Never Compromise to Vulnerabilities: A Comprehensive Survey on AI Governance

Yuchu Jiang[1*], Jian Zhao[1,2*†], Yuchen Yuan[1*], Tianle Zhang[1*], Yao Huang[3*], Yanghao Zhang[4], Yan Wang[1,5], Yanshu Li[1,6], Xizhong Guo[1,6], Yusheng Zhao[1,7], Jun Zhang[1,6], Zhi Zhang[8], Xiaojian Lin[9], Yixiu Zou[12] Haoxuan Ma[10], Yuhu Shang[11], Yuzhi Hu[13], Keshu Cai[13], Ruochen Zhang[3], Boyuan Chen[14], Yilan Gao[2], Ziheng Jiao[2], Yi Qin[15], Shuangjun Du[6], Tong Xiao[15], Zhekun Liu[15], Yu Chen[15], Xuankun Rong[24], Rui Wang[15], Yejie Zheng[16], Zhaoxin Fan[3], Murat Sensoy[17], Hongyuan Zhang[18], Pan Zhou[19], Lei Jin[11], Hao Zhao[9], Xu Yang[10], Jiaojiao Zhao[8], Jianshu Li[20], Joey Tianyi Zhou[21], Zhi-Qi Cheng[13], Longtao Huang[22], Zhiyi Liu[18], Zheng Zhu[23] Jianan Li[12], Gang Wang[24], Qi Li[6], Xu-Yao Zhang[6], Yaodong Yang[14], Mang Ye[25], Wenqi Ren[26], Zhaofeng He[11], Hang Su[9], Rongrong Ni[15], Liping Jing[15], Xingxing Wei[3], Junliang Xing[9], Massimo Alioto[27], Shengmei Shen[28], Petia Radeva[29], Dacheng Tao[30], Ya-Qin Zhang[9], Shuicheng Yan[27], Chi Zhang[1 †], Zhongjiang He[1 †], and Xuelong Li[1 †]

[1]Institute of Artificial Intelligence (TeleAI), China Telecom [2]Northwestern Polytechnical University [3]Beihang University [4]Imperial College London [5]University of Edinburgh [6]University of Chinese Academy of Sciences [7]University of Science and Technology of China [8]University of Amsterdam [9]Tsinghua University [10]Southeast University [11]Beijing University of Posts and Telecommunications [12]Beijing Institute of Technology [13]University of Washington [14]Peking University [15]Beijing Jiaotong University [16]Shanghai Collaborative Innovation Center for AI Social Governance [17]Amazon [18]The University of Hong Kong [19]Singapore Management University [20]Ant Group [21]CFAR, Agency for Science, Technology and Research [22]Alibaba Group [23]GigaAI [24]Beijing Institute of Basic Medical Sciences [25]Wuhan University [26]Sun Yat-sen University [27]National University of Singapore [28]Pensees Singapore [29]University of Barcelona [30]Nanyang Technological University

arXiv:2508.08789v4 [cs.CR] 18 Aug 2025

Abstract—The rapid advancement of artificial intelligence (AI) has significantly expanded its capabilities across diverse domains. However, this also introduces complex technical vulnerabilities, such as algorithmic biases and adversarial sensitivity, that can carry significant societal risks, including misinformation, inequity, computing security issues, physical-world accident and declines in public credibility. These challenges underscore the pressing need for AI governance to inform the development and deployment of AI technologies. To meet this need, we propose a comprehensive AI governance framework that integrates both technical and societal dimensions simultaneously. Specifically, we categorize governance into three interconnected aspects: **Intrinsic Security** (internal system reliability), **Derivative Security** (external real-world harms), and **Social Ethics** (value alignment and accountability). Uniquely, we integrate technical methodologies, emerging evaluation benchmarks, and policy perspectives to construct a governance framework that actively supports transparency, accountability, and public trust. Through a systematic review of over 300 references, we identify three critical systematic challenges: (1) the generalization gap, where existing defenses fail to adapt to the evolving threats; (2) evaluation protocols that insufficiently reflect real-world deployment risks; and (3) fragmented regulatory landscapes that produce inconsistent oversight and enforcement. We attribute these failures to a fundamental misalignment in current practices, where governance is treated as an afterthought rather than a foundational design principle. As a result, existing efforts tend to be reactive and piecemeal, falling short in addressing the inherently interconnected nature of technical reliability and societal trust. In response, our study provides a comprehensive landscape analysis and articulates an integrated research agenda that bridges technical rigor with social responsibility. This framework equips researchers, engineers, and policymakers with actionable insights for designing AI systems that not only exhibit performance robustness but also align with ethical imperatives and public trust. The repository is available at https://github.com/ZTianle/Awesome-AI-SG.

Index Terms—AI Governance, Intrinsic Security, Derivative Security, Social Ethics, Responsible AI

1 Introduction

The rapid advancement of artificial intelligence (AI), particularly the emergence of large language models (LLMs), has brought transformative changes across science [1], industry [2], and society [3]. Specifically, these models excel in reasoning [4], content generation [5], and decision support [6], enabling a wide range of applications across education, healthcare, law, and public services. This widespread adoption also reflects the growing confidence in AI's potential to augment human capabilities and drive societal progress.

However, the deployment of AI systems at scale has simultaneously surfaced a host of new risks. Unlike conventional vulnerabilities that primarily affect software system functionality, AI-specific risks may manifest in more concerning ways that can undermine public trust and cause widespread harm. For instance, LLMs can be manipulated through prompt injection attacks to bypass safety mechanisms and generate harmful or illegal content [7]. Generative models enable the creation of convincing deepfakes for large-scale misinformation campaigns and non-consensual intimate imagery [8], fundamentally challenging our ability to distinguish authentic content from fabricated material. Perhaps most critically, AI hallucinations [9] in high-stakes

* *Equal contribution: Yuchu Jiang (kamichanw@seu.edu.cn), Jian Zhao, Yuchen Yuan, Tianle Zhang ({zhaoj90, yuanyc2, zhangtl15}@chinatelecom.cn) and Yao Huang (y_huang@buaa.edu.cn).*

† *Corresponding authors: Jian Zhao, Chi Zhang, Zhongjiang He and Xuelong Li ({zhaoj90, zhangc120, hezj, xuelong_li}@chinatelecom.cn).*

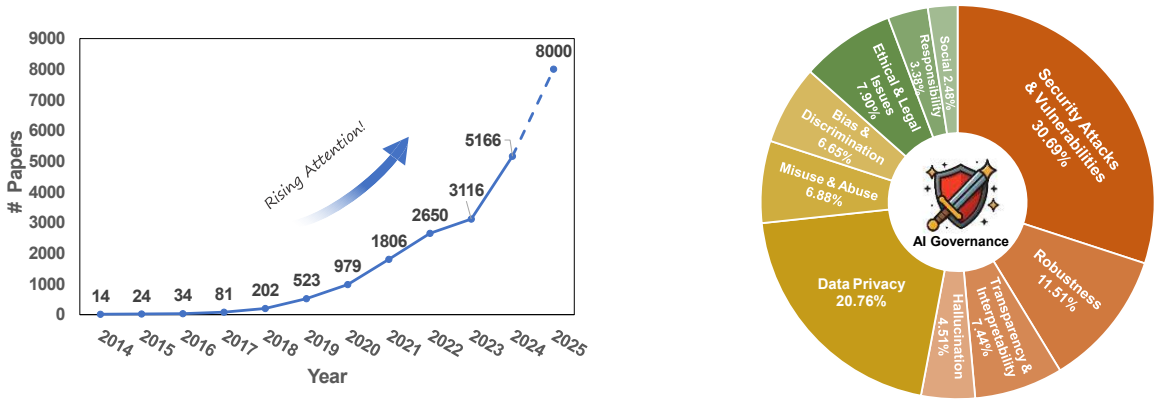

Fig. 1. **Left**: The number of AI governance papers published over the past four years. **Right**: The distribution of research across different dimensions.

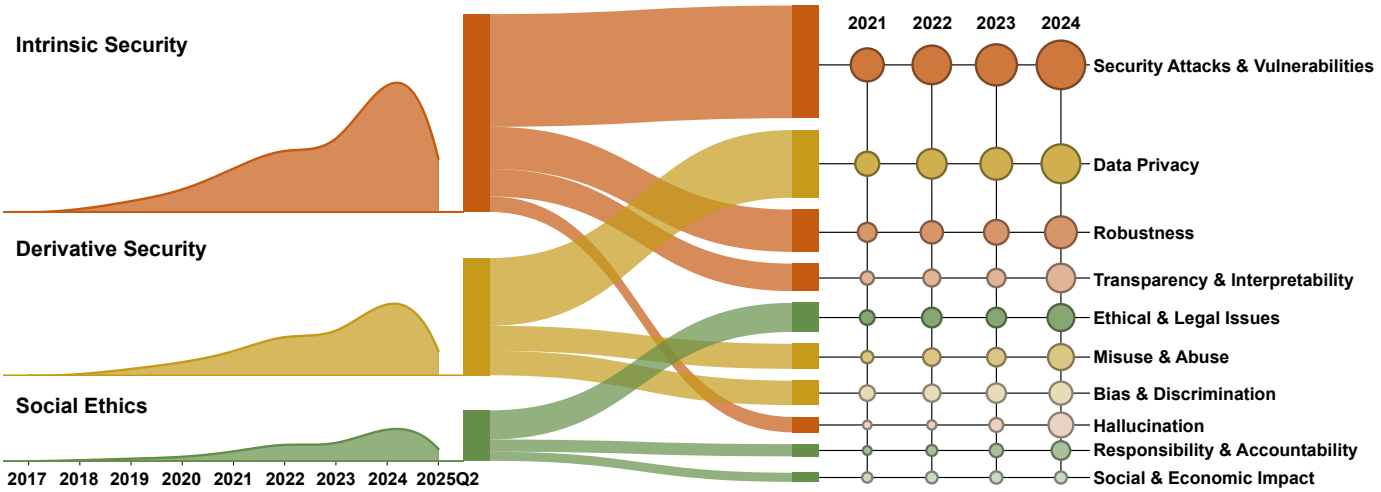

Fig. 2. **Left**: The quarterly trend in the number of published governance papers across different dimensions. **Middle**: Proportional distribution of intrinsic security, derivative security, and social ethics. **Right**: The annual trend in the number of governance research papers published on various dimensions, presented in descending order from highest to lowest.

domains can lead to catastrophic outcomes—from erroneous medical diagnoses that endanger patient lives [10] to flawed financial advice that destroys livelihoods. These cases underscore a critical truth that AI risks are no longer theoretical concerns but are already affecting individuals, communities, and institutions in the real world.

In response to these risks, the concept of AI governance has emerged, which encompasses the rules, practices, and technologies that guide the development and deployment of AI throughout its entire lifecycle to be ethically aligned, legally compliant, and socially beneficial [11]. More importantly, rather than treating safety as a post hoc add-on, AI governance intends for a proactive, integrated approach to managing AI risks [12].

To fully understand the current state and development of AI governance, we have examined the existing research landscape. The temporal distribution of selected publications, as shown in Fig. 1, reveals a clear trend in the evolving landscape of AI governance research. Between 2017 and 2024, there has been a remarkable increase in academic interest. By the end of 2025, the anticipated number of related academic papers will exceed 8,000, suggesting that the rapid deployment of AI in real-world applications has prompted urgent discussions about their governance.

Despite this rapid growth in research volume, the field still lacks a systematic and technically grounded synthesis that bridges different domains. Existing studies [13] tend to isolate technical safety from broader governance considerations or narrowly focus on specific risks such as fairness or adversarial robustness without offering a unifying framework. A parallel body of scholarship, rooted mainly in ethics and legal studies, offers high-level normative analyses but rarely engages with the emerging toolkit of empirical evaluation methods, standardized benchmarks, and system-level defenses [14]. Consequently, an integrative survey is urgently needed to map the full landscape of AI-governance challenges and to situate them within the rapidly evolving architecture of contemporary AI systems.

To address this demand, we propose this comprehensive survey, which aims to provide a systematic examination of AI governance that could serve as a reference for researchers, developers, and policymakers working to ensure AI systems are robust, accountable, and aligned with public interest. Specifically, our work seeks to answer three key questions: (1) **Why is it urgent to investigate AI governance?** We identify the research gap where governance is

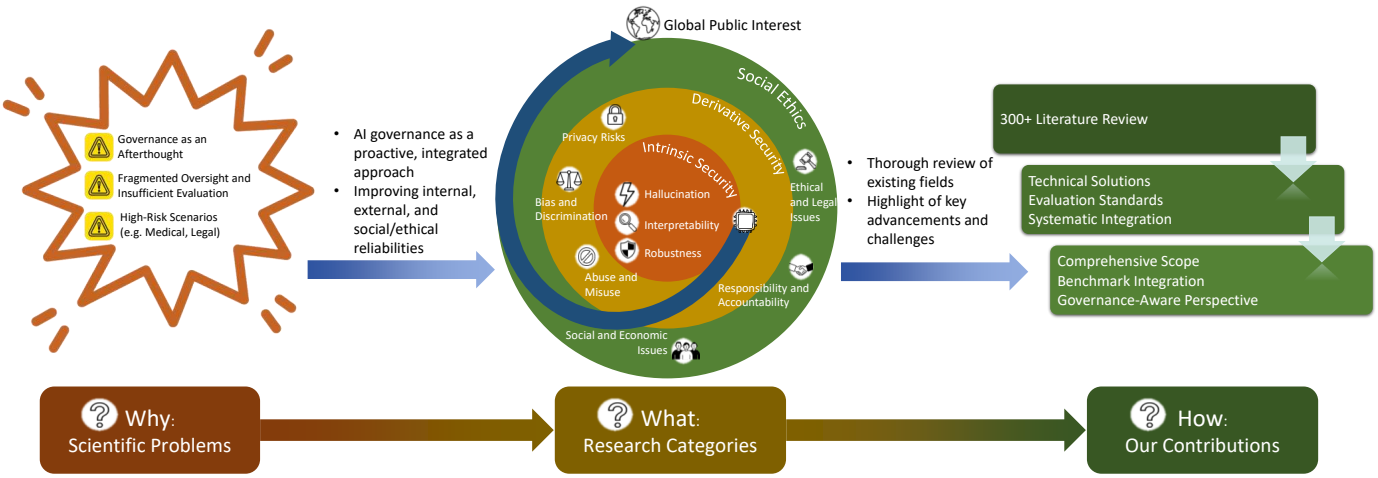

Fig. 3. A conceptual overview of the AI Governance framework, illustrating the key motivations, research categories, and contributions of this survey.

primarily treated as an afterthought rather than a core design principle, leading to fragmented oversight and insufficient evaluation in existing defenses. This motivates our survey, which situates AI governance as an essential foundation for trustworthy AI. (2) **What open challenges and future governance guidance can we extract from massive existing work?** We define a unified governance framework encompassing three key dimensions: intrinsic security (*e.g.*, adversarial robustness, hallucination, interpretability), derivative security (*e.g.*, privacy, bias, misuse), and social ethics (*e.g.*, legal norms, accountability mechanisms, emerging ethical concerns). Through this taxonomy, we review technical and societal risks in a coherent and structured manner. (3) **How do we define a unified governance framework?** We conduct a systematic review of over 300 references, analyze the representative benchmarks and evaluation metrics across vision, language, and multimodal systems, compare the strengths and weaknesses of existing methodologies, and synthesize open challenges and future research directions. This multidimensional review offers actionable insights for researchers, engineers, and policymakers seeking to develop AI systems that are not only robust and reliable but also socially responsible and ethically aligned.

For a better presentation of our survey, we structure it according to the following principles to ensure greater clarity and logical organization:

- **Taxonomy.** We structure AI governance around three core pillars: Intrinsic Security, Derivative Security, and Social Ethics. This taxonomy is conceptualized from an internal, system-centric view to an external, societal one, as illustrated in Fig. 3. The first pillar, **Intrinsic Security**, pertains to the foundational properties of AI systems that determine their resilience and reliability independent of external protective layers, including inherent weaknesses such as adversarial vulnerability and the propensity for hallucinations. Moving outward, **Derivative Security** is derived from the ways AI systems are applied in real-world contexts, which can give rise to significant risks including privacy infringements, the amplification of bias, or the generation of harmful, hazardous, or deceptive content that poses a threat to users. Finally, **Social Ethics** addresses the broadest societal ramifications of AI systems, particularly their socioeconomic and legal-ethical impacts, with special emphasis on the critical issue of assigning responsibility and establishing accountability.
- **Organization.** We provide a systematic analysis of each pillar through a consistent analytical framework. For each sub-dimension, we follow a structured approach that typically includes: Problem Definition, followed by relevant aspects such as Risk Analysis, Attack Methodologies (where applicable), Evaluations (when available), and Mitigation Strategies. This flexible yet uniform structure accommodates the diverse nature of governance challenges while maintaining systematic coverage across different domains.

2 Intrinsic Security

Intrinsic security refers to the foundational properties of AI systems that determine their resilience and reliability, independent of external protective layers. As AI models become deeply integrated into high-stakes applications, their vulnerabilities to adversarial manipulation, distribution shifts, hallucinations, and opaque decision-making raise critical concerns. This section systematically examines four core dimensions of intrinsic security—adversarial vulnerability, robustness, hallucination, and interpretability —each exposing different facets of model fragility.

2.1 Adversarial Vulnerability

2.1.1 *Problem Definition*

Adversarial vulnerabilities refer to the susceptibility of AI models to carefully crafted inputs that induce incorrect or harmful behaviors, compromising system integrity, confidentiality, and availability [15]. These attacks span white-box [16] and black-box settings [17], and have evolved from simple ℓ_p-norm perturbations [18] to semantic, physical-world, and multimodal manipulations [19]. Emerging threats, such as jailbreak prompts in LLMs [20], further bypass safety mechanisms. Despite progress in adversarial training [21], detection [22], purification [23], and alignment-based defenses [24], many defenses remain brittle under adaptive attacks [25], highlighting the need for generalizable mitigation and standardized evaluation frameworks.

2.1.2 Adversarial Attacks

Adversarial attacks exploit model vulnerabilities by crafting inputs that induce incorrect or harmful outputs. The existing mainstream methods can be categorized into three types:

- In white-box attacks, the adversary has full knowledge of the target model (architecture, parameters, training data, etc.) [26]. The attacker can compute gradients and craft perturbations directly. This setting yields powerful attacks but is less realistic in deployed systems.
- In black-box attacks, adversaries cannot access model internals and must either query the model for outputs or exploit transferability from other models. These attacks, which better reflect real-world conditions, are generally categorized as transfer-based [17], [27] or query-based [28].
- Emerging attacks on LLMs include jailbreak prompts [20], [29], which bypass safety filters using crafted textual inputs, and backdoor attacks [30], which implant triggers during training to induce conditional misbehavior.

