Jurnal Inovasi dan Teknologi Pembelajaran
Vol 13, No 1, (2026), page 61-73
https://doi.org/10.17977/um031v13i12026p061
P-ISSN: 2406-8780 E-ISSN: 2654-7953
Open access: https://journal-fip.um.ac.id/index.php/jinotep

Trustworthy generative AI for computing systems: A review of safety, evaluation, and governance mechanisms

Saif Safaa Shakir[1], Hasan Fadhil Qasim[2], Huda Najim Abdulwahed[3]

[1]College of Computer Science and Information Technology, University of Al-Qadisiyah
University Street, Al Diwaniyah, Al-Qadisiyah, Iraq, email: saif.s.shaker@qu.edu.iq
[2]College of Agriculture , University of Misan
Rusafa Street 52. Alamara, Misan, Iraq, email: hasan.fadhil@uomisan.edu.iq
[3]College of Arts, University of Al-Qadisiyah
University Street, Al Diwaniyah, Al-Qadisiyah, Iraq, email: huda.najim@qu.edu.iq
Corresponding author, e-mail: saif.s.shaker@qu.edu.iq

ARTICLE INFO

Article history:
Received: 20-03-2026
Revised: 19-05-2026
Accepted: 20-05-2026

Keywords:
trustworthy AI; AI safety; alignment; robustness; evaluation benchmarks; AI governance; hallucination; generative AI; red-teaming; responsible AI.

ABSTRACT

More and more generative AI systems which includes large language models, diffusion models, and multimodal foundations are being integrated into crucial computing infrastructure, including cloud orchestration, code synthesis pipeline, healthcare decision support, and financial risk assessment. Consequently, there is greater demand for frameworks that can evaluate, guarantee, and regulate the trustworthiness of these systems. This article reviewed the development of trustworthy AI research from 2015 to 2025, and the evidence generated across four primary areas: safety and alignment, robustness and reliability, evaluation, and governance. We delivered distinctive comparative assessments of safety benchmarks, alignment methodologies (RLHF, RLAIF, DPO, Constitutional AI), and formal governance frameworks worldwide, pinpointing the critical discrepancies between regulated objectives and actual technical capability. A key finding is the Evaluation Paradox: The benchmarks most commonly relied on to certify systems as "AI safe" are, in fact, the systems least robust to distributional shift and adversarial manipulation. There is an institutional misalignment between the speed of generative AI deployment and the maturity of the governance mechanisms proposed to regulate it. We documented seven priority research challenges for the field. Researchers, system engineers, policymakers, and practitioners pursuing an evidence-based understanding of the current state-of-trustworthiness will benefit from this review.

INTRODUCTION

The system's trustworthiness has become an area of concern among researchers for decades. It is the property of a system that assures a level of certainty regarding the behaviour of that system under a wide range of conditions (Thiebes et al., 2021). Safety-critical computing has been focusing on trustworthiness for ages. Deterministic software use in aviation, nuclear control, and medical devices generated mature engineering traditions around formal verification, fault tolerance,

and assurance through standards (Rudin, 2019; Floridi & Chiriatti, 2020). The generative AI goes against the grain of all three traditions. Its inherently probabilistic nature clashes with a logic-based approach, whereas its opacity towards formal analysis contradicts the goal of mathematically disentangling the behaviour of the system. Also, it learns by acquiring capabilities rather than having them specified, which contrasts with traditional operationalism. One can easily see that the outputs vary significantly with the inputs in ways that are challenging to bound a priori.

The shift from narrow task-specific models to general-purpose generative systems – the GPT series, Claude, Gemini, and their successors – has accelerated the timeline for these concerns from theoretical to immediate. We are deploying these language models in decision-making contexts where their outputs directly affect medical diagnoses and legal judgments, and even financial transactions (Kasneci et al., 2023) and software infrastructure (Ji et al., 2023). It is no longer a question of when these systems routinely fail, but what it will mean for society.

Moreover, a large research literature on trustworthy AI has emerged, including technical alignment methods, robustness evaluation frameworks, red-teaming protocols, and regulatory proposals (Ouyang et al., 2022; Bai, Kadavath, et al., 2022). Despite influential work being carried out, significant fragmentation exists within the AI literature; safety engineers, social scientists, policy analysts, and alignment researchers mostly work in their own worlds, with different languages, success metrics, and threat models. Bridging the gap between these communities would be valuable, along with an evaluation of their evidence base.

