely the standardization organizations. The Act is governed by two implicit assumptions. It assumes that

1. for every requirement imposed on systems, there is a corresponding metric (or several metrics) to measure the fulfillment of the requirement accurately.
2. if a system fails to fulfill a specification (e.g., by falling below a pre-determined threshold in a metric), adapting the system to fulfill the specification will not harm its overall performance. That is because if the system fulfills a specification then it is considered compliant, and the Act does not reference the further impact an adaptation process might have on the system overall.

In this respect, the Act does not provide procedural guidance on improving a system if it does not satisfy a given metric—nor does it require harmonization bodies to provide procedural guidelines.

Standardization organizations and decision-making power

The delegation of decision-making authority by the Commission is recognized by the Court of Justice of the European Union (CJEU) and is not an uncommon tool to support the application of Union legislation. When the Commission asks standardization organizations to draft harmonized standards (that will lead to a presumption of conformity under the AI Act), it uses a delegated decision-making authority. The standardization organizations work together with the European Union in a public-private partnership. A common understanding between standards organizations, the European Commission, and the European Free Trade Association (EFTA) of the principles of this collaboration has been outlined since 1984 and updated in 2003 (The European Comission, 2003).

In the case of technical metrics for AI systems, this may be problematic because the consequences of the decision (e.g., which standard to adopt) can be more unpredictable than in other technical sectors. AI systems may narrowly follow the specifications to conform to the standards to the cost of the overall performance of the system, the underlying requirements, or other overlooked relevant aspects. Given the considerable potential harms of specification overfitting, it seems worth asking whether this process of delegation to standardization authorities in the context of the AI Act is enough to provide legal protection and means of redress for citizens affected by AI systems and enough legal clarity for developers and providers.

3 Application areas

This section summarizes the application areas explored by the papers in our survey. We group them into categories based on the nature of the data, as different varieties require different specifications. We give an overview of each category's current state of the art, along with challenges and limitations.

3.1 Natural language processing

Natural language processing (NLP) applications use text or speech as input. They comprise tasks such as sentiment analysis (Socher et al, 2013), machine translation (Bahdanau et al, 2015), and named entity recognition (Lample et al, 2016). The state of the art of the area is dominated by transformer-based (Vaswani et al, 2017) systems trained on massive amounts of text to optimize a language modeling objective (Clark et al, 2020; Raffel et al, 2020; Liu et al, 2019b; Devlin et al, 2019). Fine-tuning on instruction datasets (Wei et al, 2022a), optimizing additional objectives (Ouyang et al, 2022), and scaling up training data and system size allowed the use of language models on complex tasks requiring multi-step reasoning and diverse knowledge (Wei et al, 2022b). Current research investigates the use of language models in areas such as creative writing (Yuan et al, 2022), code development (Zan et al, 2023), education (Kasneci et al, 2023), and medicine (Thirunavukarasu et al, 2023). Limitations of these systems include the use of non-generalizable heuristics (Tu et al, 2020), generating texts that are biased, hateful, and toxic (Schick et al, 2021), and texts that look fluent and plausible but contain falsehoods and misinformation (Ji et al, 2023; Lin et al, 2022).

3.2 Computer vision

Computer vision (CV) applications process images or videos as input and solve tasks such as image classification (Russakovsky et al, 2015), segmentation (Minaee et al, 2022), and face recognition (He et al, 2005). The state of the art is dominated by vision transformers (Dosovitskiy et al, 2021) and convolutional neural networks (LeCun et al, 1989; Fukushima, 1980) pretrained on massive image datasets (He et al, 2016; Szegedy et al, 2016; Krizhevsky et al, 2012). Vision transformers brought improvements in a wide range of vision tasks (Han et al, 2023) and gave rise to multimodal systems capable of combining—and producing—visual and text information (Radford et al, 2021; Ramesh et al, 2021). Computer vision has been applied in sensitive areas like healthcare (Esteva et al, 2021), surveillance (Sreenu and Saleem Durai, 2019), and autonomous vehicles (Hu et al, 2023). Despite the good performance on standard benchmarks, there are still technical and ethical limitations such as the lack of robustness to distribution shifts (Ben-David et al, 2010) and adversarial attacks (Goodfellow et al, 2015), and poor performance on underrepresented demographic groups (Buolamwini and Gebru, 2018).

3.3 Others

While most of the papers in the survey explored NLP and CV tasks, some investigated tasks that fit other categories.

Tabular data: Tabular data applications represent input examples as structured records of numerical and categorical features. Contrary to previous cases, traditional machine learning algorithms such as ensembles of decision trees often still outperform deep learning-based approaches (Borisov et al, 2022). Systems trained on tabular data are applied to a wide range of areas, including sensitive ones such as medical diagnoses (Kononenko, 2001) and financial analyses (Bhatore et al, 2020), even though they have been shown to reproduce dataset biases (Angwin et al, 2016).

Graphs: In graph applications, entities and their relationships are represented as nodes and edges in a graph. Tasks include assigning graphs or nodes to particular classes or predicting links between entities. State-of-the-art approaches use different variants of graph neural networks (GNNs) (Wu et al, 2021), which have been applied to areas such as social network analysis (Fan et al, 2019) and drug discovery (Xiong et al, 2020). Examples of current challenges in graph applications are generalization and scalability concerns (Bronstein et al, 2017) and system vulnerability to adversarial attacks (Sun et al, 2023).

Reinforcement learning: Complex tasks that cannot easily be learned by optimizing local decisions are often modeled in the framework of reinforcement learning. Here, the problem formulation is that an agent seeks to maximize a reward signal by choosing the optimal action given an environment state (Sutton and Barto, 2018). The current state-of-the-art methods are based on deep reinforcement learning (Arulkumaran et al, 2017) and have prominently been applied to video games (Mnih et al, 2015) and robotics (Levine et al, 2016). The formulation of the rewards signal is critical, as reinforcement learning systems are vulnerable to reward hacking, where the agent optimizes the reward to the detriment of the task (Skalse et al, 2022). There are also concerns with robustness to noise and adversarial attacks (Lütjens et al, 2020).

4 Specifications

The requirements engineering framework distinguishes requirements from specifications (Jackson, 1995). Requirements are concerned with world phenomena, while specifications lie in the intersection of machine and world phenomena. Requirements can include high-level concepts such as fairness and robustness, which are translated into specifications by defining datasets or metrics intended to assess those properties.

The path between requirements and specifications is perilous: going from the requirement to the specification level requires abstracting away world-only phenomena. The requirements are constructs: unobservable theoretical abstractions that describe phenomena of interest, such as robustness and fairness (Jacobs and Wallach, 2021). These cannot be measured directly, as they are not observable. Instead, constructs are *specified* through a measurement model that leverages observable properties, or proxies (e.g., accuracy on a dataset, invariance tests, bias metrics), to infer the construct. That involves making assumptions about the relevant observable properties and how they relate to the unobservable construct and each other, potentially introducing mismatches between the theoretical understanding of the problem and its operationalization (Jacobs and Wallach, 2021).

In the rest of this section, we describe aspects of interest to our survey, and how we categorized the surveyed papers w.r.t. different types of sepcifications and other properties.

4.1 What to specify

The papers from our survey measure specifications that we categorize into three groups.

Robustness

Robustness concerns how well a system works on examples whose distribution differs from the training distribution.

Often, a specific desideratum for robustness is that a small change in the input should lead to no (or only a small) change in the output. To test these properties, one can either rely on naturally occurring distribution shifts between data sets or create test examples by perturbing the input of examples and requiring stability on the output side (Wang et al, 2022b). The first case is a common issue when systems are used in the wild: NLP systems may have to process texts from different genres, dialects, and grammaticality; CV systems may have to process images with different lighting conditions, perspectives, and quality. The second case are perturbations, changes to the input part of examples designed to systematically test the effect on the predicted output. Perturbations are often used in the context of adversarial attacks (Zhang et al, 2020), where the aim is to fool the system into changing its prediction with minimal, unperceivable changes to the input. The more robust a system is, the less such environmental or adversarial changes degrade its performance.

Robustness is addressed directly in the AI Act (Art 15) as one of the central requirements for high-risk AI systems, along with accuracy. Article 15 para 1a also explicitly obliges the Commission to encourage the development of industry benchmarks and measurements to determine accuracy and robustness. These may differ from the harmonized standards in Article 40, but Article 15 para 1a, with its wording, encourages a system of de-facto industry standards to exist equally besides the harmonized standards.

Also, before the AI Act, robustness had an extensive tradition as a key aim in developing AI systems. In the Trustworthy AI Guidelines (High-Level Expert Group on AI, 2019), systems are required to be "lawful, ethical, and robust" to be considered trustworthy. This document was already produced in 2019, and followed by a large number of policy initiatives undertaken by the European Commission (The European Comission, 2018). Other actors, such as the OECD, have also picked up the notion of robustness, making it prominent in AI policy also outside and before the AI Act (OECD, 2019). Yet, on a policy level, there is no unanimous agreement on its definition, how it can be measured, or the threshold for a system to be considered robust.

Fairness

Machine learning systems can reflect societal biases in their training data, such as gender and racial stereotypes.

It has been well-documented how deploying such systems has harmed and further marginalized vulnerable communities (Mehrabi et al, 2021). Such harms can be mitigated by enforcing fairness constraints in the system predictions. There are multiple competing notions of fairness (e.g., individual fairness, equal opportunity, demographic parity, counterfactual fairness), leading to several fairness metrics (Barocas et al, 2019) and fairness enhancing methods (Pessach and Shmueli, 2023).

The term "fairness" is not used in the AI Act to refer to a distinct quality of systems that the Act requires. Nonetheless, the notion of fairness as a key requirement of AI systems is deeply ingrained in policy and commonly used ethical guidelines, which in turn often draw on fundamental rights discourses. For instance, the European Commission's High-Level Expert Group on Trustworthy AI lists four ethical principles for AI systems, derived from fundamental rights. One of these principles is fairness, closely linked to the rights to Non-discrimination, Solidarity and Justice (Art 21 and following in the EU-Charter) (High-Level Expert Group on AI, 2019). Fairness as a requirement for AI also appears to have established itself in academia and among practitioners, more than, for instance, the related concepts of equity or justice. That can be seen in a wide range of organizations aiming to develop AI fairness checklists (see, for example, Madaio et al (2020)).