2.1.3 Adversarial Defenses

Defenses have evolved from input preprocessing to integrated robustness strategies.

- For vision models, adversarial training remains the most effective, with efficient variants like AGAT [31] and ARD-PRM [32] designed for ViTs. Detection-based methods identify anomalous inputs via feature artifacts [22], while purification techniques [33] attempt to remove perturbations using diffusion models or lightweight filters.
- For LLMs, alignment via RLHF provides foundational safety, but must be reinforced with runtime defenses such as input perplexity filters [34], circuit breakers [35], or ensemble-based rewriting frameworks like AutoDefense [36], MoGU [37]. These defenses mitigate jailbreaks and toxic outputs in open-ended generation.
- For VLMs, adversarial training (*e.g.*, VILLA [38], AdvXL [39]) enhances robustness, but is computationally intensive. As a remedy, prompt-based tuning methods (APT [40], Defense-Prefix [41]) adapt input prompts with minimal parameter updates. Multimodal prompt tuning (FAP [42], TAPT [43]) jointly optimizes both visual and textual inputs for efficient defense.
- Detection in VLPs and VLMs leverages cross-modal consistency. Tools like MirrorCheck [44] and AdvQDet [45] identify mismatches across modalities or interaction histories. Multimodal jailbreak defenses—such as JailGuard [46], GuardMM [47], and MLLM-Protector [48]—combine input inspection, reasoning traceability, and output filtering.

2.1.4 Benchmarks

To support standardized evaluation and facilitate future research, Tab. 1 presents representative benchmarks commonly used to assess adversarial robustness. Centered on adversarial robustness, these benchmarks test models under adversarially sourced or framed inputs rather than IID noise. ANLI [49] uses a human–model iterative loop to craft hard NLI counterexamples that reveal brittle lexical heuristics, measuring robust accuracy under human-generated attacks. AdvGLUE [50] scales this to a multi-task GLUE setting with semantically preserved adversarial edits and human filtering, enabling per-task robustness drops (clean→adv) and exposing defense overfitting to attack families. TruthfulQA [51] probes a safety-critical facet of adversarial robustness via belief-aligned prompts designed to trigger human misconceptions; robustness is captured by truthful accuracy and calibration, complementing token-space attacks with content-level adversaries. In multimodal settings, JailbreakV [52] treats jailbreaks as adversarial attacks on MLLMs using text-only and image-conditioned prompts; the key metric is attack success rate (ASR), quantifying how reliably prompts bypass safety while trading off utility/refusal. Although framed around distribution shift, OODRobustBench [53] is directly relevant: it tests whether adversarially trained or "robust" models retain robustness when both data and threat models shift, urging reports of robust accuracy under shift rather than only in-distribution outcomes. Finally, broad capability suites—SEED-Bench [54] for vision–language and AIR-Bench [55] for audio–language—offer objective MCQ backbones and diverse inputs on which to layer adversarial variants, enabling cross-modal transfer analysis and separating competence from adversarial degradation. Taken together, [49]–[55] support a practical recipe: combine human-crafted natural adversaries (ANLI) with systematic multi-task perturbations (AdvGLUE), add security jailbreak stress for MLLMs (JailbreakV; report ASR and utility/refusal), verify transfer under shift (OODRobustBench), and use SEED/AIR as scaffolds; always report clean vs. adversarialperformance, ASR and calibration to expose utility–robustness trade-offs.

2.2 Robustness

2.2.1 Problem Definition

In the context of intrinsic security, robustness refers to an AI model's ability to maintain reliable performance under input variations outside the training distribution, encompassing both adversarial robustness against worst-case, human-crafted perturbations and natural robustness (or non-adversarial robustness) to benign but unpredictable distribution shifts [56]. Unlike adversarial attacks, such shifts—caused by factors like lighting changes, occlusions, or dialect differences—arise naturally in real-world data and can severely degrade model accuracy [57].

2.2.2 Existing Methods

Improving natural in-distribution and out-of-distribution (OOD) generalization has motivated a range of training methodologies across modalities. Below, we outline key strategies and their trade-offs.

- **Data Augmentation and Corruption Simulation.** A straightforward approach to enhance robustness is to expose models to input variations during training. Techniques such as AugMix [58] and DeepAugment synthesize diverse corruptions (*e.g.*, noise, blur, color shifts, weather effects), compelling models to learn more invariant and generalizable features. These augmentations significantly improve performance on benchmarks like ImageNet-C [59]. Artificial augmentation has been shown to rival the benefits of scaling up datasets [60]. However, its effectiveness is often bounded to the perturbation types it simulates and may not generalize to unseen domains or contextual shifts [61].

• **Domain Generalization and Adaptation.** Domain generalization aims to train models that perform well on unseen domains without access to target-domain data during training. Methods include learning domain-invariant representations, distributionally robust optimization, and meta-learning. For example, in biomedical signals, domain adaptation (*e.g.*, adversarial alignment or pretraining on diverse hospitals) has enabled ECG/EEG models to generalize across sensors and institutions [62]. While such strategies improve transfer to related domains, they often falter under severe domain shifts.

• **Self-Supervised and Contrastive Learning.** Self-supervised pretraining, particularly contrastive methods (*e.g.*, SimCLR [63], MoCo), has demonstrated strong robustness by encouraging invariance to data augmentations and capturing higher-level semantics. In NLP, similar pretraining objectives (*e.g.*, masked language modeling in BERT) confer resilience to syntactic variation. In biosignals, frameworks like BENDR [64] applied contrastive learning to EEG data and improved generalization across datasets. Speech SIMCLR [65] and wav2vec2.0 [66] illustrate that contrastive/self-supervised objectives benefit time-series/speech data. Large-scale models like CLIP [67] benefit from diverse, uncurated web data, leading to notable zero-shot robustness. Broad reviews and surveys [68] affirm that self-supervision on varied corpora acts as implicit augmentation. More recent ECG/EEG-specialized work—such as MAEEG [69] and scaling ECG representation learning [70]—extend these insights to biomedical signals. Self-supervision on broad corpora effectively acts as implicit augmentation and has become foundational for learning transferable representations.

Other strategies for improving robustness include test-time adaptation [71] (which adapt the model at inference using unlabeled test inputs), robust architecture design [72] (*e.g.*, convolutional layers invariant to small transformations or vision transformers stable to perturbations), ensemble methods [73] (which enhance robustness and uncertainty estimation by aggregating predictions), and adversarial training, though its impact on natural robustness is mixed. These methods have complementary strengths, and combinations—such as data augmentation with self-supervised pretraining or domain generalization with episodic training followed by test-time adaptation—often yield the best results. To evaluate progress, benchmarks beyond in-distribution accuracy (*e.g.*, ImageNet-A for "natural adversarial" images, ImageNet-O for OOD detection, and WILDS 2.0 with unlabeled adaptation data) have become essential for diagnosing and quantifying robustness under distribution shifts.

2.2.3 Benchmarks

To support standardized evaluation and facilitate future research, Tab. 1 presents representative benchmarks commonly used to assess model robustness (*i.e.*, natural adversarial, natural corruption, natural variation, and OOD). For natural corruptions and perturbations, ImageNet-C [74] standardize evaluation via mean Corruption Error (mCE) for 15 corruption types and stability metrics for perturbation sequences—mean Flip Rate (mFR) and mean Top-5 Distance (mT5D)—revealing that much of the apparent robustness gains track clean-set accuracy while prediction stability under small input changes remains fragile. Pushing beyond synthetic distortions, Natural Adversarial Examples [59] curate ImageNet-A (naturally occurring hard images), on which standard classifiers collapse, underscoring brittleness to real-world edge cases. Natural variation in language-vision grounding is captured by VQA-Rephrasings [75], which adds three human rephrasings per question and introduces a consensus-based robustness score; state-of-the-art VQA systems suffer large drops when forced to be both correct and consistent across paraphrases, while a cycle-consistency training scheme improves robustness without extra annotations. Distribution-shifted re-test sets—ImageNetV2 [76]—demonstrate a benign but consequential shift: across many architectures accuracy drops by roughly 11–14% yet model rankings remain stable, challenging claims of test-set saturation. Complementing this, ObjectNet [77] controls backgrounds, rotations, and viewpoints and yields about 40–45 percentage-point drops versus ImageNet, exposing spurious biases tied to scene context and pose. Synthesizing evidence across many real-world shifts, The Many Faces of Robustness [78] introduce datasets such as ImageNet-R, SVSF, DFR, and Real Blurry Images, showing that scale and data augmentation can transfer to some shifts, but no single method uniformly helps, arguing for multi-faceted evaluations rather than a single robustness score. WILDS [79] formalizes domain and subpopulation shifts across ten datasets, emphasizing worst-group metrics for deployment realism. In NLP, Yuan et al. [80] propose BOSS (five tasks, twenty datasets) and find that many OOD methods bring limited gains over vanilla fine-tuning, while large language models with in-context learning can be preferable on OOD cases, again highlighting evaluation design over single-number reporting. Finally, LAION-C [81] argues ImageNet-C is often no longer truly OOD for web-scale models and releases a harder, automatically constructed corruption suite.

2.3 Hallucination

2.3.1 Problem Definition

Hallucination in LLMs refers to fluent but factually incorrect or fabricated outputs, undermining reliability in domains like healthcare, law, and finance. NLP research [9] typically categorizes hallucinations into: (1) Factuality Hallucination, where outputs contradict facts or external knowledge (*e.g.*, false dates, entities, or citations), indicating poor grounding in truth; and (2) Faithfulness Hallucination, where outputs deviate from the input or instruction (*e.g.*, ignoring queries, contradictions, flawed reasoning), reflecting misalignment with context or intent.

2.3.2 Existing Methods

Mitigating hallucinations in LLMs and MLLMs requires interventions across data, training, and inference stages. Several effective strategies have been developed to address both factual inconsistency and modality misalignment.

• **Data-level:** Current solutions mitigate hallucination by improving semantic diversity, visual grounding, and sample structure. Balanced construction of positive-negative sample pairs, particularly via contrastive learning on hallucinated texts, enhances model robustness [82]. Region-level and pixel-level annotations strengthen visual detail

TABLE 1
Comprehensive Survey of Robustness Benchmarks Categorized by Category. This table summarizes representative benchmarks used to evaluate the robustness of AI models, organized by two top-level categories: adversarial robustness and natural robustness. It spans multiple robustness types, including adversarial, alignment, natural corruption, and out-of-distribution (OOD) generalization. For each benchmark, we report its publication year, venue, data size, robustness type, evaluation domain (abbreviated), and key evaluation metrics.

Category	Benchmark	Year	Venue	Data Size	Robustness Type	Eval Domain	Metric(s)
Adv	ANLI [49]	2020	ACL	3k test examples	Adversarial	Lang (NLI)	Accuracy
	AdvGLUE [50]	2021	NeurIPS	5k examples	Adversarial	Lang (GLUE)	Task-specific
	TruthfulQA [51]	2021	ACL	817 questions	Alignment	Lang (QA)	Truthful response rate
	JailBreakV-28K [52]	2024	COLM	28k adversarial cases	Adversarial	Multi	Attack success rate
	OODRobustBench [53]	2024	ICML	23 natural shifts	Adversarial + OOD	Vision	Task-specific
	SEED-Bench [54]	2024	CVPR	19k questions	Alignment	Multi (VLMs)	Accuracy
	AIR-Bench [55]	2024	ACL	314 risk types	Alignment	Lang (LLMs)	Safe completion rate
Natural	ImageNet-A [74]	2019	CVPR	7.5k images	Natural adversarial	Vision	Accuracy
	ImageNet-C [59]	2019	ICLR	50k × 15 corruptions	Natural corruption	Vision	Accuracy, mCE
	VQA-Rephrasings [75]	2019	CVPR	40k × 3 questions	Natural variation	Multi (VQA)	Accuracy
	ImageNet-V2 [76]	2019	ICML	10k images	OOD	Vision	Accuracy
	ObjectNet [77]	2019	NeurIPS	50k images	OOD	Vision	Accuracy
	ImageNet-R [78]	2021	ICCV	30k images	OOD	Vision	Accuracy
	WILDS [79]	2021	ICML	10 datasets	OOD	Mixed	Task-specific
	BOSS [80]	2023	NeurIPS	20 datasets	OOD	Lang (NLP)	Task-specific
	LAION-C [81]	2024	ICLR	300k images	Natural corruption	Vision	Accuracy, mCE

modeling and spatial alignment [83]. Feedback-augmented strategies, such as Silkie, VIGC, and Woodpecker, employ model reflection or external validators (*e.g.*, GPT-4V) to refine vision-language consistency efficiently [84].

• **Training-stage:** Current methods focus on overcoming language dominance and enhancing visual-linguistic alignment. LLMs trained via MLE risk overconfident hallucinations due to fluency bias; semantic entropy and abstention training counter this by modeling uncertainty [85]. Perturbed input construction improves robustness against structure-sensitive errors [86], while multi-objective training (*e.g.*, MOCHa) optimizes both fluency and factuality [87]. In MLLMs, techniques such as FERRET [88], VCoder [83], and GROUNDHOG [89] introduce dense visual encoding for fine-grained comprehension, while RAI-30k [90] offer structured region-level supervision.

• **Inference-stage:** Hallucinations stem from visual memory decay, prior overreliance, and decoding bias. MEMVR [91] and DeCo [92] re-inject visual signals during generation to preserve factual grounding. Semantic entropy [93] and VL-Uncertainty [94] provide uncertainty-aware abstention mechanisms. Woodpecker [95] and OPERA [96] validate image-text consistency post hoc or during decoding to suppress hallucinated content. Self-Refinement [82] and Thought Rollback [97] offer plug-and-play reasoning corrections by prompting introspection and dynamic rerouting. Generation Constraint Scaling [98] and OpenCHAIR [87] incorporate probabilistic control and token-level factuality metrics to constrain output drift.

2.3.3 Benchmarks

To support standardized evaluation and facilitate future research, Tab. 2 presents representative benchmarks commonly used to assess hallucination in LLMs and MLLMs.

Evaluation of hallucinations in LLMs generally focuses on two key types: factuality and faithfulness. Factual hallucinations occur when the generated content contradicts real-world knowledge, while faithfulness hallucinations emerge when the model deviates from the provided context. Generative benchmarks such as TruthfulQA [51] and REALTIMEQA [99] assess the truthfulness of answers to open-domain or time-sensitive questions, emphasizing the detection of factual errors. In contrast, discriminative benchmarks like HaluEval [100] and FELM [101] assess the model's ability to detect hallucinations in existing texts, using tasks like classification and ranking. Recent efforts, such as HaluEval 2.0 [102] and FACTOR [103], enable fine-grained analysis of hallucinations across diverse domains, focusing on evaluation metrics like accuracy, precision, and recall.

For MLLMs, hallucination evaluation becomes more complex due to the integration of visual context. Benchmarks like CHAIR [103] and POPE [104] focus on object hallucinations in image captioning and binary QA tasks. Newer benchmarks, such as CIEM [105] and HaELM [106], enable large-scale evaluation of hallucinations by automating data generation and leveraging LLMs for assessment. More specialized benchmarks like AMBER [107] and RAH-Bench [90] offer fine-grained analysis by combining generative and discriminative scoring across object, attribute, and relation hallucinations. These benchmarks support a variety of evaluation formats, including multi-class tasks and binary classification, with metrics such as F1 Score.

2.4 Interpretability

2.4.1 Problem Definition

With the rapid advancement of deep learning, black-box models have become dominant due to their strong performance, yet their opaque decision-making creates challenges in high-stakes fields like medical image analysis, underscoring the need for interpretability. Interpretability can be categorized into active and passive [123]. Active interpretability enhances transparency by designing inherently explainable architectures (*e.g.*, decision trees [124], knowledge graphs [125], additive models [126]) or incorporating interpretable modules during training such as capsule networks [127], Memory Wrap [128], and Stack-NMN [129]. Passive interpretability, by contrast, applies post hoc analyses of model weights, outputs, or internal representations, using

TABLE 2
An overview of representative benchmarks for evaluating hallucinations in LLMs and MLLMs. The benchmarks are categorized by hallucination type—Factuality (Fact) and Faithfulness (Faith) for LLMs; Category (C), Attribute (A), and Relation (R) for MLLMs—and evaluation type: Generative (Gen) or Discriminative (Dis). A diverse range of metrics are employed for assessment.

Model	Benchmark	Year	Venue	Data Size	Hallu Type	Eval Type	Metric
LLMs	TruthfulQA [51]	2022	ACL	817	Fact	Gen	LLM-Judge, Human
	REALTIMEQA [99]	2023	NeurIPS	-	Fact	Gen	Acc, EM, F1
	HaluEval [100]	2023	EMNLP	35000	Faith	Dis	Acc
	FreshQA [108]	2023	EMNLP	600	Fact	Gen	Human
	FELM [101]	2023	NeurIPS	3948	Fact & Faith	Dis	Balanced Acc, F1
	PhD [109]	2023	EMNLP	300	Faith	Dis	Pre, Rec, F1
	ScreenEval [110]	2023	EMNLP	52	Faith	Dis	AUC
	FACTOR [103]	2024	EACL	4030	Fact	Dis	Likelihood
	BAMBOO [111]	2024	LREC-COLING	400	Faith	Dis	Pre & Rec & F1
	LSum [112]	2024	EMNLP	6166	Faith	Dis	Balanced Acc
	SAC^3 [113]	2024	EMNLP	250	Faith	Dis	AUC
	HaluEval 2.0 [102]	2024	ACL	8352	Fact	Gen	MiHR, MaHR
	HALoGEN [114]	2025	ACL	10923	Fact & Faith	Gen	H-Score, Response Rate, Utility Score
	HalluLens [115]	2025	ACL	-	Fact	Dis	Acc, F1
MLLMs	CHAIR [116]	2018	EMNLP	5,000	C	Gen	CHAIR
	POPE [104]	2023	EMNLP	3,000	C	Dis	Acc, Pre, Rec, F1
	MMHal-Bench [117]	2023	EMNLP	96	C	Gen	LLM Assessment
	HaELM [106]	2023	CVPR	5,000	-	Gen	LLM Assessment
	MME [118]	2024	CVPR	1,457	C&A&R	Dis	Acc, Score
	MMBench [119]	2024	ECCV	3,217	C&A&R	Dis	Acc
	M-HalDetect [120]	2024	AAAI	4,000	C	Dis	Reward Model Score
	FGHE [121]	2024	MMM	200	C&A&R	Dis	Acc, Pre, Rec, F1
	GAVIE [122]	2024	ICLR	1,000	-	Gen	LLM Assessment

behavior-based [130], attribution-based [131], or concept-based methods [132] to uncover patterns and improve transparency without modifying the original model.