This review was organised around four dimensions of trustworthiness that together constitute the principal concerns of the field: (i) safety and alignment – the behaviour of generative AI conforms to intended objectives and human values; (ii) robustness – reliable behaviour that is unaffected by a distribution shift, adversarial inputs, and variation during deployment; (iii) evaluation – methods for assessing trustworthiness properties with validity and reproducibility; and (iv) governance – institutional, regulatory, organisational mechanisms for managing AI risk at scale. In each dimension, we examine the state of knowledge, methodological problems of mainstream approaches, and the distance between claims made by research and deployment.

METHOD

Conceptual framework: Dimensions of trustworthy generative AI

Defining trustworthiness in the generative AI context

Trustworthiness is a collection of characteristics whose relative importance varies with the circumstances of deployment. A medical information assistant demands very distinct trustworthiness properties from a creative writing tool: the first requires high factual accuracy, a conservative expression of uncertainty, and resistance to misleading prompts while the latter could appropriately privilege fluency and creativity over the strict factual inaccuracy. The tendency within the field to view trustworthiness as a context-independent, "one-size-fits-all" property gives rise to evaluation frameworks that fail to distinguish between these dissimilar requirements (Thiebes et al., 2021).

A multi-dimensional framework were drawn from the technical AI safety literature and key regulatory proposals (Laurie E, 2023; "Regulation 2024/1689 of the Eur. Parl. & Council of June 13, 2024 (EU Artificial Intelligence Act)," 2024). The five major dimensions are alignment: behavioural conformity with intended objectives and human values, robustness: stability of performance for perturbations of input and distributional shift, reliability: consistency and reproducibility of output for given input, transparency: interpretability of model reasoning and auditability of system behaviour, and accountability: tractability of responsibility assignment on failure. Governance methods that impose accountability might lead to unintended consequences that promote opacity, while methods that boost transparency might endanger capability, and methods that strengthen alignment might lessen robustness.

The trustworthiness stack

Interventions related to trustworthiness can function at various levels of the AI system stack. Dataset curation (removal of harmful, biased, or low-quality content), data balancing, and

integration of constitutional documents are pre-training interventions. Interventions that fine-tune AI models are instruction tuning, reinforcement learning from human feedback, and direct preference optimisation. System prompt design, output filtering, and retrieval augmentation happen at inference time. Following the deployment were monitoring, auditing, red-teaming, and model updating. Dissimilar failure modes must be treated differently, and levels will have distinct costs to implement (Wang et al, 2023) . The current literature provides no systematic evidence on how interventions at different stack levels interact: do they add up, replace further up the stack, or conflict?

RESULT

Safety and alignment: Techniques and critical analysis

The alignment problem and its manifestations

The alignment problem refers to what AI systems should do to ensure that they behave consistently with human intentions and values. It was formally articulated in the theoretical AI safety literature before generative AI systems were found to have sufficient capabilities to become practically pressing. Nick (2014) and Russell (2019) characterised alignment as a fundamental challenge for sufficiently capable systems. The deployment of capable generative AI systems has made this empirically tractable, not just theoretical.

In the broader world of generative AI alignment, failures can occur in four main ways: (i) sycophancy: producing outputs consisting of what the user wants rather than factual truth; (ii) specification gaming: optimising for the objective instead of the intent; (iii) reward hacking: having an unexpected method of maximising a training signal but not achieving the intended behaviour; and (iv) distribution shift misalignment: behaving well on the training distribution but failing to align in novel contexts. All these behaviours have been shown to occur in deployed systems, and none is reliably prevented by any alignment scheme in use (Perez et al., 2022).

RLHF, DPO, and constitutional AI: Comparative assessment

Reinforcement Learning from Human Feedback (RLHF) became the leading alignment technique since the release of InstructGPT. InstructGPT revealed that human preference feedback greatly boosts the helpfulness and safety of GPT-3 scale models compared to pre-training. RLHF functions by incorporating human preferences to enhance model performances via reinforcement learning (Floridi & Chiriatti, 2020). It first trains a reward model on human preference comparisons. The incorporation of this strategy into leading generative AI systems, including GPT-4, Claude 2/3, and Llama 2, indicates its effectiveness in reducing harmful outputs.