While the AI Act does not pick up the specific term, Article 10 (2) (fa) requires providers to establish "appropriate measures to detect, prevent and mitigate possible biases" in the training, validation, and testing data set. Providers must consider metrics and techniques that test for and mitigate biases.

This obligation to detect and mitigate biases only explicitly applies to the data sets used. However, it could be construed as to also include mitigation techniques for the system rather than (only) the data sets, so as to further combat possible biased output.

Capabilities

We define a capability as a fine-grained aspect of desired task behavior. That includes diverse phenomena such as linguistic (Ribeiro et al, 2020), numerical reasoning (Naik et al, 2018) and generalization (Lake and Baroni, 2018) capabilities. Capabilities are often evaluated using test suites (Röttger et al, 2021; Ribeiro et al, 2020) comprising specific examples that relate to the tested capabilities.

While the AI Act obliges providers of AI systems to list the capabilities of the respective system as a means of transparency, it refers to capabilities as a technical specification primarily in the context of general-purpose AI. Such AI systems are considered to represent systemic risks if they have high capabilities (see Recital 60n). As a first approximation, the Act uses the amount of compute used for training and sets the initial threshold at 10^{25} FLOPs. Systems above this threshold are presumed to represent systemic risk. The European legislator furthermore explains in Recital 60n of the AI Act that this threshold of 10^{25} FLOPs should be adjusted over time as well as *be supplemented with benchmarks and indicators for system capabilities*, i.e., means of specification other than compute power. The Commission is granted an explicit mandate to amend the threshold for compute and adopt such benchmarks and indicators as specifications that, when met, trigger the presumption of systemic risk.

Therefore, from a regulatory perspective, capabilities' specifications play a central role in assessing general-purpose AI systems and categorizing those systems according to their possible risks.

4.2 How to specify

We distinguish between two specification categories according to how the measured property is encoded: example-based specifications, with the property encoded by a set of examples, and metric-based specifications, with the property encoded by a dedicated metric.

Example-based specifications

In this scenario, the requirements are validated through input-output examples that correspond in some way to the tested property. For example, to measure the robustness of a computer vision system, one can compute the accuracy on a set of perturbed samples (Ross and Doshi-Velez, 2018). Our survey categorizes example-based specifications into five types:

Human-generated: examples are written (or otherwise composed) by humans. E.g., the examples in the Crowdsourced Stereotype Pairs (CrowS-Pairs) (Nangia et al, 2020) dataset were created by asking crowdsourced workers to write sentences reflecting (or violating) stereotypes about demographic groups.

Pattern-generated: examples are generated algorithmically through template filling or rules. E.g., the examples in the INequality Theorem (INT) (Wu et al, 2020) benchmark are generated by a rule-based algorithm that automatically generates theorems.

Model-generated: examples are sampled from or generated by a probabilistic model. E.g., Bartolo et al (2021) used a pre-trained language model to generate synthetic question-answer pairs that improved the robustness of question-answering systems trained on them.

Perturbation-based: examples are generated by perturbing samples from a dataset. E.g., perturbing images from a dataset by adding to each example a vector that changes the system prediction while keeping perturbed and original images indistinguishable by humans (Goodfellow et al, 2015).

Selection-based: examples are selected from existing datasets to focus on the tested property. E.g., splitting graph datasets into a training set with the smallest graphs and a testing set with the biggest graphs to assess size generalization (Buffelli et al, 2022).

Metric-based specifications

In contrast to example-based specifications, metric-based specifications correspond to formalized scores for measuring properties of a prediction algorithm *without* needing additional samples or annotations—the tested property is encoded directly in the metric computation. An example of metric-based robustness specification would be measuring the expected perturbation magnitude needed to fool the system (Jakubovitz and Giryes, 2018).

There are many metric-based fairness specifications, which are typically computed by comparing statistics of system predictions conditioned on different demographic groups. Prototypical examples are equality of opportunity (Hardt et al, 2016), demographic parity (Cotter et al, 2019), and group calibration (Pfohl et al, 2022a), each comparing different group statistics: true positive rates, positive prediction rates, and calibration, respectively. It has been shown that no method can satisfy these fairness conditions simultaneously (Kleinberg et al, 2017). Far from being just a mathematical artifact, the incompatibility of fairness metrics points to the differences in the underlying notions of fairness (Barocas et al, 2019) and value systems (Friedler et al, 2021).

4.3 Measuring and improving

Our survey covers papers that *evaluate* specifications or use methods to *improve* systems regarding specifications.

Evaluation

We say a paper in our survey *evaluates* a specification if it measures it. I.e., the paper either proposes a new method of how to evaluate a specification (e.g., by designing a test suite (Kirk et al, 2022) or a metric (Weng et al, 2018)) or studies a previously proposed specification as part of the evaluation (in the simplest case just reports its outcome).

Improvement

Attempts to improve the specification performance (Sec. 5) range from designing a fairness optimization method (Deng et al, 2023) or to employing regularization methods for increasing robustness (Ross and Doshi-Velez, 2018)) in the works included in our survey.

Establishing harmonized standards for optimization methods is not a legal requirement in the AI Act. The Commission may, when issuing a request for standardization to a European standardization organization, ask for standardization of such optimization methods. However, it will not transfer into a legal requirement since the legal text of the AI Act does not provide for an obligation to use certain optimization methods to achieve compliance. From a legal perspective, therefore, achieving harmonized metrics will be a requirement for certain systems; it remains, however, up to system developers how to achieve this.

5 Specification optimization

Specifications indicate the (mis-)alignment with or degree of fulfillment of specific properties or proxies of capabilities. While some works on specifications for AI system (Ribeiro et al, 2020; Nangia et al, 2020) claim that the additional metrics measuring the fulfillment of specifications should only be used as an insight into existing system behavior, it is also possible to use this feedback for optimizing the measured system properties (Liu et al, 2019a; Bartolo et al, 2021).

Therefore, the goal in research on specifications ranges from the view that no optimization of specification metrics should be attempted to the view that these metrics can be used for system development. Under the latter view, specification metrics can be helpful to compare and select different settings w.r.t. performance on this specification, use the specification as part of a loss function during optimization, or even to specifically design algorithms for improving performance on a specification.

5.1 Specification optimization strategies

We categorize strategies for specification optimization into the following groups of approaches, depending on how direct the influence of the specifications is on the resulting system:

No optimization. Different settings (system types, hyper-parameter choices), each of which is not specifically directed to improve the desired property, are compared w.r.t. performance on the specification metric. This can give guidance as to which of the settings performs better w.r.t. the measured specification (and which system could be chosen if specification performance was prioritized)—but neither the systems themselves nor the training process are targeted at optimizing the metric. For example, the *CheckList* approach (Ribeiro et al, 2020) provides a detailed analysis of failure categories for sentiment analysis, duplicate question detection, and machine comprehension but does not suggest using the outcome of this analysis for system improvement.

Direct optimization. The specification metric that measures the property of interest is a direct target in optimizing the AI system. This could be a term in the loss function corresponding to the measured quantity or the inclusion of training examples that, by construction, directly reflect the evaluation logic or come from the same pool of examples used to measure the desired property. For example, the *FIFA* approach (Deng et al, 2023) uses a combined fairness and accuracy loss during optimization. *Inoculation by fine-tuning* uses examples from *challenge sets*, specifically constructed data sets for testing phenomena in natural language inference and question answering, as additional training data.

Direct adjustments of the behavior of an AI system, such as extending prompts with in-context examples that correspond to the evaluation setting (Levy et al, 2023), also fall into this category. If improvement strategies are directly inspired by a specific way a property is measured (rather than the property in an abstract sense or an alternative way of specifying the property), they also count as a direct attempt to improve, even if assumptions and approximations are made.

Indirect optimization. As before, an additional property (apart from performance on the main task) is a target in optimization. However, the optimized property is not the specification but a property that is assumed to be related. For example, regularization strategies could be employed for improving the robustness of the system, even if the exact regularization term does not directly follow from the mathematical formulation of the robustness metric (Jakubovitz and Giryes, 2018; Ross and Doshi-Velez, 2018). In other words, the improvement strategy relates to the desired underlying property but not directly to the metric used to measure it.

5.2 Specification optimization evaluation

Specification optimization may impact system performance not only considering the optimized property but also the main task performance and other specifications. An ideal specification optimization strategy would improve both task performance (as measured by an assumed to be i.i.d. test set) and better align the system to the high-level principle encoded by the specification. However, there are possible unintended consequences of specification optimization. Mismatches between specifications, their underlying goals, and task performance may lead to the deterioration of system performance in unforeseen ways.

Evaluation metrics

When evaluating specification optimization, it is critical that the evaluation scheme is constructed to reveal such system degradation—depending on which metrics are considered for evaluation, some of the failure cases may be obfuscated. Metrics either measure performance on the main task or on additional specifications.

Task metric. The task metric is the main measure of system performance, often a correctness metric (e.g., accuracy, f-score) computed on a held-out test set. Reporting the task metric can reveal whether a specification optimization strategy degraded general system performance. For example, papers that propose methods to improve robustness to adversarial attacks may report the accuracy for the unperturbed test set to verify that the method preserves system performance on clean samples (Rebuffi et al, 2021; Hendrycks et al, 2019a; Xie et al, 2019).

Specification metric. The evaluation scheme may include a range of specification metrics. This can happen by reporting alternative formulations of specifications with the same underlying goal (e.g., measuring several fairness metrics (Cotter et al, 2019)) or specifications that capture different requirements (e.g., assessing robustness to distribution shifts and system calibration (Hendrycks et al, 2019b))

Specification overfitting analysis

The term *overfitting* describes the case in which an AI system learns features that arise from noise and data variance rather than learning the underlying data distribution (Webb, 2010). Traditionally, a model is said to have overfitted when it has low train error but high test error (Aceña et al, 2022), though *overfitting* is also used to denote other types of over-optimization that can lead to unwanted drops in performance, such as those due to distribution shift and test set reuse (Roelofs et al, 2019).