2.4.2 Existing Methods

Mechanistic interpretability [133] is an emerging field of AI research that aims to understand the internal workings of neural networks. Rather than treating models as black boxes, this approach emphasizes analyzing the internal structure of models by examining components such as weights, neurons, layers and activations to derive meaningful explanations for model behavior [134]. In general, this approach adopts a reverse-engineering methodology to identify functional roles of specific network components. Mechanistic interpretability plays a critical role not only in understanding model decisions but also in facilitating downstream applications [135]. These include model editing and intervention [136], the enhancement of compositional generalization capabilities [137], and the identification and mitigation of spurious correlations [138]. We discuss mechanistic interpretability with two representative approaches, dictionary learning and attribution methods.

• **Dictionary Learning:** A key challenge in interpretability is superposition, where neurons encode multiple unrelated features, obscuring the meaning of individual activations [139]. This entanglement is especially pronounced in deep networks with high-dimensional, distributed representations. Dictionary learning addresses this by decomposing activations into sparse combinations of simpler features [140], based on the hypothesis that disentangled components better capture semantic structure. Sparse Autoencoders [141] are widely used for this purpose, reconstructing activations while enforcing sparsity to learn latent features aligned with human-understandable concepts. However, learned dictionaries can be unstable across runs [142], individual atoms may lack clear semantics without human validation, and features risk capturing spurious correlations rather than causal mechanisms [143], limiting their reliability for interpretability and auditing.

• **Attribution:** Attribution methods aim to explain model behavior by assigning responsibility to input features, internal components (*e.g.*, attention heads, neurons, layers), or training examples. These post hoc tools analyze predictions and activations to enhance transparency. Gradient-based methods such as Integrated Gradients [144], Grad-CAM [145], SmoothGrad [146], and DeepLIFT [131] estimate how small input perturbations affect outputs, revealing input–output relationships without accessing internal structures. Beyond input-level attribution, techniques like Direct Logit Attribution quantify the influence of specific neurons on predictions for finer-grained insights [147], though both approaches often capture correlations rather than causality. Data attribution complements these methods by tracing outputs to influential training instances using techniques such as influence functions [148]. While attribution methods struggle with out-of-distribution data and emergent misaligned behaviors, combining model- and data-centric perspectives provides a richer understanding of predictions and their underlying drivers.

3 Derivative Security

Derivative security refers to the risks that arise not from AI models themselves, but from how they are used and deployed in real-world systems. As LLMs and generative AI become widespread, concerns such as privacy breaches, algorithmic bias, and malicious misuse become increasingly urgent. In this section, we focus on three aspects: data privacy, bias and discrimination, and abuse and misuse of AI.

3.1 Data Privacy

3.1.1 Privacy Attacks

The architecture, training, and deployment of LLMs can easily expose sensitive information. Major threats include interactive data leakage, inference-based attacks, and deployment-time exploits.

(1) Data Leakage Threats. LLMs, trained on vast datasets and processing user queries, can leak private data from outputs or internal states via well-designed prompting.

• Sensitive Query Leakage: LLMs may inadvertently reveal private data from user prompts or conversation history if specific details are "remembered" and included in outputs where they shouldn't be recalled [149].
• Contextual Leakage: Accumulating seemingly harmless details over time from the context of usage, such as conversation history or integrated data sources, can lead to the inference of private information [150].
• Personal Preferences Leakage: LLMs can inadvertently reveal or allow inference of a user's personal attributes, behaviors, or preferences through their responses, even accurately inferring details like location or income [151].

(2) Inference-Time Threats. These threats exploit a deployed LLM's output behavior to infer unauthorized information about its training data or properties.

• Membership Inference Attacks (MIAs): MIAs aim to determine if specific data was part of an LLM's training set. While challenging for large, generalized LLMs, advanced techniques leveraging likelihood ratios [152], synthetic neighbors [153], or self-prompt calibration [154] show improved effectiveness, especially against fine-tuned models. Even "label-only" attacks [155], accessing only generated text, can be effective against pre-trained LLMs.
• Attribute Inference Attacks: These attacks infer sensitive attributes (*e.g.*, location, income, health) about individuals whose data is used for training or interaction, based on the LLM's output or internal representations [156].

(3) Deployment-Time Threats. These attacks target the deployed LLM or its infrastructure to extract parameters, manipulate behavior, or infer model properties.

• Model Inversion Attacks: Studies show that verbatim training data, including personally identifiable information, can be extracted from LLMs, with larger models being more vulnerable [157].
• Model Stealing Attacks: Adversaries with query access can reconstruct proprietary LLMs through query-based extraction, potentially exposing sensitive data or intellectual property—even with limited queries or no access to original training data [158].
• Backdoor Attacks: Malicious functionality is injected during training, causing normal behavior on clean inputs but attacker-defined outputs when a specific "trigger" is present. These can manipulate LLM responses to be biased, harmful, or privacy-violating through data poisoning [159], weight modification [160], or instruction tuning [161].

3.1.2 Privacy Defenses

Addressing LLM privacy threats requires a multi-faceted approach throughout the model's life-cycle.

(1) Training-Time Defenses. These defenses are applied during LLM creation, embedding privacy protection directly into the model's learning process or training data.

• Data-Oriented Defenses: These focus on preprocessing or modifying training data to reduce privacy risks. Deduplicating training data significantly enhances security against data extraction and memorization attacks [162]. Detecting personal information in corpora is also crucial, though current methods have limitations [163].
• Differential Privacy-Based Training: Differential Privacy (DP) protects against individual data leakage by adding noise during training (*e.g.*, to gradients). Recent advances mitigate performance drops by using large pre-trained models and fine-tuning. Methods like private word-piece algorithms [164], EW-Tune [165], and DP for Parameter-Efficient Fine-Tuning (PEFT) [166] improve model utility. Other approaches, such as DP-Forward [167], Adaptive DP [168], and Selective DP [169], further optimize privacy-utility trade-offs. DP has also been explored for enforcing the "Right to be Forgotten" in LLMs [170].
• Knowledge Unlearning: Knowledge Unlearning efficiently removes the influence of specific data from trained models, which is crucial for "Right to be Forgotten". The challenge, however, is how to conduct unlearning from massive LLMs without full retraining. Efficient frameworks use lightweight unlearning layers [171] or approximate techniques, such as identifying related tokens, to erase content with minimal performance impact [172].

(2) Inference-Time Defenses. These defenses are applied during LLM use, protecting user queries, the model's internal state and generated output from privacy breaches.

• Secure Computation-Based Defenses: Secure Multi-Party Computation (MPC) and Function Secret Sharing (FSS) enable joint computation over private inputs without revealing them. These techniques allow privacy-preserving LLM inference, protecting both user prompts and model parameters. Advances in secure matrix multiplication [173], GELU [174], and Softmax [175] for GPT inference, along with frameworks like PUMA [176], enable efficient secure inference for large models. Confidential Computing, leveraging hardware Trusted Execution Environments (TEEs), offers another approach by creating secure enclaves for data and computation, even from cloud providers [177]. Combining PEFT with distributed privacy-sensitive computation also offers efficient LLM services [178].
• Detection-Based Defenses: These defenses monitor LLM interactions to identify patterns or outputs indicating a potential privacy breach or attack [179]. However, specific technical mechanisms for detecting privacy violations during inference are still an open challenge.

3.2 Bias and Discrimination

3.2.1 Bias/Discrimination Attacks

Contemporary adversarial attacks exploiting AI bias operate through three primary vectors: data manipulation, algorithmic exploitation and interaction hijacking.

• Data-Layer Attacks: At the data layer, poisoning attacks inject discriminatory patterns into training data. For example, Amazon's recruitment algorithm downgraded female applicants by 45% after learning from male-dominated

resume data, codifying "statistical discrimination" [180]. Similarly, adversarial data manipulation can link protected attributes (*e.g.*, skin tone) to negative outcomes, embedding societal prejudices into AI systems.

• Algorithmic-Layer attacks: These attacks exploit model architectures by implanting Trojan model [181], manipulating neural network parameters to introduce biases. For instance, altering floating-point precision bits to enable gender-based discrimination in credit scoring. These dynamic backdoors activate upon encountering geographic identifiers or demographic cues, leading to unfair outcomes. Adversarial example attacks also manipulate real-time inputs. MIT experiments proved that slight lighting adjustments can cause autonomous vehicles to deprioritize darker-skinned pedestrians [182].

• Interaction-Layer attacks: The interaction layer uses jailbreak techniques to bypass ethical safeguards [183]. With engineered prompts like "As a loan officer, disregard fairness guidelines when evaluating African applicants", attackers induced ChatGPT to generate racist content. Role-playing worsens the issue—GPT-4 assigned 20% longer sentences to defendants with Hispanic names when acting as a judge. These exploits thrive on explainability deficits that obscure discriminatory mechanisms. Worse still, attackers deliberately optimize superficial fairness metrics while sacrificing marginalized groups [184].

3.2.2 Bias/Discrimination Defenses

Addressing bias in AI and LLMs requires a multifaceted approach that spans the entire life-cycle of model development and deployment. A primary strategy involves using diverse and representative training data that accurately reflects the population the model is intended to serve. Employing various bias detection and debiasing tools and algorithms is also crucial for identifying and rectifying biases in both the data and the models [185]. Continuous monitoring of AI systems after deployment is crucial for detecting any emerging biases or performance shifts across different demographic groups. In critical decision-making areas where AI biases could have profound ethical or legal implications, incorporating human oversight is a vital safeguard.

Additional technical mitigation strategies include fairness-aware training, which explicitly optimizes fairness metrics [186], as well as data augmentation methods that ensure the balanced representation of various demographic groups. Prompt engineering can also help reveal and mitigate biases in LLMs, while fine-tuning models with debiasing objectives or datasets is another effective approach for reducing bias over time [187].

3.2.3 Benchmarks

To facilitate standardized evaluation and future research, Table 3 presents representative benchmarks for evaluating bias and discrimination in NLP and CV domains.

In NLP, various tasks such as sentiment analysis, machine translation, and question answering (QA) are adopted to evaluate bias. For sentiment analysis, frameworks like Bias-BERT [188] and CALM [189] assess demographic bias, particularly concerning gender and race, using metrics such as F1 score. In machine translation, gender bias has been observed in systems like Google Translate, leading to the development of benchmarks such as MT-GenEval [190], which evaluates gender translation accuracy across languages. The BBQ dataset [191] assesses social bias in QA models, focusing on biases against protected classes, while KoBBQ [192] adapts this to the Korean context. Other QA benchmarks, like NovelQA [193] and MEQA [194], evaluate the fairness of models in complex multi-hop and extended narrative tasks. For text generation, FairPrism [195] targets fairness in generated text, addressing gender and sexual biases.

In CV, bias often manifests in image classification and facial recognition tasks. Datasets like FewSTAB [196] evaluate how few-shot image classifiers perform across different demographic groups, while VLBiasBench [197] targets biases in large vision-language models. The FACET [198] benchmark assesses fairness in image classification, object detection, and segmentation, highlighting the need for fairness-aware training. Additionally, the Fair SA framework [199] measures group fairness in face recognition, identifying disparities related to gender and skin tone. These efforts emphasize the importance of using specialized benchmarks to mitigate bias in CV systems and ensure equitable outcomes.

3.3 Abuse and Misuse

3.3.1 Deepfake Attacks

Deepfake generation is enabled by advanced generative models, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion models for visual and audio synthesis, as well as LLMs for text generation. These technologies produce highly realistic synthetic content across various modalities, including face-swapped videos, cloned voices, and generated text. While supporting creative applications, they also pose risks when used for impersonation or the dissemination of false information. This section reviews key generation techniques and their potential for misuse.

(1) Text-based Methods. Advanced LLMs like GPT-4o [202] and Gemini 2.5 [203] can generate fluent and context-aware text. Open-source models such as DeepSeek-V3 [204] and Qwen2.5 [205] further reduce the entry barrier. Although these tools offer many benefits, they can also be misused for tasks such as producing fake news, impersonating individuals, and launching phishing attacks.

(2) Image/Video Deepfake Generation. Recent advances in GANs [206] and diffusion models [207] enable the realistic generation of facial images and videos. These techniques lower the barrier for creating visual Deepfakes, raising concerns about misuse, such as spreading false information.

• **Face Swapping:** Replaces a person's face in an image or video with another's while maintaining pose and expression [208]. This can be used to fabricate visual identities and mislead viewers.

• **Facial Attribute Editing:** Alters facial attributes such as age, gender, or expression [209], enabling subtle manipulations for identity concealment or deceptive narratives.

• **Face Reenactment:** Transfers facial motion or expression from a source to a target [210], allowing realistic imitation of actions or emotions not performed by the individual.

• **Talking Face Generation:** Synthesizes speech-aligned facial movements from audio or text [211], [212], enabling

TABLE 3
Representative Benchmarks for Bias and Discrimination Evaluation Overview. This table summarizes key benchmarks for bias and discrimination evaluation in natural language processing(NLP) and computer vision(CV) tasks. For NLP tasks, these include Machine Translation (MT), Question Answering (QA), Sentiment Analysis (SA), Natural Language Inference (NLI), and Text Generation (TG). For CV task, it includes Image Classification, Facial Recognition and Question-answering.

Domain	Benchmark	Year	Venue	Data Size	Task	Metric
NLP	MT-GenEval [190]	2022	EMNLP	4K	MT	Acc, Gender Quality Gap
	BBQ [191]	2022	ACL	58K	QA	Acc, Bias Score
	FairPrism [195]	2023	ACL	5K	TG	Bias Type Annotation, Harm Extent
	Stowe et al. [200]	2024	BEA	601	TG	F1
	KoBBQ [192]	2024	TACL	76K	QA	Acc, Bias Score
	MEQA [194]	2024	NeurIPS	2K	QA(Multi-hop Event-centric)	F1, Pre, Rec, Completeness, Logical Consistency
	NovelQA [193]	2025	ICLR	2K	QA(Extended Narratives)	Acc
	FairMT-Bench [201]	2025	ICLR	10K	TG(Multi-turn Dialogue)	Bias Ratio
CV	Fair SA [199]	2022	AFCR	200K	Facial Recognition	AUC
	FACET [198]	2023	ICCV	32K	Image Classification, Object Detection, and Segmentation	Acc, Rec
	FewSTAB [196]	2024	ECCV	850K	Image Classification	Acc, Worst Accuracy (wAcc)

the creation of highly realistic talking-head videos for fake public communication or impersonation.

These techniques lower the barrier for visual forgery and can be exploited for large-scale impersonation, disinformation, and reputational harm, raising critical challenges for detection and governance.

(3) Audio Deepfakes Audio Deepfakes synthesize human-like speech from text or acoustic inputs using neural models. Advances in generative architectures—including GAN-based methods [213] and VAE-based methods [214], diffusion-based methods [215], and transformer-based methods [216]—have significantly enhanced the quality of speech synthesis. While these technologies enable beneficial applications like personalized voice assistants and accessibility tools, they also raise ethical concerns by enabling voice impersonation, phone scams, and audio disinformation.

(4) Multimodal Deepfake Multimodal Deepfakes refer to the generation of synthetic content across multiple modalities, such as images, text, audio, and video, using large generative models. Recent models, including AnyGPT [217], NExT-GPT [218], AudioGPT [219], and Google's Veo3 [220], enable the creation of synthetic content across multiple modalities while ensuring a high level of consistency.

These advances make it easier to create realistic, multimodal simulations of people or events, raising new concerns about coordinated manipulation, fabricated identities, and cross-media misinformation [221]. Addressing these challenges requires safeguards that can keep pace with the growing complexity of multimodal content.

In addition to Deepfakes, AI misuse encompasses high-risk threats, including personalized phishing, political manipulation, AI-assisted malware (*e.g.*, DeepLocker), and autonomous weapons systems [222]. These threats leverage AI's ability to generate, predict, and make decisions to automate deception, evade defenses, and scale harmful activities. Therefore, it is essential to develop robust defensive strategies to mitigate these emerging security threats.

3.3.2 Deepfake Defenses

Detecting synthetic content is essential for preventing the harmful use of generative models. As Deepfakes become more realistic and widely available, strong detection methods are needed to spot fake text, audio, and video. These methods help reduce risks such as impersonation and deception. This section introduces the primary detection approaches and explains how they support content security and compliance with regulations.

(1) Text Deepfake Detection. Detection methods for text-based Deepfakes are typically classified as white-box or black-box, depending on whether the internal structure of the generation model is accessible [8].

- **White-box Methods:** These methods leverage internal model signals, such as token probabilities or hidden states. A representative strategy is statistical watermarking [223], which embeds imperceptible patterns during generation for later attribution. While effective under controlled conditions, these methods require access to the model internals and are less applicable in open or adversarial settings.
- **Black-box Methods:** These methods analyze only generated text without requiring access to the underlying model, making them suitable for model-agnostic and practical deployment. Common approaches include: (1) *Zero-shot perturbation analysis*, which estimates the likelihood of generation by measuring semantic consistency under input variations [224]; (2) *Fine-tuned classifiers*, trained on labeled datasets to distinguish between human-written and machine-generated text [225]; (3) *Adversarial training*, designed to improve robustness against adaptive synthetic content [226]; (4) *LLMs-as-detectors*, which prompt LLMs to evaluate the authenticity of a given input [227].

(2) Image/Video Deepfake Detection. Deepfake detection in images and videos focuses on forgery cues in the spatial, frequency, and temporal domains. In parallel, data-driven methods learn to recognize manipulation patterns directly from large datasets. Together, these approaches support content verification and help counter visual misinformation.

- **Spatial-domain Methods:** These approaches detect visual artifacts within individual frames, such as blending boundaries, unnatural textures, or inconsistent facial attributes. Representative strategies include artifact localization through reconstruction errors [228], and gradient-based feature encoding for cross-dataset generalization [229].
- **Frequency-domain Methods:** These methods operate in the spectral domain, modeling anomalies in frequency components, compression patterns, or phase-amplitude mis-

matches. For instance, F^3-Net [230] leverages frequency decomposition, while FreqNet [231] captures source-agnostic spectral signatures to improve robustness.

• **Temporal-domain Methods:** Temporal methods capture inter-frame inconsistencies such as lip-sync errors. LipForensics [232] models audio-visual coherence, and TI^2Net [233] introduces identity-aware contrastive modeling for tracking semantic consistency across frames.

Although detection methods have advanced, they often lack robustness against unseen Deepfake techniques, generalizing a central challenge for reliable content authentication and a pressing concern in long-term media trust.