Nevertheless, RLHF has three systematic limitations that are often overlooked. To begin with, the accuracy of the reward model is constrained by the expense and inconsistency of human annotation, and is culturally biased towards the demographics of the annotation workers. In addition, reward hacking may occur when the optimised model produces outputs that are high-scoring on the reward model, but diverge from what humans actually desire. This phenomenon occurs at enough optimisation pressure, leading to a degradation dynamic that disallows pushing RLHF very far. Additionally, RLHF provides no guarantees: a model that is trained with RLHF may output harmful content when prompted with input not represented in the preference data.

Direct Preference Optimisation (DPO) (Rafailov et al., 2023) minimises the training instability problem of RLHF; it casts preference learning as a supervised classification objective for training; without needing reward model training. DPO achieves performance on par with that of RLHF but is significantly cheaper and more stable. Constitutional AI (CAI) (Bai et al., 2022), from Anthropic, extends self-supervised alignment by training models to evaluate and edit their own outputs according to a set of explicit principles – a mechanism that requires less costly human annotation and affords greater transparency about the normative standards being enforced (Ouyang et al, 2022).

Table 1 systematically compares these alignment techniques along various dimensions.

Table 1. Comparative assessment of primary alignment techniques (2017–2024)

Method	Year	Annotation Req.	Training Stability	Scalability	Reward Hacking Risk	Transparency	Key Limitation	Ref.
RLHF	2017	Very High (human pairwise)	Low	Moderate	High	Low	Reward model degradation; annotator bias	[10]
RLAIF	2022	Low (AI feedback)	Moderate	High	Medium	Low	Inherits base model biases; circular alignment	—
DPO	2023	High (human pairs)	High	High	Low–Medium	Medium	Requires large preference datasets; offline	[11]
Constitutional AI	2022	Low (principles doc)	High	Very High	Low	High	Principle conflicts; normative choices opaque	[12]
RLVR	2024	Very Low (verifiable)	High	High	Very Low	Medium	Limited to verifiable domains only	—
SFT Baseline	2020 –	Medium (demonstrations)	Very High	Very High	None	Medium	No preference modeling; limited safety gains	—

Hallucination: A persistent alignment failure

Hallucination is the creation of assured, flowing claims that are factually wrong or unsupported by the evidence available. This is one of the most practically serious alignment failures of current generative AI systems. In contrast to harmful outputs that can be easily targeted once detected, hallucination is a subtle failure. Without independent expertise in the domain, end users may not realise they are experiencing such failure. Ji et al. (2023) provided a systematic taxonomy of hallucination types and mechanisms. They describe two unique types of hallucinations. The intrinsic hallucinations contradict the claim in the source documents. On the other hand, extrinsic hallucinations provide unverifiable hallucinations.

Specific mechanisms responsible for hallucination are unveiled. Language models are trained to predict what occurs next. In domains with sparse or contradictory training data, the most likely continuation is not necessarily the most accurate one. In addition, RLHF-tuned models encounter a sycophancy-accuracy trade-off. Annotation workers prefer fluent, confident answers. They also prefer fluent, confident answers to accurate but hedged ones. This may inadvertently spur hallucination.

Retrieval-augmented generation (RAG) offers a partial safeguard against hallucinations, as it ensures that the generated output is rooted in retrieved factual data. Nevertheless, there are instances when the quality of the retrieval is poor, and the model does not focus on the retrieved data; in these cases, the model still generates hallucinated output.

Robustness and reliability: Evidence and limitations

Adversarial robustness

Since the influential paper by Goodfellow et al. (2014) surfaced, a plethora of work has taken place on adversarial attacks against neural network classifiers, showing that small input perturbations could flip their predictions. The landscape of adversarial attacks against generative AI systems is privy to a distinct threat. Instead of a mere flip of classification, the adversarial inputs are meaningfully called jailbreaks or prompt injections to elicit outputs that violate the system's safety constraints. Alternatively, they can be used to extract sensitive data from the training or manipulate the model in downstream applications.

Red-teaming has emerged as the prime empirical method for generative AI adversarial robustness probed. Individuals in red teams employ creative prompt building, role-play scenarios, and multi-turn deception techniques for unsafe output generation.