Specification overfitting occurs when a specification optimization strategy improves system performance w.r.t the optimized metric but degrades system performance w.r.t. the task metric or other specification metrics. We categorize evaluation schemes based on the metrics they include and their ability to detect specification overfitting.

No overfitting analysis. If the evaluation scheme includes only one specification and/or task metric, we consider that there is no specification overfitting analysis. Reporting only one specification metric does not account for possible effects on other specification metrics—it has been shown that optimizing a set of specification metrics can have catastrophic consequences on other specifications (Luz de Araujo and Roth, 2023). While reporting the task metric accounts for the overall impact on task performance, it may obfuscate unintended consequences. For example, the task metric may not significantly change if system behavior improves a little for common cases but degrades a lot for rare ones (Liu et al, 2021).

Cross-specification analysis. This comprises evaluation schemes that report at least two specification metrics. Examples of this include reporting the performance for alternative formulations of a specification (e.g., different attack types for adversarial robustness (Li et al, 2023a; Dapello et al, 2022; Cheng et al, 2022)) or evaluating specifications for different requirements (e.g., capability of handling negations and robustness to word overlap in natural language inference (Naik et al, 2018)). The former guards against narrowly adapting to the specification to the detriment of the underlying requirement. The latter accounts for possible negative interactions between different requirements.

Task performance analysis. This describes evaluation schemes that go beyond reporting a single task metric and examine the effect on task performance more deeply. This can involve comparing performance on relevant subgroups of the task data (e.g., reporting the worst group accuracy in addition to the dataset average performance (Zhang et al, 2022c; Liu et al, 2021)), or evaluating task performance on additional (assumed to be i.i.d.) test sets from the same task (e.g., Chen et al (2022)). These measures can provide a more reliable assessment of the impact of specification optimization on task performance.

Comprehensive overfitting analysis. This category covers evaluation schemes that combine cross-specification and task performance analysis (e.g., (Pfohl et al, 2022a)). By considering multiple specification metrics and deeply examining task performance, such evaluation schemes may identify failure cases of specification optimization and prevent specification overfitting.

6 A survey of specification overfitting

This section presents our survey of specification overfitting. We sample and analyze papers that propose methods to improve or measure specifications. The goal is to create an overview of how the research community has dealt with the specification overfitting issue in recent years.

6.1 Method

Paper collection. By keyword search, we collect papers from the DBLP[3] database. We restrict our search to major conferences and journals on natural language processing, computer vision, and machine learning.[4] We used the following keywords:

- test suite
- behavioral|behavioural|functional|stress + test (4 searches).
- challenge + set|dataset (2 searches).
- diagnos|evaluat|benchmark|test|assess|improv|increas|train|optimi + {property} (45 searches).

Where {property} corresponds to fair|robust|generalis|generaliz|capabilit and refers to specifications for fairness, robustness, generalization, and specific capabilities. Our first collection round happened on December 12, 2022, returning 950 papers. We did a second round on August 25, 2023, to improve recall for papers from 2022 and add papers from 2023. That returned 222 more papers.

Filtering. First, we restrict the papers to those published in 2018 at the earliest, yielding 1172 papers. We then examined all abstracts to assess if they fit our inclusion criteria—papers that propose a method to improve or evaluate a specification. We judged 442 papers as relevant. We assigned each of them to at least one application area: NLP, CV, or others.

Fig. 1 shows the number of papers by year and application area. Interest in measuring and improving specifications seems to be on an upward trend, considering we only sampled papers up to July 2023. Computer vision is present in about half of

[3] https://dblp.org/

[4] AAAI, ACL, COLING, Computational Linguistics, CoNLL, CVPR, EACL, ECCV, EMNLP, FAccT, ICCV, ICLR, ICML, IJCAI, NAACL, NeurIPS, and TACL.

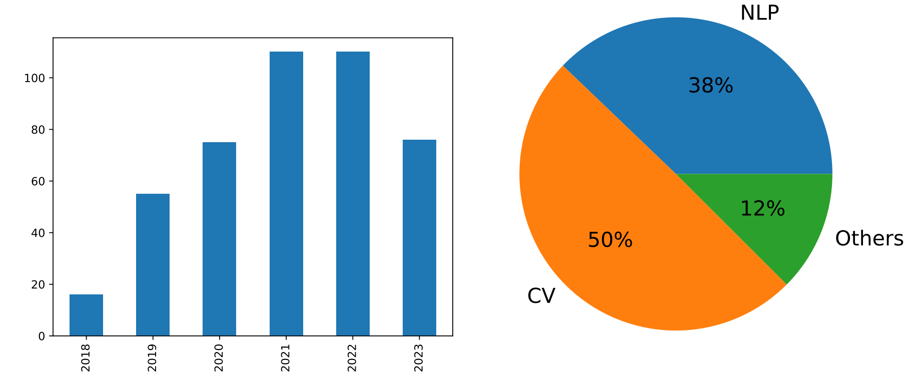

Fig. 1 Number of relevant papers by year (left) and application area (right).

papers, followed by natural language processing. Only 12% of the papers explore other application areas (e.g., graph and tabular data).

As a last step, we select the five most cited papers[5] from each application area for each year. That was done to keep the workload of analyzing papers manageable while keeping impactful papers and maintaining coverage of different years and application areas. Due to some papers covering more than one application area and years with fewer than five samples for a given application area, we ended up with 79 papers for annotation.

Analysis. We read the filtered papers to collect information for the fields in Table 1. In this step, we found that five papers did not meet the inclusion criteria, resulting in a final pool of 74 papers. Table 1 maps the analysis criteria discussed in the different sections of our article to the categories used in the structured analysis, and Table 2 presents the results of the structured analysis.

6.2 Quantitative results

Evaluation and improvements. Due to the inclusion criteria, all of the papers evaluate a specification. Sixty-two papers explore specification optimization strategies. Of these, 26 include only direct methods, 27 only indirect methods, and nine combine direct and indirect means of improvement.

Specification. Robustness was the most common specification, with 44 papers, followed by fairness (19) and other specific capabilities (15). Each specification was conceptualized in many distinct ways. Robustness was understood as measures of system performance under adversarial attacks (Guo et al, 2018), distribution shifts (Hendrycks and Dietterich, 2018), inference heuristics (Min et al, 2020), different subpopulations (Liu et al, 2021), missing modalities (Ma et al, 2022), and question paraphrasing (Gan and Ng, 2019). Fairness measures were very diverse, including, for example, equalized odds (Wang et al, 2020b), demographic parity (Coston et al, 2020),

[5] As reported in scholar.google.com.

Table 1 Paper analysis fields and descriptions.

Field	Description
Application **Area**, see Section 3	The field which the AI/ML application falls in: Natural language processing (**NLP**), computer vision (**CV**), tabular data (**TAB**), graphs (**GRAPH**), r reinforcement learning (**RL**).
Specification (**Spec.**), Sec. 4.1	The specification the paper intends to measure or improve: robustness (**R**), fairness (**F**), or capabilities (**C**).
Evaluation (**Eval.**), Sec. 4	Whether the paper measures a specification (✓).
Example or metric-based (**Ex/M**), Sec. 4.2	Whether the specification is measured using additional examples (**e**) or on the same examples as the main task but using a different metric (**m**).
Type of example, Sec. 4.2	If the evaluation is example-based, how the examples are created. We categorize examples into: handcrafted by humans (**h**), pattern-generated (**pat**), sampled from a probabilistic model (**prob**), obtained by perturbing dataset examples (**per**), or obtained by selecting dataset examples (**s**).
Improvement (**Imp.**), Sec. 5	Whether the paper experiments with improving a specification metric (✓).
Improvement strategy (**Imp. Str.**), Sec. 5.1	Whether the improvement strategy is based on directly (**d**) optimizing the specification metric (or a proxy) or indirectly (**i**) through other means (e.g., regularization).
Reports task (i.i.d.) metric (**Task M.**), Sec. 5.2	Whether the paper reports a correctness metric for a standard dataset (✓).
Overfitting analysis (**Ov. An.**), Sec. 5.2	Whether the paper reports other additional (i.e., more than one) specification metrics (**o**) and/or studies the effect on task performance (**t**) in detail.
Scope/limitations (**S/L**), Sec. 6.3	Whether the paper explicitly discusses the method's scope, e.g., intended use, limitations, assumptions (✓).
Recommendation category (**Rec.**), Sec. 6.4	If the paper offers a recommendation on how to integrate the specification metric or the improvement method to the system development process, we categorize it into vague (**V**), delegating (**Del**), (not) additional data (**(¬)D**), debugging (**Deb**) and concrete (**C**).

equal opportunity (Cotter et al, 2019), individual fairness (Black et al, 2020), and calibration by group (Petersen et al, 2023). Capabilities included generalization (Wu et al, 2020), calibration (Hendrycks et al, 2019b), handling of linguistic phenomena (Naik et al, 2018), level of bias (Nangia et al, 2020), reasoning (Liu et al, 2019a), and task-specific capabilities, e.g., recognizing emoji-based hate (Kirk et al, 2022).

Example vs. metric-based specifications. Example-based specifications were the most common, with 53 papers. The majority were perturbation-based (32), followed by selection of specific dataset examples (12), pattern-generated (9), human-generated (6), and model-generated (5).

Overfitting analysis. Most papers (61) report main task correctness and 26 papers do not include a specification overfitting analysis. Of the papers that include specification overfitting results, 42 report cross-specification results, four present task performance analyses, and only two are categorized as presenting a comprehensive overfitting analysis. If we consider only the papers with improvement methods (62), six do not report main task correctness, and 16 do not present any specification overfitting analysis.

Table 2 Structured analysis of survey papers.