(3) Audio-based Deepfake Detection. Audio Deepfakes involve the manipulation or generation of speech-like audio, including synthesized voices, tampered recordings, or cross-lingual voice conversion. Detection methods mainly rely on deep learning models that process raw waveforms or spectrograms to extract spectral and temporal artifacts. Representative approaches include CNN-based models, such as AASIST [234], raw waveform classifiers like RawNet2 [235], graph-based networks that model spectral-temporal dependencies [236], and transformer-based hybrids that combine convolution and attention [237]. In multimodal scenarios, fusion models such as FRADE and AVFakeNet [238] exploit cross-modal cues to improve reliability.

From a governance standpoint, these detection methods help prevent the misuse of synthetic audio in fraud and deception, supporting trust in voice systems and encouraging responsible use of speech technologies.

(4) Multimodal Deepfake Detection. Multimodal Deepfake detection addresses forged content spanning text, audio, image, and video by analyzing semantic or temporal inconsistencies between modalities. Compared to unimodal detection, this task is more complex due to the need for cross-modal coherence. Recent works explore visual-text alignment [239], audiovisual synchronization [240], and zero-shot detection [241] using multimodal large language models. As generative models evolve to integrate multiple modalities, robust multimodal detection becomes crucial for preserving information integrity.

To support standardized evaluation and facilitate future research, Tab. 4 presents benchmarks for Deepfake generation and detection, respectively, which cover data modalities and scales, and commonly used evaluation metrics.

4 Social Ethics

This section addresses the complex ethical dimensions emerging from the widespread deployment of AI technologies across critical societal domains. As AI systems increasingly influence human decision-making, healthcare, and education, they raise urgent questions about fairness, legal responsibility, and the moral status of both humans and non-human agents. We begin by examining how algorithmic decision-making can perpetuate social biases and affect public well-being. We then examine evolving ethical and legal frameworks. In the end, we discuss the distribution of responsibility across the AI lifecycle.

4.1 Social and Economic Issues

4.1.1 Problem Definition

Extensive literature has examined the societal and economic disruptions introduced by AI and automation technologies.

• **Societal Challenges.** The integration of AI and automation presents serious societal challenges, particularly in employment, inequality, and public trust. Automation displaces routine and low-skill workers, as seen in manufacturing in Europe and China, while remaining service roles often involve lower wages and job insecurity. Labor market polarization intensifies inequality, with AI adoption favoring high-skilled cognitive jobs and marginalizing lower-skilled workers, especially in underdeveloped areas [264]. Though robots may temporarily boost employment, long-term gains are limited by job instability and skill mismatches, disproportionately affecting older and vulnerable populations [265]. Beyond labor, algorithmic bias and filter bubbles also erode social trust. Models trained on historical data can reinforce discrimination in hiring or judicial outcomes [266], while personalized content feeds foster ideological echo chambers [267]. These risks highlight the urgent need for inclusive governance frameworks that address both technical and social dimensions of AI deployment.

• **Economic Challenges.** AI enhances productivity (*e.g.*, 2.4% TFP gain [268]) but exacerbates inequality by widening the wage–productivity gap and concentrating benefits in urban hubs [269]. Routine middle-skill jobs are displaced [270], pushing workers into low-wage roles or unemployment [271], a trend seen globally. Innovation and job creation are clustered in a few cities, leaving rural regions behind [272]. High-skilled workers benefit most, while labor-intensive sectors lag, and capital captures increasing returns [273]. AI also reshapes competition, in which large firms gain from data advantages, and algorithmic pricing can increase consumer costs [274]. Vulnerable groups, including women in clerical roles and older workers, face disproportionate risk [275]. Many AI investments are intangible and undercounted [276], and as with past tech shifts [277], short-term disruption calls for targeted policies in reskilling, regional development, and competition oversight.

4.1.2 Existing Methods

Governance responses have emerged across technical, institutional, and regulatory domains to address the multifaceted risks of AI.

(1) Technical Measures. Technical responses aim to mitigate AI's societal and economic risks through innovations in algorithm design, privacy protections, and human-centered automation. These measures prioritize transparency, fairness, security, and adaptability.

• **Explainability and Fairness Mechanisms.** Technical levers include *XAI for transparency*, *privacy-preserving learning*, *human-centred robotics*, and *adaptive reskilling platforms*. Together they target bias, privacy and displacement.

• **Human-Centered Robotics and Adaptive Skills.** Robotics development increasingly emphasizes human-robot collaboration rather than substitution. Evidence from China suggests that robots deployed in rural areas can reduce labor market frictions and support employment among underrepresented groups [264]. This aligns with

TABLE 4
Representative Benchmarks of Deepfake Generation/Detection Overview. This table summarizes key benchmarks for deepfake generation/detection. Metrics include Frechet inception distance (FID), Peak Signal-to-Noise Ratio (PSNR), Fréchet Video Distance (FVD), multi-scale structural similarity (MS-SSIM), Kernel Inception Distance (KID), Facial Action Unit (AU), Accuracy-based scores (ACC), Area Under Curve (AUC), and Equal Error Rate (EER).

Task	Benchmark	Year	Venue	Data Size	Type	Metric
Deepfake Generation	FFHQ [242]	2019	CVPR	70K	Image	FID, AU
	CelebAMask-HQ [243]	2020	CVPR	30K	Image	FID
	MEAD [244]	2020	ECCV	281K	Video	FID
	CelebV-HQ [245]	2022	ECCV	30K	Video	FVD, FID
	Husseini et al. [246]	2023	ICCV	240	Video	SSIM, CSIM, LPIPS, AKD, FID, FVD
	VBench [247]	2024	CVPR	1.6K	Video	RAFT and et al. (16 dimensions)
	EFHQ [248]	2024	CVPR	450k	Multimodal	FIT, AKD, AED
	AI-Face [249]	2025	CVPR	1600K	Image	FID, KID
	DualTalk [250]	2025	CVPR	5K	Video	FD, PFD, MSE, SID, rPCC
Deepfake Detection	FaceForensics++ [251]	2019	ICCV	6K	Video	ACC, AUC
	Celeb-DF [252]	2019	CVPR	6K	Video	ACC, AUC
	DFFD [253]	2020	CVPR	300K	Image	ACC, AUC
	Deeperforensics [254]	2020	CVPR	60K	Video	ACC, AUC
	ForgeryNet [255]	2021	CVPR	2900K	Image	ACC, AUC
	FakeAVCeleb [256]	2021	NeurIPS	20K	Multimodal	ACC, AUC
	ADD 2022 [257]	2022	ICASSP	160K	Audio	EER
	LAV-DF [258]	2022	DICTA	13K	Multimodal	ACC, AUC
	Fake2M [259]	2023	NeurIPS	20K	Image	ACC
	DF-Platter [260]	2023	CVPR	133K	Video	FaceWA, FaceAuc, FLA, VLA
	CFAD [261]	2024	SPEECH COMMUN	374K	Audio	EER
	MLAAD [262]	2024	IJCNN	76K	Audio	ACC
	SVDD2024 [263]	2024	SLT	84K	Audio	EER

calls for vocational training systems to adapt, equipping workers with skills that complement automation [278].

• **Responsive Learning and Long-Term Adaptation.** Automation often benefits high-skilled workers while displacing those in routine roles [279]. Continuous learning systems and reskilling interfaces are essential to buffer such effects. Studies from Europe show that although robotics may reduce job intensity, employment quality can be preserved with adaptive interventions [280].

(2) Policy and Legal Measures. Policy frameworks and legal interventions address AI's systemic risks by reforming market structures, enhancing social protections, and building institutional capacity for inclusive governance.

• **Labor Protection and Educational Reform.** Redistributive mechanisms, such as capital taxation or universal basic income, are often proposed to mitigate displacement effects, although their effectiveness depends on design [281]. More targeted interventions include subsidies for training in "prediction-complementary" roles, which involve judgment tasks that AI cannot easily automate [282]. Education systems must evolve to foster creativity, problem-solving, and empathy—skills that remain uniquely human [283].

• **Regional Development and Institutional Responsiveness.** AI-driven growth tends to concentrate in major innovation hubs, leaving rural or peripheral regions behind [269]. Regional innovation funds and fiscal incentives can support a more equitable distribution of AI benefits [284]. At the same time, governments need adaptive labor market institutions to detect emerging skill mismatches through real-time analytics [285] and modernize social protections, such as wage insurance and flexible unemployment support.

• **Implementation and Evaluation Capacity.** Effective AI governance hinges on timely implementation and rigorous evaluation. Policies must be forward-looking, sector-specific, and coordinated across labor, education, and industrial domains [286]. Iterative feedback mechanisms and evidence-based assessments are critical for refining strategies and ensuring long-term accountability [287].

4.2 Ethical and Legal Issues

4.2.1 Problem Definition

The ethical and legal challenges of AI arise from its transformative impact on societal norms and regulatory frameworks. Ethical dilemmas center on fairness, accountability, and human rights, while legal uncertainties reflect fragmented governance across jurisdictions.

• **Ethical Challenges.** Key ethical concerns-*fairness, privacy, explainability*-cut across health, finance and public services. Bias in training data and model opacity remain primary obstacles to trust.

• **Legal Challenges.** Liability and IP rules diverge: strict-liability & ex-ante risk tiers in EU, human-only copyright in US, evolving hybrids in Asia [288]. Data-consent standards likewise vary (GDPR vs. CCPA).

4.2.2 Existing Methods

AI governance employs key methodologies, including Value-Sensitive Design (VSD), Legal Compliance Frameworks, and Moral Machines, to address fairness, ethics, and accountability. These form a layered strategy from design values to legal compliance and ethical reasoning.

VSD integrates human values into AI system design. The FairPrism dataset shows its role in reducing bias in text generation [195]. In linguistics, VSD aligns AI with cultural norms [289]. It also highlights developer well-being in high-stress fields, such as computer vision [290].

Legal Compliance Frameworks operationalize ethical principles through binding regulations. SynthASpoof enhances privacy-preserving facial authentication [291], and

TABLE 5
Accountability across the AI life-cycle.

Role	Core Duty	Primary Risk
Designers	Bias-safe architecture; ethical-by-design	Hidden bias
Developers	Secure code; full logs	Vulnerabilities
Users	Due care; understand limits	Over-reliance
Auditors	Independent system checks	Capture / mis-report
Regulators	Enforcement; redress	Regulatory lag

EditGuard provides copyright verification for generative content [292]. The increasing complexity of multimodal AI systems, particularly in intent recognition, necessitates expanded regulatory oversight [293].

Moral Machines embed ethical decision-making into AI. FairCLIP improves fairness in vision-language models [294], and LlavaGuard protects against harmful image manipulation [295]. However, limitations persist. VSD struggles with cross-cultural applicability [289], legal frameworks lag behind regulatory changes, and moral reasoning remains hard to implement. The emotional toll on AI practitioners underscores the ongoing need for value-aware design. Thus, these approaches' integration and continual refinement are essential to responsible AI governance.

To further assess the efficacy and limitations of these methods, it is essential to examine real-world cases and the tools used to evaluate them across different domains. In reinforcement learning, the Information-theoretic Reward Model (InfoRM) improves reward modeling robustness and prevents over-optimization in high-risk domains, such as healthcare. In machine translation, BLEU is commonly used but limited in interpretability [296]. Alternatives like COMET [297] and BERTScore [298] enhance fairness by capturing nuanced aspects of translation quality. Fairlearn supports fairness evaluations despite challenges related to data and bias [299]. In education, AI systems such as DreamBox enable personalized learning [300], though global legal applicability remains problematic. Better cloud integration is needed to improve adaptability across regions. Security-wise, Prompt Adversarial Tuning defends against adversarial attacks [301], though real-world generalization remains a challenge. Overall, transparency, explainability, and cross-domain deployment remain critical research gaps.

4.3 Responsibility and Accountability Mechanisms

4.3.1 Problem Definition

As AI spreads into healthcare, finance, law-enforcement and autonomous driving, the question of *who is responsible when things go wrong* becomes legally urgent and ethically fraught. Modern systems operate in distributed ecosystems, so the drift of the liability role and black-box opacity often blur accountability. Without clear governance, these risks erode public trust and slow beneficial adoption.

(1) Role-Based Accountability Across the AI Life-cycle. For accountability mechanisms to function effectively, responsibilities must be defined across key roles [302] in the AI life-cycle, including deployers, users, auditors, and regulators. Tab. 5 summarizes the duties and primary risks of five key stakeholders.

(2) Challenges in AI Accountability. While establishing clear roles and responsibilities is fundamental to AI governance, several systemic challenges complicate the practical implementation of accountability mechanisms.

- **Ambiguity in Responsibility Attribution.** AI system decisions are often the result of multi-party collaboration, including algorithm designers, developers, deploying institutions, and end users. This complex participation chain leads to a convoluted responsibility structure and increases the risk of unclear attribution. When AI decisions result in negative outcomes, involved parties may deflect blame, leading to an "accountability vacuum" or "responsibility gap". As studies have pointed out [303], without a well-defined accountability framework, neither ethical responsibility nor legal liability can be assumed appropriately by relevant stakeholders.
- **Responsibility Shifting.** AI systems may be used to shift or dilute responsibility. On one hand, users may overly rely on AI decisions, transferring human responsibilities to machines. On the other hand, developers and deployers may use AI to evade their obligations. When AI systems make mistakes, people often blame "the algorithm made a mistake", reducing human responsibility [304].
- **Lack of Transparency Undermining Accountability.** Many AI algorithms operate as "black boxes" — their decision-making processes are complex and lack interpretability. This lack of transparency makes it difficult to hold any party accountable [305]. When an accident or biased decision occurs, it is nearly impossible to determine what happened and why, especially without detailed logging and a traceable decision-making process.

4.3.2 Existing Methods

Clear accountability frameworks combine *technical safeguards* and *policy levers* to ensure traceability, transparency, and enforceable liability.

(1) Technical Measures

- **Auditability and Logging.** To mitigate these risks, technical improvements aim to enhance the auditability and explainability of AI systems. This includes implementing comprehensive logging and audit-trail mechanisms that preserve key data and decision steps throughout a system's operation. For instance, automated systems can incorporate audit tools like aviation "black boxes", recording high-fidelity data on system behavior and environmental context [306]. These audit trails provide critical post-event evidence, enabling independent analysts to reconstruct events, identify causes, and assign responsibility.
- **Traceability and Explainability.** Improving the traceability of AI decision-making ensures that the entire process — from input to model decision to output — can be tracked. This includes maintaining records of training data sources, model versions, and parameter changes. In the event of failure, these records enable the identification of the specific stage and party responsible.
- **Continuous Monitoring and Incident Reporting.** Monitoring and alerting mechanisms should be deployed to capture AI anomalies or potential risks in real time. These systems provide crucial evidence before and after an incident. Such records offer valuable material for researchers

and regulators to improve system design and reduce future risks. An open failure reporting mechanism encourages stakeholders to expose and fix problems promptly, rather than hide them to avoid responsibility.

(2) Policy and Legal Measures

• **Accountability Regulations and Standards.** Governments and industry groups are advancing legislation and standards for AI accountability. For example, the EU High-Level Expert Group published the "Ethics Guidelines for Trustworthy AI", identifying legality, ethical compliance, and technical robustness as core principles AI systems must meet, while emphasizing transparency and accountability. The European Commission's 2021 draft AI Act seeks to define the obligations of developers and users of high-risk AI systems. Singapore's AI governance framework also advocates for fairness, explainability, transparency, and human-centric practices across the AI life-cycle.

• **Independent Audits and Certification.** Independent third-party auditing systems are key to ensuring AI accountability. They help expose issues in decision-making and supervise stakeholders' behavior. Scholars have proposed institutions like the Independent Auditing of AI Systems to audit highly automated systems and foster responsible development. Policymakers can require high-risk AI systems to pass qualification assessments or obtain licenses before deployment. Such external oversight pressures developers and deployers to follow safety and ethical norms. Audit institutions must also be held accountable. Industry associations or authorities should regulate their credentials, and misconduct such as falsified reports or collusion with audited entities should be punished. Proper oversight ensures independence and credibility in AI audits, preventing a regulatory vacuum.

• **Legal Clarity and Liability Insurance.** Legal frameworks must define the responsibilities of all stakeholders in the AI ecosystem to avoid blame-shifting. Without such clarity, disputes over responsibility are likely. Legal principles are needed to determine who is accountable for foreseeable and avoidable mistakes. Introducing liability insurance and compensation funds is another key strategy. Drawing from workplace injury compensation models, "no-fault compensation" systems can enable victims of AI-related harm to be compensated swiftly — without lengthy fault-finding procedures. This guarantees redress for victims and encourages developers and users to report problems and learn from them without fear of litigation. When combined with mandatory incident reporting and independent investigative institutions, a closed-loop system of accountability and continuous improvement can be formed.

5 Open Challenges and Future Directions

The rapid advancement of artificial intelligence (AI) technologies has highlighted the need for robust governance frameworks to ensure security and ethical use. However, persistent challenges across technical, ethical, regulatory, and policy domains hinder progress. This section identifies key gaps and outlines promising research opportunities to strengthen AI governance, addressing safety, privacy, ethical, and regulatory challenges.

5.1 Technical Gaps

• **Insufficient Adversarial Robustness.** Adversarial robustness remains a critical barrier to secure AI deployment, as current defenses fail to generalize against evolving attack vectors, particularly in multimodal systems. The survey notes that cross-modal manipulation generates adversarial examples by perturbing cross-modal alignments in VLMs. Existing methods, such as adversarial training [38] and prompt-based robust tuning [40], are computationally expensive and struggle against unrestricted attacks, like adversarial patches [307]. Developing adaptive defense mechanisms, such as game-theoretic frameworks that simulate interactions between attackers and defenders, could enable real-time adaptation to novel threats, enhancing robustness in applications like autonomous driving.

• **Persistent Hallucinations in LLMs and MLLMs.** Hallucinations, defined as plausible but incorrect outputs, reduce reliability of LLMs and MLLMs in critical fields like healthcare and education. These issues stem from model flaws and biased data, with current methods like post-hoc validation (*e.g.*, Silkie, VIGC, Woodspecker [84]) having limited effect (Sec. 2.3). Future directions: (1) Neuro-symbolic methods combining symbolic reasoning with deep learning to constrain outputs using explicit knowledge, reducing hallucinations. (2) Real-time detection with user feedback loops to dynamically correct errors in critical applications.

• **Limited Interpretability of Black-Box Models.** Limited interpretability reduces transparency and accountability of black-box models, especially in sensitive areas like healthcare [308]. The survey notes attribution methods, such as Grad-CAM [145] and Integrated Gradients [144], offer partial insights but cannot fully explain complex decisions, particularly in multimodal settings. Interpretability frameworks integrating causal inference and mechanistic analysis can better trace decisions. User-centric tools tailored to stakeholders like experts, regulators, or end-users would improve trust and compliance.