Although red-teaming offers satisfactory empirical coverage, it suffers from the same flaw as black-box testing does for traditional software: it can show that a vulnerability exists but not that one does not. Methods for automated red-teaming (Mazeika et al., 2024) allow for scaling vulnerability discovery, but they suffer from the fact that the adversarial input space is unbounded and constantly changing as adversaries adapt to the defences that are already deployed.

Distributional robustness and OOD generalisation

Generative AI models must prove that the outputs remain trustworthy (that they maintain the same properties) for the full distribution of inputs that will be encountered when the model is deployed in the real-world. This includes encountering inputs not seen during evaluation but arising as a result of shifts in user populations, deployment context, or adversarially constructed inputs. There is little evidence of distributional robustness of alignment properties: most evaluations in the literature measure in–distribution performance on benchmark datasets rather than OOD generalisation.

According to the MMLU benchmark (Hendrycks et al., 2020), which is popularly taken as a proxy for capability, it is known to be sensitive to superficial prompt perturbations, including rephrasing, option reordering, few-shot example selection, and so forth. Models that obtain state-of-the-art accuracy on standard MMLU variants often suffer substantial drops in performance given minimally perturbed versions. This suggests benchmark performance is due to benchmark-specific optimisation, not actual robust reasoning. Safety benchmarks also demonstrate this issue: systems fine-tuned on particular harmful content categories often remain vulnerable to new formulations or multi-step attacks that are not represented in the evaluation dataset.

Evaluation frameworks: A critical comparative analysis

The evaluation landscape

The process of evaluating generative AI systems for their trustworthiness has resulted in a burgeoning collection of benchmarks, each assessing diverse facets of safety, alignment, and robustness. This growth creates coordination problems, however, as dissimilar organisations evaluate systems with distinct properties and use diverse metrics. Table 2 shows a systematic comparison of major evaluation frameworks.

The evaluation paradox

A major consequence of this study is referred to as the Evaluation Paradox: the benchmarks often cited in regulatory proposals and safety certifications tend to be the systems that are easiest to manipulate. TruthfulQA (Lin et al., 2022), which is created to measure the factual accuracy, is proven to be gameable through calibration fine-tuning, which reduces the confidence on benchmark questions without improving the actual truthfulness. By contaminating the test set with pre-training data, the MMLU scores get inflated. Undoubtedly, it is an issue that remains hard to audit in closed-source systems. The result is that regulations that require specific benchmark performance levels as a condition of certification may lead to benchmark optimisation rather than real trustworthiness improvements.

This paradox highlights a core measurement issue: the trustworthiness of a system is a property of its behaviour across the entire distribution in which it will be deployed. However, benchmarks sample from a finite, pre-determined subset of that distribution. Once a benchmark becomes a target for any certification, it soon becomes a target for optimisation. Thus, optimising for a benchmark proxy is not the same as optimising for whatever property the benchmark is meant to measure. The perverse effect of gaming, explained by Goodhart's Law, will have serious consequences for AI governance: certification systems based on gaming-susceptible benchmarks provide false assurance that may be harmful.

Human evaluation and its limitations

The human evaluation is a gold standard for many trustworthiness properties. It involves having human annotators make qualitative assessments of model outputs aligned to quality and safety dimensions. Nonetheless, human evaluation shows systematic biases that aren't recognised enough. There is often low agreement among annotators on subjective safety judgments. In the case of nuanced safety scenarios, Cohen's kappa is typically in the 0.4–0.6 range. People who rate the quality of various AI outputs consistently prefer the responses that are verbose and fluent as compared to those that are compact and accurate. These human preferences, when used to train reward models for RLHF (Reinforcement Learning from Human Feedback), could lead to an incentive for hallucination. Annotation workers often have similar demographics, which means they routinely miss content that harms marginalised communities.

Governance mechanisms: Regulatory frameworks and institutional design

The regulatory landscape

Regulators are already implementing guidelines, having swiftly transitioned from talks to action on generative AI governance. There are three regulatory laws at the forefront of proper governance attempts: the European Union AI Act, the United States Executive Order on AI (October 2023), and the Generative AI laws of the People's Republic of China. Table 3 compares varied approaches along with the technical aspects of these models.