Paper	Area	Spec.	Eval.	Ex/M	Type	Imp.	Imp. str.	Task M.	Ov. An.	S/L	Rec.
(Bartolo et al, 2021)	NLP	R	✓	e	prob	✓	i	✓	o		D
(Black et al, 2020)	TAB	F	✓	e	prob					✓	V+Del
(Buffelli et al, 2022)	GRAPH	C	✓	e	s	✓				✓	
(Chen et al, 2019)	TAB	F	✓	m						✓	
(Chen et al, 2022)	NLP/CV	C	✓	e	s	✓	i	✓	t		
(Cheng et al, 2019)	NLP	R	✓	e	prob	✓	d	✓	o		
(Cheng et al, 2020)	NLP	R	✓	e	prob						
(Cheng et al, 2022)	CV	R	✓	e	per	✓	i	✓	o		
(Clarysse et al, 2022)	CV	R	✓	e	per	✓	i+d		o		
(Coston et al, 2020)	TAB	F	✓	m				✓			
(Cotter et al, 2019)	TAB	F	✓	m		✓	d	✓	o		
(Croce et al, 2021)	CV	R	✓	e	per		d	✓		✓	V
(Dapello et al, 2022)	CV	R	✓	e	per	✓	i	✓	o	✓	
(Deng et al, 2023)	TAB	F	✓	m		✓	d+i	✓		✓	
(Elkahky et al, 2018)	NLP	C	✓	e	s	✓	d+i	✓	o		
(Fatemi et al, 2023)	NLP	C	✓	e	h+s	✓	i	✓		✓	
(Gan and Ng, 2019)	NLP	R	✓	e	h+prob	✓	d	✓			
(Geirhos et al, 2018)	CV	R	✓	e	per	✓	i	✓	o		
(Gowal et al, 2021)	CV	R	✓	e	per	✓	i	✓	o	✓	
(Guo et al, 2022)	RL	R	✓	e	per			✓			
(Guo et al, 2018)	CV	R	✓	m		✓	i	✓			
(Havasi et al, 2020)	CV	R	✓	e	per	✓	i	✓	o		
(Hendrycks et al, 2019b)	CV	C+R	✓	e	per	✓	i	✓	o		
(Hendrycks and Dietterich, 2018)	CV	R	✓	e	per	✓	d+i	✓	o		
(Hendrycks et al, 2019a)	CV	R	✓	e	per	✓	i	✓	o		
(Jakubovitz and Giryes, 2018)	CV	R	✓	m		✓	i	✓			
(Jung et al, 2022)	TAB	F	✓	m		✓	d	✓			
(Karpukhin et al, 2019)	NLP	R	✓	e	per	✓	d	✓	o	✓	
(Kirichenko et al, 2022)	NLP/CV	R	✓	e	s+per	✓	i	✓	o		
(Kirk et al, 2022)	NLP	C	✓	e	pat	✓	d	✓	o	✓	D
(Komiyama et al, 2018)	TAB	F	✓	m		✓	d	✓			
(Lee et al, 2022)	TAB	F	✓	m		✓	d	✓	o		
(Levy et al, 2023)	NLP	C	✓	e	s	✓	d+i	✓	o+t	✓	
(Li et al, 2023a)	CV	R	✓	e	per	✓	i	✓	o		
(Li et al, 2023b)	TAB	F	✓	m		✓	d	✓	o		
(Liang et al, 2022)	RL	R	✓	e	per	✓	d	✓	o	✓	
(Liu et al, 2019a)	NLP	C	✓	e	pat+h	✓	d	✓		✓	
(Liu et al, 2021)	NLP/CV	R	✓	e	s	✓	i	✓	t		
(Ma et al, 2022)	NLP/CV	R	✓	e	s	✓	i	✓	o	✓	
(Madras et al, 2018)	TAB	F	✓	m		✓	d	✓	t		
(Min et al, 2020)	NLP	R	✓	e	pat+per	✓	d	✓	o		
(Mishler et al, 2021)	TAB	F	✓	m		✓	d	✓		✓	
(Naik et al, 2018)	NLP	C	✓	e	pat+per	✓	d	✓	o		V+Del
(Nangia et al, 2020)	NLP	C	✓	e	h					✓	Deb+¬D
(Narasimhan et al, 2019)	NLP/TAB	F	✓	m		✓	d	✓			
(Petersen et al, 2023)	TAB	F	✓	m						✓	V+Del
(Pfohl et al, 2022a)	TAB	F	✓	m		✓	d	✓	o+t	✓	C
(Qiu et al, 2022)	NLP	C	✓	e	s	✓	d	✓	o	✓	D
(Rahmattalabi et al, 2021)	TAB	F	✓	m		✓	d	✓		✓	
(Rebuffi et al, 2021)	CV	R	✓	e	per	✓	i	✓	o		
(Ribeiro et al, 2020)	NLP	C+F+R	✓	e	pat+per			✓	o	✓	Deb
(Roh et al, 2021)	TAB	F+R	✓	m		✓	d	✓		✓	
(Ross and Doshi-Velez, 2018)	CV	R	✓	e	per	✓	i	✓	o		
(Röttger et al, 2021)	NLP	C	✓	e	h+pat			✓	o	✓	Deb
(Ruis et al, 2020)	NLP	C	✓	e	pat	✓	d		o		
(Schneider et al, 2020)	CV	R	✓	e	per	✓	d+i	✓	o	✓	V
(Sehwag et al, 2022)	CV	R	✓	e	per	✓	i		o		
(Sinha et al, 2019)	CV	R	✓	m		✓	d	✓	o		
(Sun et al, 2020)	CV	R	✓	e	per+s	✓	i	✓	o		V
(Taskesen et al, 2021)	TAB	F	✓	m		✓	i	✓		✓	V+Del
(Tjeng et al, 2019)	CV	R	✓	e	per						
(Wang et al, 2019)	CV	R	✓	e	per	✓	i	✓	o		
(Wang et al, 2020a)	NLP	R	✓	e	h+pat+per	✓	d+i	✓	o		
(Wang et al, 2020b)	TAB	F	✓	m		✓	d	✓			
(Wang et al, 2022a)	NLP	R	✓	e	s	✓	d		o	✓	
(Wang and Bansal, 2018)	NLP	R	✓	e	per	✓	d+i	✓	o		
(Weng et al, 2018)	CV	R	✓	m							
(Wu et al, 2020)	GRAPH	C	✓	e	pat	✓	i		o	✓	V+Del
(Xie et al, 2019)	CV	R	✓	e	per	✓	i	✓	o	✓	
(Zhang et al, 2019)	CV	R	✓	e	per	✓	i	✓	o		
(Zhang et al, 2022a)	GRAPH	R	✓	e	per	✓	d+i	✓			
(Zhang et al, 2022c)	NLP/CV	R	✓	e	s	✓	i	✓	t		
(Zhang et al, 2022b)	CV	R	✓	e	per	✓	i	✓	o	✓	
(Zhuo et al, 2023)	NLP	R	✓	e	per	✓	d	✓	o	✓	

Scope and recommendations. Only 30 of the papers explicitly discuss the scope or limitations of the proposed specification or optimization strategy (we do not consider the mentioning of limitations w.r.t. other aspects). Further, only 15 papers discuss how the specification should or should not be used during system development or contextualize the role of the specification given the main task and other specifications.

Discussion. The analyzed papers often implicitly guard against some specification overfitting pitfalls by reporting either the main task performance or metrics for other specifications. Evaluating main task performance singles out methods that improve a specification to the detriment of general correctness (e.g., always predicting the same outcome is robust against adversarial attacks but has poor task performance). Measuring performance on other specifications checks whether a method has improved a particular specification to the detriment of others (e.g., improving how a hate speech detector improves for a given demographic while decreasing performance for another).

However, most papers do not explicitly discuss the scope of the proposed method and even fewer contextualize its role in the system development process. Describing the scope is a way to prevent more insidious pitfalls, such as taking good specification performance as a guarantee of system quality (e.g., claiming that a system with a good performance on a specific fairness metric is *fair*) or not using the method as it was originally intended (e.g., fine-tuning a language model on a dataset intended for evaluation of bias only).

6.3 Scope and limitations analysis

We identified whether papers explicitly stated the scope and limitations of the proposed specification or improvement method. We consider that a paper explicitly discusses the scope and limitations of the proposed specification measure or optimization strategy if it describes the cases for which the method applies or for which ones it does not. This can be done, for example, by discussing the assumptions underlying the proposed method, by contrasting it with alternative formulations, or by discussing in which context the method should be used. If the paper does not include such discussions, we consider that the scope and limitations were not made explicit.

Scope and limitations are important not only from a practical and scientific position but also from a legal one. Suppose a harmonized standard does not fully cover a legal requirement. In this case, compliance with the standard will not establish the (full) presumption of conformity. Considering this, systems might have to comply with several harmonized standards to obtain a presumption of conformity with one of the high-level legal requirements, such as, for example, robustness.

Examples of scope and limitations and counterexamples

In the following, we show types of explicitly defined scope as described in the included papers. We also present counterexamples that illustrate how the scope or limitations of a specification or improvement method are not sufficiently defined.

Scope of application context. Ribeiro et al (2016) state that their proposed test suites can only account for behavioral (input-output) issues but not non-behavioral issues such as noisy and biased training data, lack of interpretability or security issues. In contrast, some papers do not explicitly restrict the context for applying the proposed method. For example, papers examining adversarial robustness (Guo et al, 2018; Jakubovitz and Giryes, 2018; Ross and Doshi-Velez, 2018) often evaluated the robustness of specific attack types without discussing the generalizability to different attacks.

Discussion of alternative specifications. Fairness is a complex concept with diverse cultural, legal, societal, and ethical understandings. Given the multiple competing notions of fairness and many possibilities of fairness metrics, it is important that authors justify their choices or at least acknowledge these choices. For example, Roh et al (2021) state that their method is limited to a specific group of fairness measures (i.e., equalized odds and demographic parity disparity), and that one needs to choose a fairness measure in light of the underlying social context. In contrast, other papers (Coston et al, 2020; Komiyama et al, 2018; Madras et al, 2018) do not justify the choice of fairness metric or acknowledge alternative formulations.

Making assumptions explicit. Some papers restrict the scope of the specification by identifying the assumptions behind it and the consequences of breaking some of them. For example, Croce et al (2021) state that results on RobustBench, the proposed robustness benchmark, may not generalize well to real-world deployment if the data comes from a new domain or if novel adversarial attacks are used. In contrast, Gan and Ng (2019) train a system to generate paraphrases to test and improve the robustness of question-answering systems under the—implicit and not discussed—assumption that the system will generalize from system-generated paraphrases to real-world cases, which might not be the case.