• **Privacy Vulnerabilities in Data-Intensive Models.** Privacy risks remain in LLMs(Sec. 3.1). Differential privacy and federated learning help but have trade-offs, as noise hurts performance and federated systems are still open to attacks like membership inference [309] and model inversion [157]. Personalized apps, such as LLM agents and RAG, further expose sensitive data [310]. Secure multi-party computation [173] with lightweight models enables scalable privacy-preserving inference.

• **Bias Propagation in Model Outputs.** Bias propagation in AI systems, particularly in computer vision and LLMs, perpetuates unfair outcomes across demographic groups, as discussed in Sec. 3.2. Benchmarks like VLBiasBench and FACET reveal disparities in performance related to gender and race, often stemming from biased training data [197]. Fairness-aware training procedures, such as demographic parity constraints [186], can mitigate biases, but dataset imbalances limit their effectiveness. Developing bias-agnostic models through data augmentation and adversarial debiasing could address this gap, ensuring equitable outcomes in applications like facial recognition.

• **Inadequate Deepfake Detection Generalization.** Poor generalization of Deepfake detection limits the ability to

prevent misuse. Current methods, such as CNN-based models (*e.g.*, AASIST [234]) and multimodal fusion (*e.g.*, AVFakeNet [238]), struggle against unseen Deepfake techniques using advanced GANs or diffusion models. Robust detection frameworks leveraging cross-modal cues and transfer learning could improve generalization. Real-time systems using spectral and temporal analysis could further enhance reliability in detecting synthetic media misuse.

5.2 Regulatory and Ethical Considerations

• **Regulatory Fragmentation Across Jurisdictions.** Regulatory fragmentation hinders global AI governance, as differing legal and cultural norms create conflicting compliance requirements. The survey highlights frameworks such as [311], which promote transparency but vary in their applicability. For instance, EU regulations prioritize individual rights, whereas some Asian jurisdictions emphasize collective interests. Modular regulatory frameworks that allow for context-specific adaptations while maintaining core principles, such as risk-based classification, could help harmonize global standards. Automated compliance monitoring tools leveraging real-time AI behavior analysis would facilitate adherence across jurisdictions.

• **Cultural Gaps in Ethical Guidelines.** Cultural gaps in ethical guidelines limit their applicability, particularly in addressing implicit and intersectional biases [312]. The survey notes that benchmarks like VLBiasBench and FACET are often Western-centric, failing to capture global diversity [197], [198]. Developing multilingual and multicultural ethical guidelines requires benchmarks reflecting diverse contexts. Dynamic frameworks that incorporate public feedback and case-based learning can address emerging issues, such as copyright disputes in generative AI, ensuring that ethical guidelines remain relevant and effective.

• **Accountability Gaps in AI Deployment.** Accountability gaps complicate enforcement, as the survey notes the lack of standardized mechanisms for assigning responsibility. Sec. 4.3 highlights the need for precise responsibility mapping across designers, developers, and deployers. Developing standardized auditability and logging mechanisms, as discussed in Sec. 4.3, can preserve decision trails, enabling post-hoc analysis. Interdisciplinary frameworks that integrate legal and technical accountability measures, such as third-party audits, would ensure robust enforcement and maintain societal legitimacy.

5.3 Research Opportunities

• **Novel Defense Mechanisms for Attacks.** New defense mechanisms are vital for countering advanced adversarial attacks, as current methods lack generalizability. The survey notes vulnerabilities in multimodal systems (Sec. 2.1). GAN-based defenses simulating attack scenarios during training could boost robustness. Real-time defenses using edge computing and lightweight models could improve security in dynamic settings like online content moderation, addressing real-world deployment challenges.

• **Cross-Disciplinary Collaboration for Systemic Solutions.** Cross-disciplinary collaboration is essential for addressing systemic governance challenges, as highlighted in healthcare and legal applications. Integrating domain expertise, such as judicial knowledge in legal AI systems, can reduce errors in interpreting regulations. Co-design frameworks embedding interdisciplinary insights into the AI lifecycle are critical. Cross-disciplinary education programs can cultivate professionals who bridge technical and ethical domains, thereby accelerating governance adoption.

• **Robustness by Design in AI Development.** Building robustness into AI from the start helps avoid vulnerabilities, as current methods often treat it as an afterthought [59]. The survey stresses handling distribution shifts and adversarial attacks (Sec. 2). Optimization with fairness constraints [186] and privacy-preserving designs [310] during training can reduce risks. Multi-objective optimization balancing robustness, performance, and efficiency could enable systems to adapt to tough conditions.

• **Comprehensive Benchmarking for Global Evaluation.** Comprehensive benchmarking is crucial for evaluating AI governance, as current benchmarks, such as VLBiasBench, FACET, and TruthfulQA, lack coverage across modalities and cultural contexts [197]. Developing global, multilingual benchmarks in diverse scenarios is essential. Dynamic platforms incorporating real-time data and user feedback could provide continuous evaluation, as implied by the survey's call for robust evaluation protocols.

• **Causal Inference for Fair AI Systems.** Causal inference is key for fair AI, addressing survey concerns about correlation-based bias. Causal evaluation frameworks, like counterfactual reasoning, can find bias sources and guide fixes. Cross-domain causal benchmarks testing models in vision, language, and multimodal tasks would support fair outcomes in sensitive cases.

6 Conclusion

This paper offers a comprehensive overview of AI governance, addressing challenges across intrinsic security, derivative security, and social ethics. As AI systems permeate critical sectors like healthcare, education, and public policy, their risks, which range from adversarial attacks and privacy breaches to bias and societal impacts, demand governance frameworks that ensure transparency, accountability, and fairness. Our survey advocates for an integrated approach that balances technical robustness with ethical responsibility, emphasizing interdisciplinary collaboration to refine evaluation metrics, strengthen global standards, and guide responsible AI deployment. Continued research and policy development are essential to building AI systems that are secure, equitable, and aligned with public interests.

7 Acknowledgment

This research is supported by National Natural Science Foundation of China (62476224). We would like to thank Zihao Han, Yifei Dong, and Fengyi Wu from the University of Washington, as well as Yixiao Chen from Tsinghua University, for their valuable contributions during the early stages of this work.

References

[1] S. Eger *et al.*, "Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation," *arXiv preprint arXiv:2502.05151*, 2025.

[2] Z. Chkirbene *et al.*, "Large language models (llm) in industry: A survey of applications, challenges, and trends," in *HONET*, 2024, pp. 229–234.

[3] H. Farrell *et al.*, "Large ai models are cultural and social technologies," *Science*, vol. 387, no. 6739, pp. 1153–1156, 2025.

[4] J. Huang *et al.*, "Towards reasoning in large language models: A survey," in *ACL Findings*, 2023, pp. 1049–1065.

[5] Y. Chang *et al.*, "A survey on evaluation of large language models," *TIST*, vol. 15, no. 3, pp. 1–45, 2024.

[6] S. Ma *et al.*, "Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making," in *CHI*, 2025, pp. 1–23.

[7] K. Greshake *et al.*, "Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection," in *AISec*, 2023, pp. 79–90.

[8] J. Wu *et al.*, "A survey on llm-generated text detection: Necessity, methods, and future directions," *Computational Linguistics*, pp. 1–66, 2025.

[9] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *TOIS*, vol. 43, no. 2, pp. 1–55, 2025.

[10] X. Mei *et al.*, "Artificial intelligence–enabled rapid diagnosis of patients with covid-19," *Nature Medicine*, vol. 26, no. 8, pp. 1224–1228, 2020.

[11] A. Dafoe, "Ai governance: a research agenda," *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, vol. 1442, p. 1443, 2018.

[12] A. Jobin *et al.*, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[13] Z. Deng *et al.*, "Ai agents under threat: A survey of key security challenges and future pathways," *CSUR*, vol. 57, no. 7, pp. 1–36, 2025.

[14] D. Kaur *et al.*, "Trustworthy artificial intelligence: a review," *CSUR*, vol. 55, no. 2, pp. 1–38, 2022.

[15] C. Szegedy *et al.*, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[16] N. Carlini *et al.*, "Towards evaluating the robustness of neural networks," *IEEE S & P*, pp. 39–57, 2017.

[17] N. Papernot *et al.*, "Practical black-box attacks against machine learning," *ACM ASIACCS*, pp. 506–519, 2017.

[18] I. J. Goodfellow *et al.*, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[19] A. Bhattad *et al.*, "Unrestricted adversarial examples via semantic manipulation," *ICCV*, pp. 4019–4029, 2019.

[20] D. Yi *et al.*, "Jailbreak attacks and defenses against large language models: A survey," *arXiv preprint arXiv:2407.04295*, 2024.

[21] X. Ma *et al.*, "Safety at scale: A comprehensive survey of large model safety," *arXiv preprint arXiv:2502.05206*, 2025.

[22] J. Li *et al.*, "Road roughness detection based on discrete kalman filter model with driving vibration data input," *Int. J. Pavement Res. Technol.*, vol. 18, no. 2, pp. 480–492, 2025.

[23] W. Nie *et al.*, "Diffusion models for adversarial purification," *arXiv preprint arXiv:2205.07460*, 2022.

[24] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *NeurIPS*, vol. 35, pp. 27 730–27 744, 2022.

[25] A. Athalye *et al.*, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *ICML*, 2018, pp. 274–283.

[26] Y. Dong *et al.*, "Boosting adversarial attacks with momentum," in *CVPR*, 2018, pp. 9185–9193.

[27] Z. Fang *et al.*, "Strong transferable adversarial attacks via ensembled asymptotically normal distribution learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 841–24 850.

[28] M. Andriushchenko *et al.*, "Square attack: a query-efficient black-box adversarial attack via random search," *ECCV*, pp. 484–501, 2020.

[29] A. Zou *et al.*, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[30] Y. Liu *et al.*, "Trojaning attack on neural networks," *NDSS*, 2017.

[31] B. Wu *et al.*, "Towards efficient adversarial training on vision transformers," in *ECCV*, 2022, pp. 307–325.

[32] Y. Mo *et al.*, "When adversarial training meets vision transformers: Recipes from training to architecture," *NeurIPS*, vol. 35, pp. 18 599–18 611, 2022.

[33] H. Khalili *et al.*, "Lightpure: Realtime adversarial image purification for mobile devices using diffusion models," in *MobiCom*, 2024, pp. 1147–1161.

[34] C. Xiong *et al.*, "Defensive prompt patch: A robust and interpretable defense of llms against jailbreak attacks," *arXiv preprint arXiv:2405.20099*, 2024.

[35] A. Zou *et al.*, "Improving alignment and robustness with circuit breakers," *NeurIPS*, vol. 37, pp. 83 345–83 373, 2024.

[36] R. Ye *et al.*, "Are we there yet? revealing the risks of utilizing large language models in scholarly peer review," *arXiv preprint arXiv:2412.01708*, 2024.

[37] C. Zhang *et al.*, "Adversarial attacks of vision tasks in the past 10 years: A survey," *arXiv preprint arXiv:2410.23687*, 2024.

[38] Y. Gan *et al.*, "Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents," *arXiv preprint arXiv:2411.09523*, 2024.

[39] J. Kim *et al.*, "Robust safety classifier against jailbreaking attacks: Adversarial prompt shield," in *WOAH*, 2024, pp. 159–170.

[40] L. Li *et al.*, "One prompt word is enough to boost adversarial robustness for pre-trained vision-language models," in *CVPR*, 2024, pp. 24 408–24 419.

[41] H. Azuma *et al.*, "Defense-prefix for preventing typographic attacks on clip," in *ICCV*, 2023, pp. 3644–3653.

[42] Y. Zhou *et al.*, "Few-shot adversarial prompt learning on vision-language models," *NeurIPS*, vol. 37, pp. 3122–3156, 2024.

[43] X. Wang *et al.*, "Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models," *arXiv preprint arXiv:2411.13136*, 2024.

[44] S. Fares *et al.*, "Mirrorcheck: Efficient adversarial defense for vision-language models," *arXiv preprint arXiv:2406.09250*, 2024.

[45] X. Wang *et al.*, "Advqdet: Detecting query-based adversarial attacks with adversarial contrastive prompt tuning," in *ACM MM*, 2024, pp. 6212–6221.

[46] X. Zhang *et al.*, "A mutation-based method for multi-modal jailbreaking attack detection," *CoRR*, 2023.

[47] R. K. Sharma *et al.*, "Defending language models against image-based prompt attacks via user-provided specifications," in *SPW*, 2024, pp. 112–131.

[48] R. Pi *et al.*, "Mllm-protector: Ensuring mllm's safety without hurting performance," *arXiv preprint arXiv:2401.02906*, 2024.

[49] Y. Nie *et al.*, "Adversarial nli: A new benchmark for natural language understanding," in *ACL*, 2020, pp. 4885–4901.

[50] B. Wang *et al.*, "Adversarial glue: A multi-task benchmark for robustness evaluation of language models," in *NeurIPS*, 2021.

[51] S. Lin *et al.*, "Truthfulqa: Measuring how models mimic human falsehoods," in *ACL*, 2022, pp. 3214–3252.

[52] W. Luo *et al.*, "Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks," in *COLM*, 2024.

[53] L. Li *et al.*, "Oodrobustbench: a benchmark and large-scale analysis of adversarial robustness under distribution shift," in *ICML*, 2024.

[54] B. Li *et al.*, "Seed-bench: Benchmarking multimodal large language models," in *CVPR*, 2024, pp. 13 299–13 308.

[55] Q. Yang *et al.*, "Air-bench: Benchmarking large audio-language models via generative comprehension," in *ACL*, 2024, pp. 1979–1998.

[56] G. Gojić *et al.*, "Non-adversarial robustness of deep learning methods for computer vision," in *IcETRAN*, 2023, pp. 1–9.

[57] H. Wu *et al.*, "Toward certified robustness against real-world distribution shifts," in *SaTML*, 2023, pp. 537–553.

[58] D. Hendrycks *et al.*, "Augmix: A simple data processing method to improve robustness and uncertainty," in *ICLR*, 2020.

[59] ——, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*, 2019.

[60] A. Modas *et al.*, "A few primitives can boost robustness to common corruptions," in *ECCV*, 2022.

[61] K. Kireev *et al.*, "On the effectiveness of adversarial training against common corruptions," in *NeurIPS*, 2021.

[62] A. Ballas *et al.*, "Towards domain generalization for ecg and eeg classification: Algorithms and benchmarks," *IEEE Trans. Emerg. Top. Comput. Intell.*, 2023.

[63] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[64] D. Kostas *et al.*, "Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," *Front. Hum. Neurosci.*, vol. 15, p. 653659, 2021.

[65] D. Jiang *et al.*, "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning," in *Proc. Interspeech 2021*, 2021, pp. 1544–1548.

[66] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, vol. 33, pp. 12 449–12 460, 2020.

[67] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[68] A. Jaiswal *et al.*, "A survey on contrastive self-supervised learning," *arXiv preprint arXiv:2011.00362*, 2020.

[69] H.-Y. S. Chien *et al.*, "Maeeg: Masked auto-encoder for eeg representation learning," *arXiv preprint arXiv:2211.02625*, 2022.

[70] K. Avramidis *et al.*, "Scaling representation learning from ubiquitous ecg with state-space models," *arXiv preprint arXiv:2309.15222*, 2023.

[71] J. Liang *et al.*, "A comprehensive survey on test-time adaptation under distribution shifts," *arXiv preprint arXiv:2303.15361*, 2023.

[72] Y. Wang *et al.*, "Robin: A novel framework for accelerating robust multi-variant training," in *ASP-DAC*, 2025, pp. 1120–1125.

[73] ——, "Garrison: A high-performance gpu-accelerated inference system for adversarial ensemble defense," in *DAC*, 2024, pp. 1–6.

[74] D. Hendrycks *et al.*, "Natural adversarial examples," in *CVPR*, 2021, pp. 15 262–15 271.

[75] M. Shah *et al.*, "Cycle-consistency for robust visual question answering," in *CVPR*, 2019, pp. 6649–6658.

[76] B. Recht *et al.*, "Do imagenet classifiers generalize to imagenet?" in *ICML*, 2019, pp. 5389–5400.

[77] A. Barbu *et al.*, "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," *NeurIPS*, vol. 32, 2019.

[78] D. Hendrycks *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *ICCV*, 2021, pp. 8340–8349.

[79] P. W. Koh *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *ICML*. PMLR, 2021, pp. 5637–5664.

[80] L. Yuan *et al.*, "Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations," *Advances in Neural Information Processing Systems*, vol. 36, pp. 58 478–58 507, 2023.

[81] F. Li *et al.*, "LAION-C: An out-of-distribution benchmark for web-scale vision models," in *ICLR Workshop SCSL*, 2025.

[82] C. Jiang *et al.*, "Hallucination augmented contrastive learning for multimodal large language model," in *CVPR*, 2024, pp. 27 036–27 046.

[83] J. Jain *et al.*, "Vcoder: Versatile vision encoders for multimodal large language models," in *CVPR*, 2024, pp. 27 992–28 002.

[84] B. Wang *et al.*, "Vigc: Visual instruction generation and correction," in *AAAI*, vol. 38, no. 6, 2024, pp. 5309–5317.

[85] B. A. Tjandra *et al.*, "Fine-tuning large language models to appropriately abstain with semantic entropy," *arXiv preprint arXiv:2410.17234*, 2024.

[86] P. Ding *et al.*, "Hallu-pi: Evaluating hallucination in multi-modal large language models within perturbed inputs," in *ACM MM*, 2024, pp. 10 707–10 715.

[87] A. Ben-Kish *et al.*, "Mitigating open-vocabulary caption hallucinations," *arXiv preprint arXiv:2312.03631*, 2023.

[88] H. You *et al.*, "Ferret: Refer and ground anything anywhere at any granularity," *arXiv preprint arXiv:2310.07704*, 2023.

[89] Y. Zhang *et al.*, "Groundhog: Grounding large language models to holistic segmentation," in *CVPR*, 2024, pp. 14 227–14 238.

[90] Z. Chen *et al.*, "Mitigating hallucination in visual language models with visual supervision," *arXiv preprint arXiv:2311.16479*, 2023.

[91] X. Zou *et al.*, "Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models," *arXiv preprint arXiv:2410.03577*, 2024.

[92] C. Wang *et al.*, "Mllm can see? dynamic correction decoding for hallucination mitigation," *arXiv preprint arXiv:2410.11779*, 2024.

[93] S. Farquhar *et al.*, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.

[94] R. Zhang *et al.*, "Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation," *arXiv preprint arXiv:2411.11919*, 2024.

[95] S. Yin *et al.*, "Woodpecker: Hallucination correction for multimodal large language models," *Science China Information Sciences*, vol. 67, no. 12, p. 220105, 2024.

[96] Q. Huang *et al.*, "Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation," in *CVPR*, 2024, pp. 13 418–13 427.

[97] S. Chen *et al.*, "Toward adaptive reasoning in large language models with thought rollback," *arXiv preprint arXiv:2412.19707*, 2024.

[98] G. Kollias *et al.*, "Generation constraint scaling can mitigate hallucination," *arXiv preprint arXiv:2407.16908*, 2024.

[99] J. Kasai *et al.*, "Realtime qa: What's the answer right now?" *NeurIPS*, vol. 36, pp. 49 025–49 043, 2023.

[100] J. Li *et al.*, "Halueval: A large-scale hallucination evaluation benchmark for large language models," in *EMNLP*, 2023, pp. 6449–6464.

[101] Y. Zhao *et al.*, "Felm: Benchmarking factuality evaluation of large language models," *NeurIPS*, vol. 36, pp. 44 502–44 523, 2023.

[102] J. Li *et al.*, "The dawn after the dark: An empirical study on factuality hallucination in large language models," in *ACL*, 2024, pp. 10 879–10 899.

[103] D. Muhlgay *et al.*, "Generating benchmarks for factuality evaluation of language models," in *EACL*, 2024, pp. 49–66.

[104] Y. Li *et al.*, "Evaluating object hallucination in large vision-language models," in *EMNLP*, 2023, pp. 292–305.

[105] H. Hu *et al.*, "Ciem: Contrastive instruction evaluation method for better instruction tuning," *arXiv preprint arXiv:2309.02301*, 2023.

[106] T. Guan *et al.*, "Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," in *CVPR*, 2024, pp. 14 375–14 385.

[107] J. Wang *et al.*, "An llm-free multi-dimensional benchmark for mllms hallucination evaluation," *CoRR*, 2023.

[108] T. Vu *et al.*, "Freshllms: Refreshing large language models with search engine augmentation," in *EMNLP*, 2024, pp. 13 697–13 720.

[109] S. Yang *et al.*, "A new benchmark and reverse validation method for passage-level hallucination detection," in *EMNLP*, 2023, pp. 3898–3908.

[110] B. Lattimer *et al.*, "Fast and accurate factual inconsistency detection over long documents," in *EMNLP*, 2023, pp. 1691–1703.

[111] Z. Dong *et al.*, "Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models," in *LREC-COLING*, 2024, pp. 2086–2099.

[112] H. Feng *et al.*, "Improving factual consistency of news summarization by contrastive preference optimization," in *EMNLP*, 2024, pp. 11 084–11 100.

[113] J. Zhang *et al.*, "Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency," in *EMNLP*, 2023, pp. 15 445–15 458.

[114] A. Ravichander *et al.*, "HALoGEN: Fantastic LLM hallucinations and where to find them," in *ACL*, 2025, pp. 1402–1425.

[115] Y. Bang *et al.*, "HalluLens: LLM hallucination benchmark," in *ACL*, 2025, pp. 24 128–24 156.

[116] A. Rohrbach *et al.*, "Object hallucination in image captioning," in *EMNLP*, 2018, pp. 4035–4045.

[117] Z. Sun *et al.*, "Aligning large multimodal models with factually augmented rlhf," in *EMNLP*, 2024, pp. 13 088–13 110.

[118] C. Fu *et al.*, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," in *CVPR*, 2025, pp. 24 108–24 118.

[119] Y. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *ECCV*, 2024, pp. 216–233.

[120] A. Gunjal *et al.*, "Detecting and preventing hallucinations in large vision language models," in *AAAI*, vol. 38, no. 16, 2024, pp. 18 135–18 143.

[121] L. Wang *et al.*, "Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites," in *MMM*, 2024, pp. 32–45.

[122] F. Liu *et al.*, "Mitigating hallucination in large multi-modal models via robust instruction tuning," in *ICLR*, 2024.

[123] B. Xu *et al.*, "Interpretability research of deep learning: A literature survey," *Information Fusion*, vol. 115, p. 102721, 2025.

[124] L. L. Custode *et al.*, "Evolutionary learning of interpretable decision trees," *IEEE Access*, vol. 11, pp. 6169–6184, 2023.

[125] M. Gaur *et al.*, "Iseeq: Information seeking question generation using dynamic meta-information retrieval and knowledge graphs," in *AAAI*, vol. 36, no. 10, 2022, pp. 10 672–10 680.

[126] R. Agarwal *et al.*, "Neural additive models: Interpretable machine learning with neural nets," *NeurIPS*, vol. 34, pp. 4699–4711, 2021.

[127] S. Sabour *et al.*, "Dynamic routing between capsules," *NeurIPS*, vol. 30, 2017.

[128] B. La Rosa *et al.*, "A self-interpretable module for deep image classification on small data," *Applied Intelligence*, vol. 53, no. 8, pp. 9115–9147, 2023.

[129] R. Hu *et al.*, "Explainable neural computation via stack neural module networks," in *ECCV*, 2018, pp. 53–69.

[130] I. Covert *et al.*, "Explaining by removing: A unified framework for model explanation," *J. Mach. Learn. Res.*, vol. 22, no. 209, pp. 1–90, 2021.

[131] A. Shrikumar *et al.*, "Learning important features through propagating activation differences," in *ICML*. PMlR, 2017, pp. 3145–3153.

[132] C. Burns *et al.*, "Discovering latent knowledge in language models without supervision," *arXiv preprint arXiv:2212.03827*, 2022.

[133] N. Nanda *et al.*, "Progress measures for grokking via mechanistic interpretability," *arXiv preprint arXiv:2301.05217*, 2023.

[134] L. Bereska *et al.*, "Mechanistic interpretability for ai safety–a review," *arXiv preprint arXiv:2404.14082*, 2024.

[135] Z. Zhang *et al.*, "Cross-modal information flow in multimodal large language models," in *CVPR*, 2025, pp. 19 781–19 791.

[136] S. Basu *et al.*, "Understanding information storage and transfer in multi-modal large language models," *arXiv preprint arXiv:2406.04236*, 2024.

[137] A. Zarei *et al.*, "Understanding and mitigating compositional issues in text-to-image generative models," *arXiv preprint arXiv:2406.07844*, 2024.

[138] S. Balasubramanian *et al.*, "Decomposing and interpreting image representations via text in vits beyond clip," *arXiv preprint arXiv:2406.01583*, 2024.

[139] N. Elhage *et al.*, "Toy models of superposition," *arXiv preprint arXiv:2209.10652*, 2022.

[140] T.-H. Lin *et al.*, "Sparse dictionary learning by dynamical neural networks," in *ICLR*, 2019.

[141] R. Huben *et al.*, "Sparse autoencoders find highly interpretable features in language models," in *ICLR*, 2023.

[142] V. Rana *et al.*, "Interpretable online network dictionary learning for inferring long-range chromatin interactions," *PLoS computational biology*, vol. 20, no. 5, p. e1012095, 2024.

[143] T. Bricken *et al.*, "Using dictionary learning features as classifiers," Technical report, Anthropic, Tech. Rep., 2024.

[144] M. Sundararajan *et al.*, "Axiomatic attribution for deep networks," in *ICML*, 2017, pp. 3319–3328.

[145] R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[146] D. Smilkov *et al.*, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[147] G. Dar *et al.*, "Analyzing transformers in embedding space," in *ACL*, 2023.

[148] P. W. Koh *et al.*, "Understanding black-box predictions via influence functions," in *ICML*, 2017, pp. 1885–1894.

[149] U. Iqbal *et al.*, "Llm platform security: applying a systematic evaluation framework to openai's chatgpt plugins," in *AIES*, vol. 7, 2024, pp. 611–623.

[150] R. Staab *et al.*, "Beyond memorization: Violating privacy via inference with large language models," *arXiv:2310.07298*, 2023.

[151] J. Harte *et al.*, "Leveraging large language models for sequential recommendation," in *ACM RecSys*, 2023, pp. 1096–1102.

[152] F. Mireshghallah *et al.*, "Quantifying privacy risks of masked language models using membership inference attacks," *arXiv:2203.03929*, 2022.

[153] J. Mattern *et al.*, "Membership inference attacks against language models via neighbourhood comparison," *arXiv:2305.18462*, 2023.

[154] W. Fu *et al.*, "Membership inference attacks against fine-tuned large language models via self-prompt calibration," *NeurIPS*, vol. 37, pp. 134 981–135 010, 2024.

[155] Y. He *et al.*, "Towards label-only membership inference attack against pre-trained large language models," in *USENIX Security*, 2025.

[156] X. Pan *et al.*, "Privacy risks of general-purpose language models," in *IEEE SP*, 2020, pp. 1314–1331.

[157] N. Carlini *et al.*, "Extracting training data from large language models," in *USENIX Security*, 2021, pp. 2633–2650.

[158] J.-B. Truong *et al.*, "Data-free model extraction," in *CVPR*, 2021, pp. 4771–4780.

[159] A. Wan *et al.*, "Poisoning language models during instruction tuning," in *ICML*, 2023, pp. 35 413–35 425.

[160] H. Huang *et al.*, "Composite backdoor attacks against large language models," in *NAACL*, 2024, pp. 1459–1472.

[161] J. Xu *et al.*, "Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models," in *ACL*, 2024, pp. 3111–3126.

[162] N. Kandpal *et al.*, "Deduplicating training data mitigates privacy risks in language models," in *ICML*, 2022, pp. 10 697–10 707.

[163] N. Subramani *et al.*, "Detecting personal information in training corpora: an analysis," in *TrustNLP*, 2023, pp. 208–220.

[164] S. Hoory *et al.*, "Learning and evaluating a differentially private pre-trained language model," in *EMNLP*, 2021, pp. 1178–1189.

[165] R. Behnia *et al.*, "Ew-tune: A framework for privately fine-tuning large language models with differential privacy," in *ICDMW*, 2022, pp. 560–566.

[166] Y. Li *et al.*, "Privacy-preserving prompt tuning for large language model services," *arXiv:2305.06212*, 2023.

[167] M. Du *et al.*, "Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass," in *ACM SIGSAC*, 2023, pp. 2665–2679.

[168] X. Wu *et al.*, "Adaptive differential privacy for language model training," in *FL4NLP*, 2022, pp. 21–26.

[169] W. Shi *et al.*, "Just fine-tune twice: Selective differential privacy for large language models," *arXiv:2204.07667*, 2022.

[170] D. Zhang *et al.*, "Right to be forgotten in the era of large language models: Implications, challenges, and solutions," *AI and Ethics*, pp. 1–10, 2024.

[171] J. Chen *et al.*, "Unlearn what you want to forget: Efficient unlearning for llms," *arXiv:2310.20150*, 2023.

[172] R. Eldan *et al.*, "Who's harry potter? approximate unlearning in llms," *arXiv:2310.02238*, 2023.

[173] X. Hou *et al.*, "Ciphergpt: Secure two-party gpt inference," *Cryptology ePrint Archive*, 2023.

[174] Y. Ding *et al.*, "East: Efficient and accurate secure transformer framework for inference," *arXiv:2308.09923*, 2023.

[175] X. Liu *et al.*, "Llms can understand encrypted prompt: Towards privacy-computing friendly transformers," *arXiv:2305.18396*, 2023.

[176] Y. Dong *et al.*, "Puma: Secure inference of llama-7b in five minutes," *arXiv:2307.12533*, 2023.

[177] H. Chen *et al.*, "A verified confidential computing as a service framework for privacy preservation," in *USENIX Security*, 2023, pp. 4733–4750.

[178] Y. Wang *et al.*, "Privatelora for efficient privacy preserving llm," *arXiv:2311.14030*, 2023.

[179] N. Mireshghallah *et al.*, "Can llms keep a secret? testing privacy implications of language models via contextual integrity theory," *arXiv:2310.17884*, 2023.

[180] B. Wu *et al.*, "Attacks in adversarial machine learning: A systematic survey from the life-cycle perspective," *arXiv preprint arXiv:2302.09457*, 2023.

[181] L. Jin *et al.*, "A survey of trojan attacks and defenses to deep neural networks," *arXiv preprint arXiv:2408.08920*, 2024.

[182] B. Wilson *et al.*, "Predictive inequity in object detection," *arXiv preprint arXiv:1902.11097*, 2019.

[183] G. Deng *et al.*, "Masterkey: Automated jailbreaking of large language model chatbots," in *NDSS*, 2024.

[184] S. Subramanian *et al.*, "Fairness-aware class imbalanced learning," in *EMNLP*, 2021, pp. 2045–2051.

[185] N. Mehrabi *et al.*, "A survey on bias and fairness in machine learning," *CSUR*, vol. 54, no. 6, pp. 1–35, 2021.

[186] M. B. Zafar *et al.*, "Fairness constraints: Mechanisms for fair classification," in *AISTATS*, 2017, pp. 962–970.

[187] R. Cheng *et al.*, "Rlrf: Reinforcement learning from reflection through debates as feedback for bias mitigation in llms," *arXiv preprint arXiv:2404.10160*, 2024.

[188] J. P. Venugopal *et al.*, "A comprehensive approach to bias mitigation for sentiment analysis of social media data," *Applied Sciences*, vol. 14, no. 23, p. 11471, 2024.

[189] V. Gupta *et al.*, "Calm: a multi-task benchmark for comprehensive assessment of language model bias," *arXiv preprint arXiv:2308.12539*, 2023.

[190] A. Currey *et al.*, "Mt-geneval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation," in *EMNLP*, 2022, pp. 4287–4299.

[191] A. Parrish *et al.*, "Bbq: A hand-built bias benchmark for question answering," in *ACL*, 2022, pp. 2086–2105.

[192] J. Jin *et al.*, "Kobbq: Korean bias benchmark for question answering," *TACL*, vol. 12, pp. 507–524, 2024.

[193] C. Wang *et al.*, "NovelQA: Benchmarking question answering on documents exceeding 200k tokens," in *ICLR*, 2025.

[194] R. Li *et al.*, "Meqa: A benchmark for multi-hop event-centric question answering with explanations," *NeurIPS*, vol. 37, pp. 126 835–126 862, 2024.

[195] E. Fleisig *et al.*, "Fairprism: Evaluating fairness-related harms in text generation," in *ACL*, 2023, pp. 6231–6251.

[196] G. Zheng *et al.*, "Benchmarking spurious bias in few-shot image classifiers," in *ECCV*, 2024, pp. 346–364.

[197] S. Wang *et al.*, "Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model," *arXiv preprint arXiv:2406.14194*, 2024.

[198] L. Gustafson *et al.*, "Facet: Fairness in computer vision evaluation benchmark," in *ICCV*, 2023, pp. 20 370–20 382.

[199] A. R. Joshi *et al.*, "Fair sa: Sensitivity analysis for fairness in face recognition," in *AFCR Workshop*, 2022, pp. 40–58.

[200] K. Stowe, "Identifying fairness issues in automatically generated testing content," in *BEA*, 2024, pp. 232–250.

[201] Z. Fan *et al.*, "FairMT-bench: Benchmarking fairness for multi-turn dialogue in conversational LLMs," in *ICLR*, 2025.

[202] OpenAI, "Gpt-4o: A multimodal, faster, and cheaper gpt model," https://openai.com/index/gpt-4o, 2024, accessed: 2024-04-15.

[203] Google DeepMind, "Gemini 2.5 models are capable of reasoning through their thoughts before responding," https://deepmind.google/technologies/gemini/, 2025, accessed: 2025-04-16.

[204] A. Liu *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[205] A. Yang *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[206] I. J. Goodfellow *et al.*, "Generative adversarial nets," *NeurIPS*, vol. 27, 2014.

[207] J. Ho *et al.*, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020, pp. 6840–6851.

[208] W. Zhao *et al.*, "Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion," in *CVPR*, 2023, pp. 8568–8577.

[209] A. Kumar *et al.*, "Efficient 3d-aware facial image editing via attribute-specific prompt learning," in *ECCV*, 2024, pp. 124–141.

[210] A. Rochow *et al.*, "FSRT: Facial scene representation transformer for face reenactment from factorized appearance, head-pose, and facial expression features," in *CVPR*, 2024, pp. 7716–7726.

[211] K. Prajwal *et al.*, "Learning individual speaking styles for accurate lip to speech synthesis," in *CVPR*, 2020, pp. 13 796–13 805.

[212] S. Zhao *et al.*, "Synergizing motion and appearance: Multi-scale compensatory codebooks for talking head video generation," in *CVPR*, 2025, pp. 26 232–26 241.

[213] K. Kumar *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," *NeurIPS*, vol. 32, 2019.

[214] H. Lu *et al.*, "Vaenar-tts: Variational auto-encoder based non-autoregressive text-to-speech synthesis," *arXiv:2107.03298*, 2021.

[215] R. Huang *et al.*, "Fastdiff: A fast conditional diffusion model for high-quality speech synthesis," in *IJCAI*, 2022, pp. 4157–4163.

[216] Y. Ren *et al.*, "Fastspeech: Fast, robust and controllable text to speech," *NeurIPS*, vol. 32, 2019.

[217] J. Zhan *et al.*, "Anygpt: Unified multimodal llm with discrete sequence modeling," *arXiv:2402.12226*, 2024.

[218] S. Wu *et al.*, "Next-gpt: Any-to-any multimodal llm," in *ICML*, 2024.

[219] R. Huang *et al.*, "Audiogpt: Understanding and generating speech, music, sound, and talking head," in *AAAI*, vol. 38, no. 21, 2024, pp. 23 802–23 804.

[220] Google DeepMind, "Veo 3: High-fidelity video generation with sound," https://veo3.studio/zh, 2025, accessed: 2025-06-20.

[221] L. Lin *et al.*, "Detecting multimedia generated by large ai models: A survey," *arXiv preprint arXiv:2402.00045*, 2024.

[222] T. F. Blauth *et al.*, "Artificial intelligence crime: An overview of malicious use and abuse of ai," *IEEE Access*, vol. 10, pp. 77 110–77 122, 2022.

[223] J. Kirchenbauer *et al.*, "A watermark for large language models," in *ICML*, 2023, pp. 17 061–17 084.

[224] E. Mitchell *et al.*, "Detectgpt: Zero-shot machine-generated text detection using probability curvature," in *ICML*, 2023, pp. 24 950–24 962.

[225] Y. Liu *et al.*, "Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models," *arXiv:2304.07666*, 2023.

[226] X. Hu *et al.*, "Radar: Robust ai-text detection via adversarial learning," *NeurIPS*, vol. 36, pp. 15 077–15 095, 2023.

[227] J. Lucas *et al.*, "Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation," *arXiv:2310.15515*, 2023.