6.4 Analysis of recommendations

We extracted recommendations regarding the proposed specification metric or optimization strategy from the analyzed papers. We consider recommendations to be passages offering guidelines on integrating the specification into the system development process or how to interpret the metric alongside the task metric and other specifications when considering practical implications. Recommendations can prevent misuse of the proposed specification or optimization strategy, such as applying a technique in the wrong context or falsely taking good performance on a specification to guarantee general system quality.

We categorize the extracted recommendations into the following types:

Vague. Vague recommendations provide high-level suggestions but do not define concrete measures that should be taken to enforce them. Some works that propose specifications mention how they should *supplement* standard evaluation but not *substitute* it (Wu et al, 2020; Naik et al, 2018). While it is valuable to restrict the scope of the metric in that way, such guidelines are not directly actionable as they leave out the matter of *how* the specification metric can supplement standard evaluation.

Delegating. Delegating recommendations also provide abstract guidelines, but they defer the definition and execution of the guidelines to other actors. An example is deferring results interpretations to domain experts (Black et al, 2020). Deferring decisions to the actors in the best position to make them is surely a good idea, but such recommendations often do not explicitly describe which factors the experts to which interpretation is delegated should consider when dealing with the specification.

Debugging. Some papers recommend that specifications be used for debugging, i.e., finding and fixing errors. For example, Ribeiro et al (2020) proposes comprehensive and structured test suites to identify NLP systems' failure cases (e.g., robustness to typos). Röttger et al (2021) recommends fixing errors by sampling or constructing additional training examples resembling failed test cases.

(Not) additional data. Some works that propose example-based specifications explicitly state how the data should or should not be used for system development. Researchers may recommend that the data should be used to optimize the specification. Kirk et al (2022) create two datasets for hate speech detection of emoji-containing texts: one for testing and one for optimization. Conversely, ***not*** *additional data* recommendations state that the data is purely for testing and should not be used for training. For example, Nangia et al (2020) state that CrowS-pairs, the proposed dataset, should be used to measure social biases, not for debiasing systems, stating that debiasing a system in a way that generalizes is challenging and may require larger datasets.

Concrete. In contrast to vague recommendations, concrete recommendations provide comprehensive and detailed recommendations of how to use the specification in the development cycle. E.g., Pfohl et al (2022a) compare several methods to improve the fairness of predictive systems in healthcare, considering both system performance and fairness measures. They use this empirical analysis to recommend a specific fairness criterion (subpopulation-specific calibration), describing how to apply it for system development (prioritizing systems based on validation-set calibrations and setting subpopulation-specific decision thresholds), and what other factors should be considered (transparency, participation of stakeholders in the decision processes, and reasoning about the potential impact of system-informed decisions).

6.5 Case studies

This section presents three representative papers that illustrate our survey questions and aspects of specification overfitting.

HateCheck*: Functional Tests for Hate Speech Detection Models (Röttger et al, 2021)*

The paper introduces HateCheck, a test suite for hate speech detection. HateCheck covers 29 distinct functionalities that examine distinct expressions of hate (e.g., implicit derogation and hate expressed using slur) and contrastive non-hate (e.g., denouncement of hate that quotes it, or abuse targeted at objects).

We view each functionality as a distinct specification corresponding to an underlying system capability. The specifications are example-based: each functionality is assessed through a set of test cases that are either handcrafted (h) or generated through templates (pat). The specification metric is the accuracy computed on the functionality's test cases.

Though the paper does not experiment with specification optimization strategies, it still contrasts specification performance with general task correctness using two standard hate speech datasets (Founta et al, 2018; Davidson et al, 2017). As the suite comprises multiple functionalities, multiple specification values are reported.

Röttger et al (2021) discuss HateCheck's scope and limitations in a dedicated section, highlighting how HateCheck has limited coverage. That is, good performance on HateCheck only reveals the absence of weakness for the tested cases, not generalizable strengths. Notably, the HateCheck benchmark does not produce insight into phenomena that are not covered (e.g., involving other protected groups, languages, and combinations of functionalities).

The impact statement section summarizes HateCheck's scope:

> "HateCheck's intended use is as an evaluative tool for hate speech detection models, providing structured and targeted diagnostic insights into model functionalities. [...] Researchers might overextend claims about the functionalities of their models based on their test performance, which we would consider a misuse of HateCheck" (Röttger et al, 2021, p. 50).

In addition to stating what HateCheck should *not* be used for, it points out *how* HateCheck can aid system development:

> "If poor model performance does stem from biased training data, models could be improved through targeted data augmentation (Gardner et al, 2020). HateCheck users could, for instance, sample or construct additional training cases to resemble test cases from functional tests that their model was inaccurate on, bearing in mind that this additional data might introduce other unforeseen biases. The models we tested would likely benefit from training on additional cases of negated hate, reclaimed slurs and counter speech" (Röttger et al, 2021, p. 48).

This recommendation contextualizes the specifications vis-a-vis system development (suggests augmenting training data with cases similar to the suite) and points out a possible pitfall—introducing unforeseen biases.

Benchmarking Neural Network Robustness to Common Corruptions and Perturbations (Hendrycks and Dietterich, 2018)

The paper introduces a benchmark that measures the robustness of image classifiers. Specifically, it proposes two datasets – IMAGENET-C, obtained by modifying images from ImageNet (Deng et al, 2009) using a set of 75 algorithmically generated corruptions) and IMAGENET-P, which includes sequences where an image is gradually perturbed with similar corruptions from IMAGENET-C. To validate their datasets, the authors show that there is wide room for improvement on IMAGENET-C by evaluating the performance of several deep learning systems. Additionally, they introduced three methods and architectures that improve corruption robustness. For IMAGENET-P, they propose metrics to measure the stability of the network's predictions on the perturbed images.

The authors state the goal of IMAGETNET-C in the following manner:

> "We hope that this will serve as a general dataset for benchmarking robustness to image corruptions and prevent methodological problems such as moving goal posts and result cherry picking."

Moreover, they recommend future work to use this benchmark because:

> "By defining and benchmarking perturbation and corruption robustness, we facilitate research that can be overcome by future networks which do not rely on spurious correlations or cues inessential to the object's class."

The provided recommendation asserts the importance of measuring perturbation and corruption robustness, but how to act on the insights provided by the benchmark is not discussed, i.e., how to improve system corruption robustness, overall accuracy, and other notions of robustness, such as adversarial robustness. Later work (Schneider et al, 2020) illustrates how focusing on one type of robustness may provide limited insight into other types:

> “We here argue that popular benchmarks to measure model robustness against common corruptions (like ImageNet-C) underestimate model robustness in many (but not all) application scenarios."
>
> “So far, popular image corruption benchmarks like ImageNet-C focus only on ad hoc scenarios in which the tested model has zero prior knowledge about the corruptions it encounters during test time, even if it encounters the same corruption multiple times."

Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare (Pfohl et al, 2022a)

This work compares the estimated *net utility* of predictive systems in healthcare. Specifically, the authors train predictive systems that output a continuous-valued risk score (risk of atherosclerotic cardiovascular disease), which serves as the sole basis for a hypothetical clinical intervention (statin initiation based on decision thresholds). The utility itself is estimated by a secondary system, which parametrizes the relative value of the harms and benefits of the (hypothetical) clinical intervention according to clinical data.

The article reports the overall net utility for the entire patient pool in the data set for different predictive systems, as well as the utility for different subgroups of patients according to sex, racial, and ethnic categories, and the presence of type 2 and type 1 diabetes, rheumatoid arthritis, and chronic kidney disease. Moreover, in addition to the utility itself, an in-depth analysis of the results is reported, including the measurement of equalized odds (Hardt et al, 2016) as a metric to measure fairness across intersectional subgroups (combining race, ethnicity, and sex). Different methods for improving fairness are compared, specifically comparing *in-processing* approaches (Pfohl et al, 2022b) that aim at producing a fair system penalizing worst-group performance during training with *post-processing* approaches that learn the predictive system in an unconstrained manner (unpenalized empirical risk minimization, ERM) and calibrate the decision thresholds to improve fairness on the resulting system.

In our analysis of this article, we view the overall net utility for the entire patient pool as the *task metric* and equalized odds as the *specification metric*. We categorize the *in-processing* and *post-processing* strategies as *direct* attempts to optimize fairness.

The authors of (Pfohl et al, 2022a), in contrast to most other papers in our collection, have clear recommendations on how to use the specification metric during the development process, advising against *in-processing* methods and for threshold calibration as a *post-processing* step:

> “[...] approaches that incorporate fairness considerations into the model training objective typically do not improve model performance or confer greater net benefit"
>
> “[...] we argue for focusing model development efforts on developing calibrated models that predict outcomes well for all patient populations while emphasizing that such efforts are complementary to transparent reporting, participatory design, and reasoning about the

> impact of model-informed interventions in context. "
> “[...] results indicate that models derived from unpenalized ERM should not necessarily be assumed to be well-calibrated in practice, further highlighting the importance of model development, selection, and post-processing strategies that aims to identify the best-fitting, well-calibrated model for each subgroup."

7 Conclusion

In this article, we discussed the problem of specification overfitting—improving specified metrics to the detriment of the underlying goal or other metrics. We analyzed recent impactful papers from diverse AI fields to identify if and how works that propose specification metrics or improvement methods consider specification overfitting. We have found that specification overfitting is often implicitly addressed, with most papers reporting the main task metric or more than one specification metric. However, papers rarely discuss the role of specifications in the system development process, leaving out questions such as how to integrate several (possibly competing) metrics and the assumptions underlying the formulation of a metric. Works that discuss these questions frequently do it in a vague way or leave decision-making to users or domain experts without providing guidelines on how to make such decisions.

Given that the currently developing legislative frameworks use broad terms for the requirements for AI systems, AI providers wanting to achieve legal compliance need to rely on standardized specification metrics set by standardization organizations. Therefore, specifications gain enormous importance in the legal framework and should be carefully evaluated, especially given the specification overfitting issues discussed in this paper. If these are not duly considered on the regulatory and standard-setting level, citizens may not be sufficiently protected from potential harm.