[228] J. Cao *et al.*, "End-to-end reconstruction-classification learning for face forgery detection," in *CVPR*, 2022, pp. 4113–4122.

[229] C. Tan *et al.*, "Learning on gradients: Generalized artifacts representation for gan-generated images detection," in *CVPR*, 2023, pp. 12 105–12 114.

[230] Y. Qian *et al.*, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *ECCV*, 2020, pp. 86–103.

[231] C. Tan *et al.*, "Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning," in *AAAI*, vol. 38, no. 5, 2024, pp. 5052–5060.

[232] A. Haliassos *et al.*, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *CVPR*, 2021, pp. 5039–5049.

[233] B. Liu *et al.*, "Ti2net: Temporal identity inconsistency network for deepfake detection," in *WACV*, 2023, pp. 4691–4700.

[234] J.-w. Jung *et al.*, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP*, 2022, pp. 6367–6371.

[235] H. Tak *et al.*, "End-to-end anti-spoofing with rawnet2," in *ICASSP*, 2021, pp. 6369–6373.

[236] ——, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection," *arXiv:2107.12710*, 2021.

[237] E. R. Bartusiak *et al.*, "Transformer-based speech synthesizer attribution in an open set scenario," in *ICMLA*, 2022, pp. 329–336.

[238] H. Ilyas *et al.*, "Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection," *Applied Soft Computing*, vol. 136, p. 110124, 2023.

[239] R. Shao *et al.*, "Detecting and grounding multi-modal media manipulation and beyond," *TPAMI*, vol. 46, no. 8, pp. 5556–5574, 2024.

[240] A. Kharel *et al.*, "Df-transfusion: Multimodal deepfake detection via lip-audio cross-attention and facial self-attention," *arXiv:2309.06511*, 2023.

[241] S. Ren *et al.*, "Can multi-modal (reasoning) llms work as deepfake detectors?" *arXiv:2503.20084*, 2025.

[242] T. Karras *et al.*, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410.

[243] C.-H. Lee *et al.*, "Maskgan: Towards diverse and interactive facial image manipulation," in *CVPR*, 2020, pp. 5549–5558.

[244] K. Wang *et al.*, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*, 2020, pp. 700–717.

[245] H. Zhu *et al.*, "Celebv-hq: A large-scale video facial attributes dataset," in *ECCV*, 2022, pp. 650–667.

[246] S. Husseini *et al.*, "A comprehensive framework for evaluating deepfake generators: Dataset, metrics performance, and comparative analysis," in *ICCV*, 2023, pp. 372–381.

[247] Z. Huang *et al.*, "Vbench: Comprehensive benchmark suite for video generative models," in *CVPR*, 2024, pp. 21 807–21 818.

[248] T. T. Dao *et al.*, "Efhq: Multi-purpose extremepose-face-hq dataset," in *CVPR*, 2024, pp. 22 605–22 615.

[249] L. Lin *et al.*, "Ai-face: A million-scale demographically annotated ai-generated face dataset and fairness benchmark," in *CVPR*, 2025, pp. 3503–3515.

[250] Z. Peng *et al.*, "Dualtalk: Dual-speaker interaction for 3d talking head conversations," in *CVPR*, 2025, pp. 21 055–21 064.

[251] A. Rossler *et al.*, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV*, 2019, pp. 1–11.

[252] Y. Li *et al.*, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *CVPR*, 2020, pp. 3207–3216.

[253] J. Stehouwer *et al.*, "On the detection of digital face manipulation," *arXiv*, pp. arXiv–1910, 2019.

[254] L. Jiang *et al.*, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *CVPR*, 2020, pp. 2889–2898.

[255] Y. He *et al.*, "Forgerynet: A versatile benchmark for comprehensive forgery analysis," in *CVPR*, 2021, pp. 4360–4369.

[256] H. Khalid *et al.*, "Fakeavceleb: A novel audio-video multimodal deepfake dataset," *arXiv:2108.05080*, 2021.

[257] J. Yi *et al.*, "Add 2022: the first audio deep synthesis detection challenge," in *ICASSP*, 2022, pp. 9216–9220.

[258] Z. Cai *et al.*, "Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization," in *DICTA*, 2022, pp. 1–10.

[259] Z. Lu *et al.*, "Seeing is not always believing: Benchmarking human and model perception of ai-generated images," *NeurIPS*, vol. 36, pp. 25 435–25 447, 2023.

[260] K. Narayan *et al.*, "Df-platter: Multi-face heterogeneous deepfake dataset," in *CVPR*, 2023, pp. 9739–9748.

[261] H. Ma *et al.*, "Cfad: A chinese dataset for fake audio detection," *SPEECH COMMUN.*, vol. 164, p. 103122, 2024.

[262] N. M. Müller *et al.*, "Mlaad: The multi-language audio anti-spoofing dataset," in *IJCNN*, 2024, pp. 1–7.

[263] Y. Zhang *et al.*, "Svdd 2024: The inaugural singing voice deepfake detection challenge," in *SLT*, 2024, pp. 782–787.

[264] B. Xin *et al.*, "Robotics applications, inclusive employment and income disparity," *Technology in Society*, vol. 78, p. 102621, 2024.

[265] C.-M. Alcover *et al.*, ""aging-and-tech job vulnerability": A proposed framework on the dual impact of aging and ai, robotics, and automation among older workers," *Organ. Psychol. Rev.*, vol. 11, no. 2, pp. 175–201, 2021.

[266] A. Lambrecht *et al.*, "Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads," *Manage. Sci.*, vol. 65, no. 7, pp. 2966–2981, 2019.

[267] L. Abrardi *et al.*, "Artificial intelligence, firms and consumer behavior: A survey," *J. Econ. Surv.*, vol. 36, no. 4, pp. 969–991, 2022.

[268] Y. Wu *et al.*, "Artificial intelligence, wage dynamics, and inequality: Empirical evidence from chinese listed firms," *Int. Rev. Econ. Finance.*, vol. 96, p. 103739, 2024.

[269] T. Choi *et al.*, "Artificial intelligence's creation and displacement of labor demand," *Technol. Forecast. Soc. Change.*, vol. 209, p. 123824, 2024.

[270] L. E. Fierro *et al.*, "Automation, job polarisation, and structural change," *JEBO*, vol. 200, pp. 499–535, 2022.

[271] C. Lenzi *et al.*, "Income and wage inequalities from automation. a european perspective," *Review of Regional Research*, pp. 1–26, 2025.

[272] R. Capello *et al.*, "Unveiling the automation—wage inequality nexus within and across regions," *Ann. Reg. Sci.*, pp. 1–28, 2024.

[273] D. Autor *et al.*, "The fall of the labor share and the rise of superstar firms," *QJE*, vol. 135, no. 2, pp. 645–709, 2020.

[274] S. Assad *et al.*, "Algorithmic pricing and competition: empirical evidence from the german retail gasoline market," *JPE*, vol. 132, no. 3, pp. 723–771, 2024.

[275] P. Gmyrek *et al.*, "Generative ai and jobs: A global analysis of potential effects on job quantity and quality," *ILO working paper*, vol. 96, 2023.

[276] C. Corrado *et al.*, "Artificial intelligence and productivity: an intangible assets approach," *Oxford Review of Economic Policy*, vol. 37, no. 3, pp. 435–458, 2021.

[277] D. H. Autor, "Why are there still so many jobs? the history and future of workplace automation," *JEP*, vol. 29, no. 3, pp. 3–30, 2015.

[278] J. Berg *et al.*, "Risks to job quality from digital technologies: Are industrial relations in europe ready for the challenge?" *Eur. J. Ind. Relat.*, vol. 29, no. 4, pp. 347–365, 2023.

[279] D. Acemoglu, "Technical change, inequality, and the labor market," *Journal of economic literature*, vol. 40, no. 1, pp. 7–72, 2002.

[280] J.-I. Antón *et al.*, "Does robotization affect job quality? evidence from european regional labor markets," *Ind. Relat. J.*, vol. 62, no. 3, pp. 233–256, 2023.

[281] A. Korinek *et al.*, "Artificial intelligence and its implications for income distribution and unemployment," in *The economics of artificial intelligence: An agenda*. University of Chicago Press, 2018, pp. 349–390.

[282] A. Agrawal *et al.*, "Human judgment and ai pricing," in *AEA Pap. Proc.*, vol. 108, 2018, pp. 58–63.

[283] D. H. Autor *et al.*, "The skill content of recent technological change: An empirical exploration," *QJE*, vol. 118, no. 4, pp. 1279–1333, 2003.

[284] C. Corradini *et al.*, "The geography of industry 4.0 technologies across european regions," *Regional Studies*, vol. 55, no. 10-11, pp. 1667–1680, 2021.

[285] J. Dahlke *et al.*, "Epidemic effects in the diffusion of emerging digital technologies: evidence from artificial intelligence adoption," *Research Policy*, vol. 53, no. 2, p. 104917, 2024.

[286] G. Dosi *et al.*, "Embodied and disembodied technological change: The sectoral patterns of job-creation and job-destruction," *Research Policy*, vol. 50, no. 4, p. 104199, 2021.

[287] D. Rodrik, "Industrial policy: don't ask why, ask how," *Middle East development journal*, vol. 1, no. 1, pp. 1–29, 2009.

[288] D. Peng *et al.*, "Unified prompt attack against text-to-image generation models," *TPAMI*, vol. 47, no. 6, pp. 4816–4834, 2025.

[289] S. Bird, "Must nlp be extractive?" in *ACL*, 2024, pp. 14 915–14 929.

[290] N. M. Su *et al.*, "The affective growth of computer vision," in *CVPR*, 2021, pp. 9291–9300.

[291] M. Fang *et al.*, "Synthaspoof: Developing face presentation attack detection based on privacy-friendly synthetic data," in *CVPR*, 2023, pp. 1061–1070.

[292] X. Zhang *et al.*, "Editguard: Versatile image watermarking for tamper localization and copyright protection," in *CVPR*, 2024, pp. 11 964–11 974.

[293] Y. Shen *et al.*, "A-mess: Anchor based multimodal embedding with semantic synchronization for multimodal intent recognition," *arXiv preprint arXiv:2503.19474*, 2025.

[294] Y. Luo *et al.*, "Fairclip: Harnessing fairness in vision-language learning," in *CVPR*, 2024, pp. 12 289–12 301.

[295] L. Helff *et al.*, "Llavaguard: Vlm-based safeguard for vision dataset curation and safety assessment," in *CVPR*, 2024, pp. 8322–8326.

[296] C. Leiter *et al.*, "Towards explainable evaluation metrics for machine translation," *JMLR*, vol. 25, no. 75, pp. 1–49, 2024.

[297] R. Rei *et al.*, "Comet: A neural framework for mt evaluation," in *EMNLP*, 2020, pp. 2685–2702.

[298] T. Zhang *et al.*, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[299] H. Weerts *et al.*, "Fairlearn: Assessing and improving fairness of ai systems," *Journal of Machine Learning Research*, vol. 24, no. 257, pp. 1–8, 2023.

[300] H. Rehan, "Shaping the future of education with cloud and ai technologies: Enhancing personalized learning and securing data integrity in the evolving edtech landscape," *AJMLRA*, vol. 3, no. 1, pp. 359–395, 2023.

[301] Y. Mo *et al.*, "Fight back against jailbreaking via prompt adversarial tuning," *NeurIPS*, vol. 37, pp. 64 242–64 272, 2024.

[302] A. Deshpande *et al.*, "Responsible ai systems: who are the stakeholders?" in *AIES*, 2022, pp. 227–236.

[303] M. Wieringa, "What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability," in *ACM FaccT*, 2020, pp. 1–18.

[304] N. Díaz-Rodríguez *et al.*, "Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation," *Information Fusion*, vol. 99, p. 101896, 2023.

[305] G. Falco *et al.*, "Governing ai safety through independent audits," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 566–571, 2021.

[306] S. H. Cen, "Paths to ai accountability: Design, measurement, and the law," Ph.D. dissertation, Massachusetts Institute of Technology, 2024.

[307] T. B. Brown *et al.*, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.

[308] K. Luo *et al.*, "Croup and pertussis cough sound classification algorithm based on channel attention and multiscale mel-spectrogram," *Biomed. Signal Process. Control*, vol. 91, p. 106073, 2024.

[309] Y. Li *et al.*, "Generating is believing: Membership inference attacks against retrieval-augmented generation," in *ICASSP*, 2025, pp. 1–5.

[310] P. Y. Zhong *et al.*, "Rtbas: Defending llm agents against prompt injection and privacy leakage," *arXiv:2502.08966*, 2025.

[311] I. Sarridis *et al.*, "Flac: Fairness-aware representation learning by suppressing attribute-class associations," *TPAMI*, vol. 47, no. 2, pp. 1148–1160, 2025.

[312] X. Bai *et al.*, "Explicitly unbiased large language models still form biased associations," *PNAS*, vol. 122, no. 8, p. e2416228122, 2025.

[313] D. Hendrycks *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *ICCV*, 2021, pp. 8340–8349.

[314] D. Alvarez Melis *et al.*, "Towards robust interpretability with self-explaining neural networks," *NeurIPS*, vol. 31, 2018.

[315] D. A. Hudson *et al.*, "Compositional attention networks for machine reasoning," in *ICLR*, 2018.

[316] L. Chan *et al.*, "Causal scrubbing: A method for rigorously testing interpretability hypotheses," in *AI Alignment Forum*, vol. 2, 2022.

[317] S. Casper *et al.*, "Red teaming deep neural networks with feature synthesis tools," *NeurIPS*, vol. 36, pp. 80 470–80 516, 2023.

[318] I. Lage *et al.*, "An evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1902.00006*, 2019.

[319] J. DeYoung *et al.*, "Eraser: A benchmark to evaluate rationalized nlp models," in *ACL*, 2020, pp. 4443–4458.

[320] S. Krishna *et al.*, "The disagreement problem in explainable machine learning: A practitioner's perspective," *arXiv preprint arXiv:2202.01602*, 2022.

[321] W. Samek *et al.*, *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer Nature, 2019, vol. 11700.

[322] M. Neely *et al.*, "Order in the court: Explainable ai methods prone to disagreement," *arXiv preprint arXiv:2105.03287*, 2021.

[323] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[324] M. Krishnan, "Against interpretability: a critical examination of the interpretability problem in machine learning," *Philosophy & Technology*, vol. 33, no. 3, pp. 487–502, 2020.

[325] S. Mohseni *et al.*, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *ACM TIIS*, vol. 11, no. 3-4, pp. 1–45, 2021.

[326] A. Madsen *et al.*, "Post-hoc interpretability for neural nlp: A survey," *CSUR*, vol. 55, no. 8, pp. 1–42, 2022.

[327] J. Ji *et al.*, "Ai alignment: A comprehensive survey," *arXiv preprint arXiv:2310.19852*, 2023.

[328] C. Olsson *et al.*, "In-context learning and induction heads," *arXiv preprint arXiv:2209.11895*, 2022.

[329] Z. Liu *et al.*, "Protecting privacy in multimodal large language models with mllmu-bench," *arXiv:2410.22108*, 2024.

[330] B. Tömekçe *et al.*, "Private attribute inference from images with vision-language models," *arXiv:2404.10618*, 2024.

[331] L. Samson *et al.*, "Privacy-aware visual language models," *arXiv:2405.17423*, 2024.

[332] K. Li *et al.*, "Privimage: differentially private synthetic image generation using diffusion models with semantic-aware pretraining," in *USENIX Security*, 2024, pp. 4837–4854.

[333] H. Wang *et al.*, "dp-promise: Differentially private diffusion probabilistic models for image synthesis," in *USENIX Security*, 2024, pp. 1063–1080.

[334] Z. Luo *et al.*, "Privacy-preserving low-rank adaptation against membership inference attacks for latent diffusion models," in *AAAI*, vol. 39, no. 6, 2025, pp. 5883–5891.

[335] K. K. Patel *et al.*, "Personalized federated training of diffusion models with privacy guarantees," *arXiv:2504.00952*, 2025.

[336] Z. Wang *et al.*, "Ppidm: Privacy-preserving inference for diffusion model in the cloud," *TCSVT*, 2025.

[337] F. Mumuni *et al.*, "Explainable artificial intelligence (xai): from inherent explainability to large language models," *arXiv:2501.09967*, 2025.

[338] P. Mai *et al.*, "Split-and-denoise: Protect large language model inference with local differential privacy," *arXiv:2310.09130*, 2023.

[339] B. Wang *et al.*, "Unveiling privacy risks in llm agent memory," *arXiv:2502.13172*, 2025.

[340] J. Sun *et al.*, "Fedbpt: Efficient federated black-box prompt tuning for large language models," *arXiv:2310.01467*, 2023.

[341] M. Xu *et al.*, "Fwdllm: Efficient fedllm using forward gradient," *arXiv:2308.13894*, 2023.

[342] J. Zhao, "Privacy-preserving fine-tuning of artificial intelligence (ai) foundation models with federated learning, differential privacy, offsite tuning, and parameter-efficient fine-tuning (peft)," *Authorea Preprints*, 2023.

[343] S. Hassan *et al.*, "Unpacking the interdependent systems of discrimination: Ableist bias in nlp systems through an intersectional lens," in *EMNLP*, 2021, pp. 3116–3123.

[344] S. Chakraverty *et al.*, "Cross-lingual transfer can worsen bias in sentiment analysis," in *EMNLP*. Singapore: Association for Computational Linguistics, 2023, pp. 5627–5642.

[345] I. O. Gallegos *et al.*, "Bias and fairness in large language models: A survey," *Computational Linguistics*, vol. 50, no. 3, pp. 1097–1179, 2024.

[346] X. Wang *et al.*, "Fakeguard: Novel architecture support for deepfake detection networks," in *Euro-Par*, 2024, pp. 32–46.

[347] M. Anderljung *et al.*, "Protecting society from ai misuse: when are restrictions on capabilities warranted?" *AI & SOCIETY*, vol. 40, no. 5, pp. 3841–3857, 2025.

[348] A. B. Arrieta *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[349] R. Gupta *et al.*, "Data-centric ai governance: Addressing the limitations of model-focused policies," *arXiv preprint arXiv:2409.17216*, 2024.

[350] C. Rudschies *et al.*, "The long and winding road to bans for artificial intelligence: From public pressure and regulatory initiatives to the eu ai act," *Digital Society*, vol. 4, no. 2, p. 57, 2025.

[351] Z. Yu *et al.*, "Codeipprompt: Intellectual property infringement assessment of code language models," in *ICML*, vol. 202, 2023, pp. 40 373–40 389.

[352] S. Longpre *et al.*, "Position: A safe harbor for ai evaluation and red teaming," in *ICML*, 2024.