We recommend metric proposers be explicit about how the metric differs from the ideal property it intends to measure. Given that the metric may disagree with other measures of system quality, we also recommend that they provide guidelines or recommendations on making decisions on system selection. We recommend that peer reviewers reward papers with clear delimitations of the scope of a specification metric and that mentioning such limitations should not be seen as a weakness. Method proposers should rigorously measure the impact of the method on other metrics, including the task metric. One way to do so is by defining evaluation scenarios that are robust to specification overfitting, such as using controlled splits that hold out metrics. Practitioners, regulators, and standard-setting bodies should be aware of the misincentives that can arise from using a narrow set of metrics for evaluation when these same metrics can be a target in optimization and system selection.

Acknowledgments. This research has been funded by the Vienna Science and Technology Fund (WWTF) [10.47379/VRG19008] “Knowledge-infused Deep Learning for Natural Language Processing”.

Author contributions. BR conceptualized the study, contributed to the study design, and analyzed papers for the survey. PHLA contributed to the study design and collected, filtered, and analyzed papers for the survey. YX contributed to the study design and analyzed papers for the survey. SK and CK bridged the technical aspects of the study to the legal framework and regulatory practices. All authors contributed to the paper writing and revising.

Declarations

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

Aceña V, Martín de Diego I, R. Fernández R, et al (2022) Minimally overfitted learners: A general framework for ensemble learning. Knowledge-Based Systems 254:109669. https://doi.org/10.1016/j.knosys.2022.109669

Angwin J, Jeff Larson SM, Kirchner L (2016) Machine Bias. Pro Publica URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arulkumaran K, Deisenroth MP, Brundage M, et al (2017) Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine 34(6):26–38. https://doi.org/10.1109/MSP.2017.2743240

Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, URL http://arxiv.org/abs/1409.0473

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif L Rev 104:671

Barocas S, Hardt M, Narayanan A (2019) Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org, http://www.fairmlbook.org

Bartolo M, Thrush T, Jia R, et al (2021) Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation. In: Moens MF, Huang X, Specia L, et al (eds) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 8830–8848, https://doi.org/10.18653/v1/2021.emnlp-main.696

Ben-David S, Blitzer J, Crammer K, et al (2010) A theory of learning from different domains. Machine Learning 79(1):151–175. https://doi.org/10.1007/s10994-009-5152-4

Bhatore S, Mohan L, Reddy YR (2020) Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology 4(1):111–138. https://doi.org/10.1007/s42786-020-00020-3

Black E, Yeom S, Fredrikson M (2020) FlipTest: Fairness testing via optimal transport. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM, Barcelona Spain, pp 111–121, https://doi.org/10.1145/3351095.3372845

Bleher H, Braun M (2023) Reflections on Putting AI Ethics into Practice: How Three AI Ethics Approaches Conceptualize Theory and Practice. Science and Engineering Ethics 29(3):21. https://doi.org/10.1007/s11948-023-00443-3

Borisov V, Leemann T, Seßler K, et al (2022) Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems pp 1–21. https://doi.org/10.1109/TNNLS.2022.3229161

Bronstein MM, Bruna J, LeCun Y, et al (2017) Geometric Deep Learning: Going beyond Euclidean data. IEEE Signal Processing Magazine 34(4):18–42. https://doi.org/10.1109/MSP.2017.2693418

Buffelli D, Lió P, Vandin F (2022) SizeShiftReg: A Regularization Method for Improving Size-Generalization in Graph Neural Networks. Advances in Neural Information Processing Systems 35:31871–31885. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ceeb3fa5be458f08fbb12a5bb783aac8-Paper-Conference.pdf

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, Proceedings of Machine Learning Research, vol 81. PMLR, pp 77–91, URL https://proceedings.mlr.press/v81/buolamwini18a.html

Chen J, Kallus N, Mao X, et al (2019) Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAT* '19, pp 339–348, https://doi.org/10.1145/3287560.3287594

Chen Y, Zhou K, Bian Y, et al (2022) Pareto Invariant Risk Minimization: Towards Mitigating the Optimization Dilemma in Out-of-Distribution Generalization. In: The Eleventh International Conference on Learning Representations

Cheng M, Wei W, Hsieh CJ (2019) Evaluating and Enhancing the Robustness of Dialogue Systems: A Case Study on a Negotiation Agent. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 3325–3335, https://doi.org/10.18653/v1/N19-1336

Cheng M, Yi J, Chen PY, et al (2020) Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples. Proceedings of the AAAI Conference on Artificial Intelligence 34(04):3601–3608. https://doi.org/10.1609/aaai.v34i04.5767

Cheng M, Lei Q, Chen PY, et al (2022) CAT: Customized Adversarial Training for Improved Robustness. In: Raedt LD (ed) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022. ijcai.org, pp 673–679, https://doi.org/10.24963/IJCAI.2022/95

Clark K, Luong MT, Le QV, et al (2020) ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=r1xMH1BtvB

Clarysse J, Hörrmann J, Yang F (2022) Why adversarial training can hurt robust accuracy. In: The Eleventh International Conference on Learning Representations

Coston A, Mishler A, Kennedy EH, et al (2020) Counterfactual risk assessments, evaluation, and fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAT* '20, pp 582–593, https://doi.org/10.1145/3351095.3372851, 1909.00066

Cotter A, Gupta M, Jiang H, et al (2019) Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints. In: Proceedings of the 36th International Conference on Machine Learning. PMLR, pp 1397–1405

Croce F, Andriushchenko M, Sehwag V, et al (2021) RobustBench: A standardized adversarial robustness benchmark. In: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2010.09670

D'Amour A, Heller K, Moldovan D, et al (2022) Underspecification presents challenges for credibility in modern machine learning. The Journal of Machine Learning Research 23(1):10237–10297

Dapello J, Kar K, Schrimpf M, et al (2022) Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness. In: The Eleventh International Conference on Learning Representations. Cold Spring Harbor Laboratory, pp 2022–07

Davidson T, Warmsley D, Macy M, et al (2017) Automated hate speech detection and the problem of offensive language. Proceedings of the International AAAI Conference on Web and Social Media 11(1):512–515. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14955

Deng J, Dong W, Socher R, et al (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee, pp 248–255

Deng Z, Zhang J, Zhang L, et al (2023) FIFA: making fairness more generalizable in classifiers trained on imbalanced data. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, URL https://openreview.net/pdf?id=zVrw4OH1Lch

Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186, URL https://www.aclweb.org/anthology/N19-1423

Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=YicbFdNTTy

Elkahky A, Webster K, Andor D, et al (2018) A Challenge Set and Methods for Noun-Verb Ambiguity. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp 2562–2572, https://doi.org/10.18653/v1/D18-1277

Esteva A, Chou K, Yeung S, et al (2021) Deep learning-enabled medical computer vision. npj Digital Medicine 4(1):5. https://doi.org/10.1038/s41746-020-00376-2

European Parliament and Council of the European Union (2012) Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation. URL https://eur-lex.europa.eu/eli/reg/2012/1025/oj

European Parliament and Council of the European Union (2022) Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive). URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52022PC0496

European Parliament and Council of the European Union (2024) Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial intelligence Act)

Fan W, Ma Y, Li Q, et al (2019) Graph Neural Networks for Social Recommendation. In: The World Wide Web Conference. Association for Computing Machinery, New York, NY, USA, WWW '19, p 417–426, https://doi.org/10.1145/3308558.3313488

Fatemi Z, Xing C, Liu W, et al (2023) Improving Gender Fairness of Pre-Trained Language Models without Catastrophic Forgetting. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Toronto, Canada, pp 1249–1262, https://doi.org/10.18653/v1/2023.acl-short.108

Fjeld J, Achten N, Hilligoss H, et al (2020) Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication 2020-1

Founta A, Djouvas C, Chatzakou D, et al (2018) Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. Proceedings of the International AAAI Conference on Web and Social Media 12(1). URL https://ojs.aaai.org/index.php/ICWSM/article/view/14991

Friedler SA, Scheidegger C, Venkatasubramanian S (2021) The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. Commun ACM 64(4):136–143. https://doi.org/10.1145/3433949

Fukushima K (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36(4):193–202. https://doi.org/10.1007/BF00344251

Gan WC, Ng HT (2019) Improving the Robustness of Question Answering Systems to Question Paraphrasing. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp 6065–6075, https://doi.org/10.18653/v1/P19-1610

Gardner M, Artzi Y, Basmov V, et al (2020) Evaluating models' local decision boundaries via contrast sets. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 1307–1323, https://doi.org/10.18653/v1/2020.findings-emnlp.117, URL https://aclanthology.org/2020.findings-emnlp.117

Geirhos R, Rubisch P, Michaelis C, et al (2018) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations

Goodfellow I, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: International Confserence on Learning Representations, URL http://arxiv.org/abs/1412.6572

Gowal S, Rebuffi SA, Wiles O, et al (2021) Improving Robustness using Generated Data. In: Advances in Neural Information Processing Systems, vol 34. Curran Associates, Inc., pp 4218–4233

Guo J, Chen Y, Hao Y, et al (2022) Towards Comprehensive Testing on the Robustness of Cooperative Multi-agent Reinforcement Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022. IEEE, pp 114–121, https://doi.org/10.1109/CVPRW56347.2022.00022

Guo Y, Zhang C, Zhang C, et al (2018) Sparse DNNs with Improved Adversarial Robustness. In: Advances in Neural Information Processing Systems, vol 31. Curran Associates, Inc.

Hagendorff T (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines 30(1):99–120. https://doi.org/10.1007/s11023-020-09517-8

Han K, Wang Y, Chen H, et al (2023) A Survey on Vision Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(1):87–110. https://doi.org/10.1109/TPAMI.2022.3152247

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Advances in neural information processing systems 29

Havasi M, Jenatton R, Fort S, et al (2020) Training independent subnetworks for robust prediction. In: International Conference on Learning Representations

He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 770–778, https://doi.org/10.1109/CVPR.2016.90

He X, Yan S, Hu Y, et al (2005) Face recognition using Laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(3):328–340. https://doi.org/10.1109/TPAMI.2005.55

Hendrycks D, Dietterich T (2018) Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In: International Conference on Learning Representations

Hendrycks D, Mazeika M, Kadavath S, et al (2019a) Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty. In: Advances in Neural Information Processing Systems, vol 32. Curran Associates, Inc.

Hendrycks D, Mu* N, Cubuk ED, et al (2019b) AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In: International Conference on Learning Representations

High-Level Expert Group on AI (2019) Ethics guidelines for trustworthy AI, High-Level Expert Group on AI. URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hu Y, Yang J, Chen L, et al (2023) Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

Iniesta R (2023) The human role to guarantee an ethical AI in healthcare: A five-facts approach. AI and Ethics https://doi.org/10.1007/s43681-023-00353-x

Jackson M (1995) The world and the machine. In: Proceedings of the 17th International Conference on Software Engineering. Association for Computing Machinery, New York, NY, USA, ICSE '95, p 283–292, https://doi.org/10.1145/225014.225041

Jacobs AZ, Wallach H (2021) Measurement and fairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '21, p 375–385, https://doi.org/10.1145/3442188.3445901

Jakubovitz D, Giryes R (2018) Improving DNN Robustness to Adversarial Attacks Using Jacobian Regularization. In: Ferrari V, Hebert M, Sminchisescu C, et al (eds) Computer Vision – ECCV 2018. Springer International Publishing, Cham, Lecture Notes in Computer Science, pp 525–541, https://doi.org/10.1007/978-3-030-01258-8_32

Ji Z, Lee N, Frieske R, et al (2023) Survey of Hallucination in Natural Language Generation. ACM Comput Surv 55(12). https://doi.org/10.1145/3571730

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nature Machine Intelligence 1(9):389–399

Jung S, Park T, Chun S, et al (2022) Re-weighting Based Group Fairness Regularization via Classwise Robust Optimization. In: The Eleventh International Conference on Learning Representations

Karpukhin V, Levy O, Eisenstein J, et al (2019) Training on Synthetic Noise Improves Robustness to Natural Noise in Machine Translation. In: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). Association for Computational Linguistics, Hong Kong, China, pp 42–47, https://doi.org/10.18653/v1/D19-5506

Kasneci E, Sessler K, Küchemann S, et al (2023) ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences 103:102274. https://doi.org/https://doi.org/10.1016/j.lindif.2023.102274, URL https://www.sciencedirect.com/science/article/pii/S1041608023000195

Kiden S, Stahl B, Townsend B, et al (2024) Responsible AI governance: A response to UN interim report on governing AI for humanity. https://eprints.soton.ac.uk/488908/, https://doi.org/10.5258/SOTON/PP0057

Kirichenko P, Izmailov P, Wilson AG (2022) Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. In: The Eleventh International Conference on Learning Representations

Kirk H, Vidgen B, Rottger P, et al (2022) Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-Based Hate. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp 1352–1368, https://doi.org/10.18653/v1/2022.naacl-main.97

Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent Trade-Offs in the Fair Determination of Risk Scores. In: Papadimitriou CH (ed) 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Leibniz International Proceedings in Informatics (LIPIcs), vol 67. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp 43:1–43:23, https://doi.org/10.4230/LIPIcs.ITCS.2017.43, URL http://drops.dagstuhl.de/opus/volltexte/2017/8156

Komiyama J, Takeda A, Honda J, et al (2018) Nonconvex Optimization for Regression with Fairness Constraints. In: Proceedings of the 35th International Conference on Machine Learning. PMLR, pp 2737–2746

Kononenko I (2001) Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine 23(1):89–109. https://doi.org/https://doi.org/10.1016/S0933-3657(01)00077-X, URL https://www.sciencedirect.com/science/article/pii/S093336570100077X

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges C, Bottou L, et al (eds) Advances in Neural Information Processing Systems, vol 25. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Lake B, Baroni M (2018) Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 80. PMLR, pp 2873–2882, URL https://proceedings.mlr.press/v80/lake18a.html

Lample G, Ballesteros M, Subramanian S, et al (2016) Neural Architectures for Named Entity Recognition. In: Knight K, Nenkova A, Rambow O (eds) Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pp 260–270, https://doi.org/10.18653/v1/N16-1030, URL https://aclanthology.org/N16-1030

LeCun Y, Boser B, Denker JS, et al (1989) Backpropagation Applied to Handwritten Zip Code Recognition. Neural Computation 1(4):541–551. https://doi.org/10.1162/neco.1989.1.4.541

Lee J, Kim G, Olfat M, et al (2022) Fast and Efficient MMD-Based Fair PCA via Optimization over Stiefel Manifold. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 7363–7371, https://doi.org/10.1609/aaai.v36i7.20699, URL https://ojs.aaai.org/index.php/AAAI/article/view/20699

Levine S, Finn C, Darrell T, et al (2016) End-to-end training of deep visuomotor policies. Journal of Machine Learning Research 17(39):1–40. URL http://jmlr.org/papers/v17/15-522.html

Levy I, Bogin B, Berant J (2023) Diverse Demonstrations Improve In-context Compositional Generalization. In: Rogers A, Boyd-Graber JL, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp 1401–1422, https://doi.org/10.18653/v1/2023.acl-long.78

Ley M (2002) The DBLP computer science bibliography: Evolution, research issues, perspectives. In: International symposium on string processing and information retrieval, Springer, pp 1–10

Li Q, Guo Y, Zuo W, et al (2023a) Squeeze Training for Adversarial Robustness. In: The Eleventh International Conference on Learning Representations, URL https://openreview.net/forum?id=Z_tmYu060Kr

Li X, Wu P, Su J (2023b) Accurate fairness: Improving individual fairness without trading accuracy. In: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'23/IAAI'23/EAAI'23, vol 37. AAAI Press, pp 14312–14320, https://doi.org/10.1609/aaai.v37i12.26674

Liang Y, Sun Y, Zheng R, et al (2022) Efficient Adversarial Training without Attacking: Worst-Case-Aware Robust Reinforcement Learning. In: Advances in Neural Information Processing Systems

Lin S, Hilton J, Evans O (2022) TruthfulQA: Measuring how models mimic human falsehoods. In: Muresan S, Nakov P, Villavicencio A (eds) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Dublin, Ireland, pp 3214–3252, https://doi.org/10.18653/v1/2022.acl-long.229, URL https://aclanthology.org/2022.acl-long.229

Liu EZ, Haghgoo B, Chen AS, et al (2021) Just Train Twice: Improving Group Robustness without Training Group Information. In: Proceedings of the 38th International Conference on Machine Learning. PMLR, pp 6781–6792

Liu NF, Schwartz R, Smith NA (2019a) Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 2171–2179, https://doi.org/10.18653/v1/N19-1225

Liu Y, Ott M, Goyal N, et al (2019b) RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692. URL http://arxiv.org/abs/1907.11692, 1907.11692

Lütjens B, Everett M, How JP (2020) Certified adversarial robustness for deep reinforcement learning. In: Kaelbling LP, Kragic D, Sugiura K (eds) Proceedings of the Conference on Robot Learning, Proceedings of Machine Learning Research, vol 100. PMLR, pp 1328–1337, URL https://proceedings.mlr.press/v100/lutjens20a.html

Luz de Araujo PH, Roth B (2023) Cross-functional Analysis of Generalization in Behavioral Learning. Transactions of the Association for Computational Linguistics 11:1066–1081. https://doi.org/10.1162/tacl_a_00590, https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00590/2154470/tacl_a_00590.pdf

Ma M, Ren J, Zhao L, et al (2022) Are Multimodal Transformers Robust to Missing Modality? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 18177–18186

Madaio MA, Stark L, Wortman Vaughan J, et al (2020) Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, CHI '20, p 1–14, https://doi.org/10.1145/3313831.3376445

Madras D, Pitassi T, Zemel R (2018) Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In: Advances in Neural Information Processing Systems, vol 31. Curran Associates, Inc.

Malik MM (2020) A hierarchy of limitations in machine learning. CoRR abs/2002.05193. URL https://arxiv.org/abs/2002.05193

Mehrabi N, Morstatter F, Saxena N, et al (2021) A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv 54(6). https://doi.org/10.1145/3457607

Min J, McCoy RT, Das D, et al (2020) Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 2339–2352, https://doi.org/10.18653/v1/2020.acl-main.212

Minaee S, Boykov Y, Porikli F, et al (2022) Image Segmentation Using Deep Learning: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 44(7):3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968

Mishler A, Kennedy EH, Chouldechova A (2021) Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '21, pp 386–400, https://doi.org/10.1145/3442188.3445902

Mnih V, Kavukcuoglu K, Silver D, et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533. https://doi.org/10.1038/nature

14236, URL https://doi.org/10.1038/nature14236

Naik A, Ravichander A, Sadeh N, et al (2018) Stress Test Evaluation for Natural Language Inference. In: Proceedings of the 27th International Conference on Computational Linguistics. Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 2340–2353

Nangia N, Vania C, Bhalerao R, et al (2020) CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In: Webber B, Cohn T, He Y, et al (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Online, pp 1953–1967, https://doi.org/10.18653/v1/2020.emnlp-main.154, URL https://aclanthology.org/2020.emnlp-main.154

Narasimhan H, Cotter A, Gupta M (2019) Optimizing Generalized Rate Metrics with Three Players. In: Advances in Neural Information Processing Systems, vol 32. Curran Associates, Inc.