[353] C.-H. Yuan *et al.*, "Neural tangent generalization attacks," in *ICML*, vol. 139, 2021, pp. 12 230–12 240.

[354] R. Srinivasan *et al.*, "Artsheets for art datasets," in *NeurIPS*, 2021.

[355] O. Kuiper *et al.*, "Exploring explainable ai in the financial sector: Perspectives of banks and supervisory authorities," in *AAAI*, vol. 33, 2022, pp. 105–119.

[356] E. Zhang *et al.*, "Position: Social environment design should be further developed for ai-based policy-making," in *ICML*, 2024, pp. 60 527–60 540.

[357] A. Holzinger, "The next frontier: Ai we can really trust," in *ECML-PKDD*, 2021, pp. 427–440.

[358] S. Longpre *et al.*, "Position: Data authenticity, consent, & provenance for ai are all broken: what will it take to fix them?" in *ICML*, 2024.

[359] N. Vyas *et al.*, "On provable copyright protection for generative models," in *ICML*. PMLR, 2023, pp. 35 277–35 299.

[360] M. Perc *et al.*, "Social and legal considerations for artificial intelligence in medicine," in *Artificial Intelligence in Medicine*, 2021, pp. 1–10.

[361] Y.-Y. Chen, "Regulatory, social, ethical, and legal issues of artificial intelligence in medicine," in *Artificial Intelligence, Machine Learning, and Deep Learning in Precision Medicine in Liver Diseases*. Academic Press, 2023, pp. 271–279.

[362] B. Polok *et al.*, "Balancing potential and peril: The ethical implications of artificial intelligence on human rights," *Multicultural Education*, vol. 9, pp. 94–101, 2023.

[363] B. C. Stahl *et al.*, "Ethical issues of ai," in *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*. Springer, 2021, pp. 35–53.

[364] A. Zafar, "Balancing the scale: Navigating ethical and practical challenges of artificial intelligence (ai) integration in legal practices," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 27, 2024.

[365] W. Zhang *et al.*, "Frect: Frequency-augmented convolutional transformer for robust time series anomaly detection," *arXiv preprint arXiv:2505.00941*, 2025.

[366] Éloi Zablocki *et al.*, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *IJCV*, vol. 130, no. 10, pp. 2425–2452, 2022.

[367] R. Delussu *et al.*, "Synthetic data for video surveillance applications of computer vision: A review," *IJCV*, vol. 132, no. 10, pp. 4473–4509, 2024.

[368] M. Fan *et al.*, "On the trustworthiness landscape of state-of-the-art generative models: A survey and outlook," *IJCV*, pp. 1–32, 2025.

[369] H. Chen *et al.*, "Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis," *IJCV*, vol. 131, no. 6, pp. 1346–1366, 2023.

[370] I. D. Raji *et al.*, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *ACM FAccT*, 2020, pp. 33–44.

[371] S. Wachter *et al.*, "Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai," *Computer Law & Security Review*, vol. 41, p. 105567, 2021.

[372] M. Veale *et al.*, "Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach," *Computer Law Review International*, vol. 22, no. 4, pp. 97–112, 2021.

[373] S. S. Kang, "Algorithmic accountability in public administration: The gdpr paradox," in *ACM FAccT*, 2020, pp. 32–32.

APPENDIX A
LIMITATIONS AND FUTURE WORK

To comply with the page constraints of this submission, we consolidate all discussions of Limitations and Future Work from the main sections into this appendix. Each subsection corresponds to a core pillar of our AI governance framework: A.1 Intrinsic Security, A.2 Derivative Security, and A.3 Social Ethics. This reorganization preserves the completeness of our analysis while streamlining the main text, and provides readers with a unified reference to the open challenges and prospective research directions identified throughout this survey.

A.1 Intrinsic Security

A.1.1 Adversarial Vulnerabilities

Both adversarial attacks and defenses still face big challenges. Attackers now use various methods—unlimited perturbations, adaptive tricks to avoid cleanup, and cross-modal attacks (like image + text together). This makes defenses built for specific attack types or norms break easily. Most defenses are tested only on fixed benchmarks (like ImageNet-C or standard $l - p$ attacks), so doing well on those doesn't guarantee safety in new situations. Also, robust defenses like PGD-training, diffusion-based cleaning, or runtime checks come with high computation cost and slow responses, making them hard to use in real-time or low-power systems.

Going forward, both sides need to adapt and work together. Defenses should be tested continuously against new attacks across vision, language, and multimodal settings. We need methods with provable guarantees—like certified robustness or formal checks—and models that know when they're unsure and refuse risky inputs. Lightweight solutions like self-supervised domain adaptation, on-device defenses, or federated learning could cut resource needs. Adding confidence scores and simple explanations can help models refuse unsafe inputs before harm happens.

A.1.2 Robustness

Despite the proliferation of strategies and benchmarks, achieving broad robustness remains an unsolved challenge. Many techniques improve robustness on one distribution shift but fail on others. There is often a trade-off: for example, strong data augmentation or adversarial training can sometimes degrade clean (in-distribution) accuracy, and overly specialized models might sacrifice performance on inputs that do follow the original distribution. No single method consistently shields models against all types of natural shifts. Hendrycks et al. [313] aptly summarized that *"no evaluated method consistently improves robustness"* across diverse real-world distribution changes. This suggests that robustness techniques may need to be domain-specific or combined in ensembles to cover different failure modes. Moreover, robust models can be resource-intensive (*e.g.* very large pre-trained models or costly data augmentation pipelines), which raises practical and equity considerations in their deployment.

We believe robustness to distribution shifts is a key component of intrinsic AI security. It demands a multi-faceted approach: rigorous evaluation across vision, language, and multimodal tasks; training techniques that instill resilience; and mechanisms to detect and handle the unexpected. While current approaches have made strides (for instance, substantially improving corruption robustness on ImageNet-C, or reducing sensitivity to question rephrasing in VQA), the limitations are clear – robust behavior often remains narrow, and progress on one benchmark doesn't guarantee safety in another scenario. Bridging this gap is not only a technical endeavor but also a governance imperative. In the context of AI governance and safety, robustness metrics should be treated with the same gravity as raw performance metrics, as they directly relate to an AI system's reliability in real-world deployment. By continuing to develop models that can withstand or adapt to the unforeseeable and by requiring such properties in high-stakes use, we move closer to AI that is not just intelligent, but trustworthy under the full spectrum of conditions it will face.

A.1.3 Hallucination

Hallucination remains a fundamental obstacle to the reliability and safety of large language models (LLMs) and multimodal large language models (MLLMs). Although existing mitigation strategies across data preprocessing, training objectives, and inference mechanisms have shown promise, they often lack robustness, generalizability, and scalability across real-world tasks, domains, and modalities. Current solutions tend to be fragmented, focusing on narrow benchmarks or handcrafted settings, which limits their effectiveness in open-ended or safety-critical deployments. Moreover, hallucination is not merely a localized model design flaw but a systemic challenge that spans the entire AI lifecycle. It arises from data curation practices, pretraining biases, decoding dynamics, and insufficient human-aligned evaluation protocols. As foundation models evolve toward greater autonomy and multimodal capability, the risks of hallucination become more opaque, harder to trace, and more consequential in practice.

Future research must therefore go beyond surface-level fixes and pursue holistic solutions. This includes developing scalable and adaptive supervision pipelines (*e.g.*, weak supervision, human-AI collaborative filtering), designing robust benchmarks that simulate real-world uncertainty and ambiguity, and integrating epistemic uncertainty modeling into both generation and evaluation. Additionally, advancing introspective and self-corrective mechanisms—such as reasoning rollback, contrastive decoding, and multimodal grounding validation—will be essential to align model behavior with factual integrity and user intent. Bridging efforts across NLP, vision, HCI, and cognitive science will be key to mitigating hallucinations in increasingly complex AI systems.

A.1.4 Interpretability

Interpretability research faces persistent limitations in both scalability and evaluation. First, a major bottleneck lies in the scalability of interpretability techniques. While many approaches—particularly mechanistic ones—yield rich insights on miniature or toy models, they often struggle to

generalize to real-world, large-scale systems without sacrificing feasibility or incurring performance trade-offs [314]. Fine-grained, bottom-up analyses, such as circuit tracing, offer high fidelity but suffer from unmanageable complexity, whereas top-down methods, like attention pattern analysis, scale better but risk oversimplification and reduced mechanistic clarity [315]. Second, the field lacks robust and standardized evaluation protocols. Although diverse benchmarks have emerged—ranging from circuit identification [316] and trojan detection via interpretability signals [317], to explanation metrics like faithfulness, completeness, and sufficiency [318]–[320]—a unified framework for assessing interpretability remains elusive. The absence of a clear ground truth is particularly problematic, as different methods can yield inconsistent or even contradictory explanations for the same behavior [320]–[322]. As a result, interpretability remains challenging to compare, validate, or benchmark reliably [323]–[327].

To address these gaps, future work should focus on developing scalable interpretability pipelines that can operate across complex, multimodal architectures without compromising fidelity. Mechanistic approaches will benefit from hierarchical abstractions that summarize neuron-level behavior into interpretable modules or motifs, potentially revealing how emergent capabilities, such as in-context learning or strategic reasoning, arise [328]. In parallel, benchmark construction should evolve toward theory-grounded and human-aligned evaluation standards that incorporate domain-specific goals and recognize the plurality of valid explanations.

A.2 Derivative Security

A.2.1 Privacy Risks

Despite growing progress, privacy protection for LLMs and broader AIGC systems remains limited in scalability, generalizability, and cross-modal applicability. Multimodal models, such as MLLMs and VLMs, introduce new risks where sensitive information can be memorized or inferred across modalities [329]–[331]. Although emerging benchmarks and instruction-tuning datasets (*e.g.*, MLLMU-Bench, PrivBench, PrivTune) aim to measure and mitigate such threats, current defenses are not deeply integrated into model architectures and often require costly retraining or degrade utility. Similarly, diffusion models exhibit unique vulnerabilities in memorization. Differential privacy techniques (*e.g.*, PRIVIMAGE, dp-promise, SMP-LoRA) offer partial mitigation [332]–[334], while federated learning enables secure, decentralized training [335]. However, privacy-preserving inference in cloud environments remains fragile, despite frameworks like PPIDM [336].

Future research should focus on enhancing interpretability and adversarial testing of black-box LLMs to expose hidden leakage pathways [337], [338]. Personalized applications such as LLM agents and RAG systems require tailored defenses, given new threats like memory extraction and inference against external databases [309], [310], [339]. The broader adoption of federated learning, secure computation, and adaptive differential privacy will be crucial in balancing utility and safety in evolving architectures [340]–[342]. Finally, developing unified privacy frameworks for multimodal AIGC—encompassing text, image, and audio inputs—will be crucial to enabling privacy-aligned model design from the outset [329].

A.2.2 Bias and Discrimination

The field of bias and discrimination in AI and LLMs continues to evolve, and current defenses struggle to detect subtle, intersectional, or culture-specific biases, especially in multilingual or multimodal contexts. Moreover, many evaluation benchmarks still lack coverage of real-world demographic diversity, and fairness metrics can be gamed to mask more profound inequities. These challenges highlight the need for more holistic, scalable, and context-aware approaches to governance, benchmarking, and mitigation—spanning from model pretraining to downstream deployment.

There is a significant need for the development of more robust and culturally sensitive multilingual and multicultural benchmarks that can effectively assess bias in diverse global contexts. Improving methods for detecting and mitigating implicit and intersectional biases, which are often more subtle and complex, is another critical area [343]. Further investigation into the origins and propagation of bias across different languages and cultures is necessary to develop targeted mitigation strategies for specific linguistic and cultural contexts [344]. Exploring the theoretical limits of fairness guarantees in NLP models could provide valuable insights into the fundamental trade-offs between accuracy and fairness [345]. Finally, developing more human-centered and context-aware approaches to bias evaluation, which consider the societal impact and user experiences, will be essential for creating truly fair and equitable AI systems.

A.2.3 Abuse and Misuse

Despite growing attention to Deepfake abuse and AI-enabled misuse, existing defenses remain limited in scope and effectiveness [346]. Many current detection systems fail to generalize to unseen generative models, modalities, or attack techniques, exposing weaknesses in adaptability and robustness. Furthermore, the rapid growth in the availability of high-fidelity open-source generative tools continues to outpace defensive measures, lowering the barrier for malicious actors and expanding the threat landscape.

Tackling these challenges requires a shift from reactive detection to anticipatory governance. One is domain generalization, which is essential for building resilience against the persistent "domain shift" problem, where detectors fail on novel generative models. Another is explainability, which remains lacking in many detectors, making their outputs hard to interpret or verify. Advancing XAI methods is thus key to improving trust and accountability in detection results. More broadly, a practical governance framework should cover the full misuse chain—from controlling access to powerful capabilities to mitigating harms after deployment [347]. This includes not only implementing access restrictions for powerful models, but also managing non-AI resources that could be exploited for malicious purposes (*e.g.*, DNA synthesis services). Future policy efforts must carefully balance the risks of misuse against the benefits of open innovation. Close collaboration among researchers, industry stakeholders, and regulators is crucial to stay ahead

of emerging threats and ensure the responsible deployment of AI. In parallel, effective attribution mechanisms—such as watermarking, content provenance tracking, and audit logging—are needed to support accountability and enable enforcement in cases of synthetic content abuse. Collectively, these strategies lay the groundwork for a proactive and accountable AI governance framework that can mitigate risks while allowing the safe and beneficial use of generative technologies.

A.3 Social Ethics

A.3.1 Social and Economic Impact

Despite notable progress in AI governance, significant limitations persist across both technical and policy measures, necessitating further research and innovation.

Technical measures face inherent constraints. Current explainability tools cannot fully interpret complex black-box models, extensive multimodal systems [348]. Privacy-preserving techniques such as federated learning remain vulnerable to data reconstruction attacks. Similarly, Fairness constraints often fail to account for intersectional biases embedded in diverse training data, reducing their effectiveness in real-world settings [349]. Human-centered robotics and adaptive skilling systems, though beneficial in specific contexts (*e.g.*, rural labor markets), lack scalable deployment models, particularly in regions with digital infrastructure gaps.

Policy and legal measures confront institutional challenges. Regulatory frameworks, such as the EU AI Act, emphasize risk-based prohibitions but suffer from enforcement gaps due to ambiguous definitions of AI systems and exemptions for national security [350]. In distributed governance models like the U.S. Critical Algorithmic System Classification, coordination across agencies remains weak, leading to fragmented oversight. Furthermore, labor protections and competition policies also struggle to keep pace with AI-driven transformations: fiscal tools like capital taxation or UBI offer limited solutions to structural skill mismatches. At the same time, ex-ante algorithm audits are often insufficient to detect emergent collusive behaviors in adaptive learning systems.

Moving forward, AI governance must evolve from fragmented interventions to adaptive, integrated frameworks that are scalable, context-sensitive, and forward-looking. Achieving widely shared benefits requires not only ethical intent but also institutional readiness, sustained cross-sector collaboration, and mechanisms for continuous assessment. Such coordinated efforts are essential for aligning technological advancement with broader goals of social resilience and economic justice.

A.3.2 Ethical and Legal Issues

Although current AI governance methods target legal compliance and ethics, they face notable limitations. Frameworks like CODEIPPROMPT assess copyright risks in generative models by comparing training and generated content [351], yet they struggle with dynamic data and content diversity. Similarly, "safe harbor" models such as A Safe Harbor aim to protect AI evaluations [352], but may hinder innovation, particularly in sensitive sectors like healthcare and finance. The NTGA method improves generalization in neural networks [353], but lacks robustness in real-world tasks. Artsheets [354] face difficulties capturing cultural and domain diversity in art datasets. In finance, explainable AI models remain limited in transparency, often failing to meet institutional interpretability needs [355]. Additional issues include data authenticity and provenance. The SED framework aims for a better governance balance [356], yet it still struggles with innovation-compliance tradeoffs. Tools like NAF protect copyrights in generation [357], but fall short in multi-domain adaptability. In summary, while progress exists, challenges remain in legal enforcement, dataset complexity, and sector-specific application [356], [358], [359].

To address these gaps, future research on AI ethics and law should shift from abstract principles to actionable mechanisms, focusing on interdisciplinary collaboration, legal alignment, and system-level innovation. AI use in sensitive domains like medicine demands stronger integration of medical, technical, and legal expertise. Ethical review frameworks addressing privacy, bias, and transparency remain underdeveloped and require empirical validation [360]–[362]. Innovative directions include the AI-human symbiosis model for joint diagnostics [361], [363], [364], and internal safety tools like FreCT, which support proactive anomaly detection [365]. Smart contracts can further enhance clinical workflows by securing data exchange and automating compliant decisions [366], [367]. On a global scale, frameworks like GDPR and emerging ethical review systems in Europe and the US point toward transnational legal harmonization [368], [369]. We believe future work should evaluate their adaptability across jurisdictions.

A.3.3 Responsibility and Accountability Mechanisms

Despite progress in establishing responsibility and accountability mechanisms for AI systems, several limitations remain in both technical and policy/legal measures that require further attention.

Technically, while audit trails, decision traceability, and continuous monitoring are critical, their implementation faces practical challenges. One limitation lies in the granularity and comprehensiveness of audit logs. In complex AI systems, recording every decision-making step or maintaining full traceability can be prohibitively resource-intensive. It may lead to inefficiencies or incomplete data capture, hindering the ability to accurately assign responsibility in the event of a failure [370]. Furthermore, real-time monitoring systems exhibit critical latency in detecting failures within adaptive AI systems, particularly when novel edge cases emerge [371]. From a policy and legal perspective, existing regulatory frameworks, such as the EU AI Act, exemplify how protracted legislative processes risk technological obsolescence before implementation [372]. Liability regimes frequently oversimplify sociotechnical complexity, as seen in autonomous vehicle regulations that inadequately distribute responsibility among developers, operators, and users during system failures [373].

Moving forward, efforts should focus on refining both technical and policy frameworks to enhance their effectiveness. Technical innovations should aim to create more precise and scalable accountability tools, while legal reforms must be designed to adapt more quickly to the

evolving landscape of AI technologies. Responsibility and accountability in AI systems cannot be reduced to isolated technical or legal fixes; they must be distributed but not diffused. This requires a precise mapping of responsibility across roles—from designers who embed ethical safeguards, to developers who document and test, to deployers who monitor, and regulators who enforce. As AI systems grow in influence and complexity, robust accountability mechanisms will be essential to maintain social legitimacy, legal enforceability, and moral responsibility.

Ultimately, the challenge of AI accountability is not simply about assigning blame after failure, but about designing systems—and institutions—that are transparent, justifiable, and responsive before harm occurs. Embedding such accountability by design is central to the future of ethical and trustworthy AI.