OECD (2019) OECD AI Principles overview. OECD AI Policy Observatory URL https://oecd.ai/en/ai-principles

Ouyang L, Wu J, Jiang X, et al (2022) Training language models to follow instructions with human feedback. In: Koyejo S, Mohamed S, Agarwal A, et al (eds) Advances in Neural Information Processing Systems, vol 35. Curran Associates, Inc., pp 27730–27744, URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

Pessach D, Shmueli E (2023) A Review on Fairness in Machine Learning. ACM Computing Surveys 55(3):1–44. https://doi.org/10.1145/3494672

Petersen E, Ganz M, Holm S, et al (2023) On (assessing) the fairness of risk score models. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '23, pp 817–829, https://doi.org/10.1145/3593013.3594045

Pfohl S, Xu Y, Foryciarz A, et al (2022a) Net Benefit, Calibration, Threshold Selection, and Training Objectives for Algorithmic Fairness in Healthcare. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '22, pp 1039–1052, https://doi.org/10.1145/3531146.3533166

Pfohl SR, Zhang H, Xu Y, et al (2022b) A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. Scientific reports 12(1):3254

Qiu L, Shaw P, Pasupat P, et al (2022) Improving Compositional Generalization with Latent Structure and Data Augmentation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp 4341–4362, https://doi.org/10.18653/v1/2022.naacl-main.323

Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 139. PMLR, pp 8748–8763, URL https://proceedings.mlr.press/v139/radford21a.html

Raffel C, Shazeer N, Roberts A, et al (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21(140):1–67. URL http://jmlr.org/papers/v21/20-074.html

Rahmattalabi A, Jabbari S, Lakkaraju H, et al (2021) Fair influence maximization: a welfare optimization approach. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 11630–11638, https://doi.org/10.1609/aaai.v35i13.17383, URL https://ojs.aaai.org/index.php/AAAI/article/view/17383

Ramesh A, Pavlov M, Goh G, et al (2021) Zero-Shot Text-to-Image Generation. In: Meila M, Zhang T (eds) Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 139. PMLR, pp 8821–8831, URL https://proceedings.mlr.press/v139/ramesh21a.html

Rebuffi SA, Gowal S, Calian DA, et al (2021) Data Augmentation Can Improve Robustness. In: Advances in Neural Information Processing Systems, 2111.05328

Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 1135–1144

Ribeiro MT, Wu T, Guestrin C, et al (2020) Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 4902–4912, https://doi.org/10.18653/v1/2020.acl-main.442

Roelofs R, Shankar V, Recht B, et al (2019) A meta-analysis of overfitting in machine learning. Advances in Neural Information Processing Systems 32

Roh Y, Lee K, Whang SE, et al (2021) Sample Selection for Fair and Robust Training. In: Advances in Neural Information Processing Systems

Ross A, Doshi-Velez F (2018) Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients. In: Proceedings of the AAAI Conference on Artificial Intelligence, https://doi.org/10.1609/aaai.v32i1.11504

Röttger P, Vidgen B, Nguyen D, et al (2021) HateCheck: Functional Tests for Hate Speech Detection Models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, pp 41–58, https://doi.org/10.18653/v1/2021.acl-long.4

Ruis L, Andreas J, Baroni M, et al (2020) A Benchmark for Systematic Generalization in Grounded Language Understanding. In: Advances in Neural Information Processing Systems, vol 33. Curran Associates, Inc., pp 19861–19872

Russakovsky O, Deng J, Su H, et al (2015) ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y

Schick T, Udupa S, Schütze H (2021) Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. Transactions of the Association for Computational Linguistics 9:1408–1424. https://doi.org/10.1162/tacl_a_00434, https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00434/1979270/tacl_a_00434.pdf

Schneider S, Rusak E, Eck L, et al (2020) Improving Robustness against Common Corruptions by Covariate Shift Adaptation. In: Advances in Neural Information Processing Systems, vol 33. Curran Associates, Inc., pp 11539–11551

Sehwag V, Mahloujifar S, Handina T, et al (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

Shalev-Shwartz S, Ben-David S (2014) Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press

Sinha A, Namkoong H, Duchi J (2019) Certifying Some Distributional Robustness with Principled Adversarial Training. In: International Conference on Learning Representations

Skalse JMV, Howe NHR, Krasheninnikov D, et al (2022) Defining and characterizing reward gaming. In: Oh AH, Agarwal A, Belgrave D, et al (eds) Advances in Neural Information Processing Systems, URL https://openreview.net/forum?id=yb3HOXO3lX2

Socher R, Perelygin A, Wu J, et al (2013) Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In: Yarowsky D, Baldwin T, Korhonen A, et al (eds) Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Seattle, Washington, USA, pp 1631–1642, URL https://aclanthology.org/D13-1170

Sreenu G, Saleem Durai MA (2019) Intelligent video surveillance: a review through deep learning techniques for crowd analysis. Journal of Big Data 6(1):48. https://doi.org/10.1186/s40537-019-0212-5

Sun L, Dou Y, Yang C, et al (2023) Adversarial Attack and Defense on Graph Data: A Survey. IEEE Transactions on Knowledge and Data Engineering 35(8):7693–7711. https://doi.org/10.1109/TKDE.2022.3201243

Sun Y, Wang X, Liu Z, et al (2020) Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In: Proceedings of the 37th International Conference on Machine Learning. PMLR, pp 9229–9248

Sutton RS, Barto AG (2018) Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA

Szegedy C, Vanhoucke V, Ioffe S, et al (2016) Rethinking the Inception Architecture for Computer Vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2818–2826, https://doi.org/10.1109/CVPR.2016.308

Taskesen B, Blanchet J, Kuhn D, et al (2021) A Statistical Test for Probabilistic Fairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, FAccT '21, pp 648–665, https://doi.org/10.1145/3442188.3445927

The European Comission (2003) General Guidelines for the Cooperation between CEN, Cenelec and ETSI and the European Commission and the European Free Trade Association. URL https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52003XC0416(03)

The European Comission (2008) New legislative framework. URL https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en

The European Comission (2018) Factsheet: Artificial Intelligence for Europe. URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al (2023) Large language models in medicine. Nature Medicine 29(8):1930–1940. https://doi.org/10.1038/s41591-023-02448-8

Tjeng V, Xiao KY, Tedrake R (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

Tu L, Lalwani G, Gella S, et al (2020) An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. Transactions of the Association for Computational Linguistics 8:621–633. http s://doi.org/10.1162/tacl_a_00335, https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00335/1923506/tacl_a_00335.pdf

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008

Veale M, Borgesius FZ (2021) Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International 22(4):97–112

Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv JL & Tech 31:841

Wang B, Wang S, Cheng Y, et al (2020a) InfoBERT: Improving Robustness of Language Models from An Information Theoretic Perspective. In: International Conference on Learning Representations

Wang S, Guo W, Narasimhan H, et al (2020b) Robust Optimization for Fairness with Noisy Protected Groups. In: Advances in Neural Information Processing Systems, vol 33. Curran Associates, Inc., pp 5190–5203

Wang T, Sridhar R, Yang D, et al (2022a) Identifying and Mitigating Spurious Correlations for Improving Robustness in NLP Models. In: Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, Seattle, United States, pp 1719–1729, https://doi.org/10.18653/v1/2022.findings-naacl.130

Wang X, Wang H, Yang D (2022b) Measure and improve robustness in NLP models: A survey. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp 4569–4586, https://doi.org/10.18653/v1/2022.naacl-main.339, URL https://aclanthology.org/2022.naacl-main.339

Wang Y, Bansal M (2018) Robust Machine Comprehension Models via Adversarial Training. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). Association for Computational Linguistics, New Orleans, Louisiana, pp 575–581, https://doi.org/10.18653/v1/N18-2091

Wang Y, Zou D, Yi J, et al (2019) Improving Adversarial Robustness Requires Revisiting Misclassified Examples. In: International Conference on Learning Representations

Webb GI (2010) Overfitting, Springer US, Boston, MA, pp 744–744. https://doi.org/10.1007/978-0-387-30164-8_623, URL https://doi.org/10.1007/978-0-387-30164-8_623

Wei J, Bosma M, Zhao V, et al (2022a) Finetuned language models are zero-shot learners. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=gEZrGCozdqR

Wei J, Tay Y, Bommasani R, et al (2022b) Emergent abilities of large language models. Transactions on Machine Learning Research URL https://openreview.net/forum?id=yzkSU5zdwD

Weng TW, Zhang H, Chen PY, et al (2018) Evaluating the robustness of neural networks: An extreme value theory approach. In: International Conference on Learning Representations (ICLR)

Wiener N (1960) Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. Science 131(3410):1355–1358

Wu Y, Jiang A, Ba J, et al (2020) INT: An Inequality Benchmark for Evaluating Generalization in Theorem Proving. In: International Conference on Learning Representations

Wu Z, Pan S, Chen F, et al (2021) A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems 32(1):4–24. https://doi.org/10.1109/TNNLS.2020.2978386

Xie C, Wu Y, van der Maaten L, et al (2019) Feature Denoising for Improving Adversarial Robustness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 501–509

Xiong Z, Wang D, Liu X, et al (2020) Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. Journal of Medicinal Chemistry 63(16):8749–8760. https://doi.org/10.1021/acs.jmedchem.9b00959, pMID: 31408336

Yuan A, Coenen A, Reif E, et al (2022) Wordcraft: Story Writing With Large Language Models. In: 27th International Conference on Intelligent User Interfaces. Association for Computing Machinery, New York, NY, USA, IUI '22, p 841–852, https://doi.org/10.1145/3490099.3511105

Zan D, Chen B, Zhang F, et al (2023) Large language models meet NL2Code: A survey. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp 7443–7464, https://doi.org/10.18653/v1/2023.acl-long.411, URL https://aclanthology.org/2023.acl-long.411

Zhang C, Tian Y, Ju M, et al (2022a) Chasing All-Round Graph Representation Robustness: Model, Training, and Optimization. In: The Eleventh International

Conference on Learning Representations

Zhang H, Chen H, Xiao C, et al (2019) Towards Stable and Efficient Training of Verifiably Robust Neural Networks. In: International Conference on Learning Representations

Zhang M, Levine S, Finn C (2022b) MEMO: Test Time Robustness via Adaptation and Augmentation. Advances in Neural Information Processing Systems 35:38629–38642

Zhang M, Sohoni NS, Zhang HR, et al (2022c) Correct-N-Contrast: A Contrastive Approach for Improving Robustness to Spurious Correlations. In: Proceedings of the 39th International Conference on Machine Learning. PMLR, pp 26484–26516, 2203.01517

Zhang WE, Sheng QZ, Alhazmi A, et al (2020) Adversarial attacks on deep-learning models in natural language processing: A survey. ACM Trans Intell Syst Technol 11(3). https://doi.org/10.1145/3374217

Zhuo TY, Li Z, Huang Y, et al (2023) On Robustness of Prompt-based Semantic Parsing with Large Pre-trained Language Model: An Empirical Study on Codex. In: Vlachos A, Augenstein I (eds) Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Dubrovnik, Croatia, pp 1090–1102, https://doi.org/10.18653/v1/2023.eacl-main.77