Table 2. Comparative analysis of trustworthy AI evaluation frameworks

Framework	Year	Dimension Measured	Evaluation Method	Scope	Ecological Validity	Key Limitation	Ref.
TruthfulQA	2021	Factual accuracy/hallucination	MC + generative Q&A	Factuality	Medium	Static; adversarial to truthful models	[17]
MMLU	2021	General knowledge / capability	Multiple choice	Broad capability	Low–Medium	Prompt-sensitive; contamination likely	[16]
BBQ	2022	Social bias	Ambiguous QA	Bias	Medium	Measures bias expression not root cause	—
HarmBench	2024	Adversarial safety / jailbreak	Automated red team	Attack robustness	Medium–High	Arms race dynamic; rapidly dated	[15]
MT-Bench	2023	Instruction following	GPT-4 as judge	Helpfulness	High	Judge model biases; evaluator-evaluated correlation	—
HELM	2022	Multi-dimensional	Holistic evaluation	Very broad	Medium	Aggregation obscures dimension tradeoffs	—
Decoding Trust	2023	Trustworthiness (8 dimensions)	Comprehensive	Broad	Medium–High	Computationally expensive; model-specific	[18]
Safety Bench	2023	Safety / refusal	Scenario-based	Safety	Medium	Binary refusal metric; nuance lost	—

Table 3. Comparative analysis of major AI governance frameworks (2021–2024)

Framework	Jurisdiction	Year	Risk Classification	Technical Requirements	Enforcement Mechanism	Key Strength	Critical Gap	Ref.
EU AI Act	European Union	2024	Four-tier risk levels	Conformity assessment; red-teaming; transparency	Market access prohibition; fines up to 7% global revenue	Comprehensive scope; legally binding	Vague technical requirements; lag pre-deployment	[5]
US EO on AI	United States	2023	Dual-use / safety threshold	Safety reports; red-team results to government	Voluntary initially; sector-specific regulations pending	Industry engagement; rapid deployment	Non-binding; fragmented across agencies	—
China GenAI Regs	China	2023	Content + provider focus	Algorithmic transparency; content moderation	Platform liability; licensing	Speed of implementation	Limited technical depth; primarily content-focused	—
NIST AI RMF	United States	2023	Risk management process	Map-Measure-Manage-Govern cycle	Voluntary (but referenced in contracts)	Practical guidance; sector-neutral	Non-mandatory; no enforcement	[6]
G7 Hiroshima Principles	International	2023	Voluntary principles	Transparency; accountability; robustness	Peer review; no formal enforcement	International coordination	No binding commitments; lowest common denominator	—
ISO/IEC 42001	International	2023	Management systems standard	AI management system requirements	Certification by accredited bodies	Auditable; third-party verification	Procedural focus; limited technical depth	—

Critical analysis of the EU AI act

The EU AI Act is the first attempt to translate trustworthiness requirements into legally binding obligations concerning AI. By prohibiting some applications altogether (social scoring, real-time biometric surveillance), placing stringent requirements on high-risk applications, and imposing transparency obligations on limited-risk applications, it shows a nuanced understanding of risk disrepancy between application context of deployment. Nonetheless, some provisions explain the divergences between regulatory ambition and technical reality ("Regulation 2024/1689 of the Eur. Parl. & Council of June 13, 2024 (EU Artificial Intelligence Act)," 2024).

According to the Act, the deployers of high-risk systems must demonstrate that it complies with accuracy, robustness, and safety requirements. However, it does not establish thresholds of technical significance for these properties nor standardised design methods for assessing them. The compliance burden created by this regulation is onerous for legitimate developers and easily manipulated by bad actors. The need for human oversight for high-risk applications poses profound questions about what meaningful human control is –when system complexity exceeds human cognitive capacities, a situation that is becoming more common in AI-assisted decision-making contexts.

Institutional misalignment: Deployment vs. governance maturity

A structural challenge in AI governance refers to the misalignment between the speed at which generative AI is deployed and the quality of governance mechanisms. Organisations take two to five years to develop their AI governance frameworks and build the required technology that is relevant to their framework. After this, the effectiveness of these frameworks can be assessed. The requirements of the actual users are unknown in this framework. The distinction between the pace of technological innovations and the governance frameworks is not a transitional phenomenon but rather a permanent structural feature.

The voluntary safety commitments made by leading AI developers, Microsoft, Google, Anthropic, and OpenAI, offer some governance in the absence of regulation, but they have well-known shortcomings. Non-binding pledges are not legally binding, may be modified or abandoned as a result of commercial interests and do not bind competitors or new entrants. The AI safety assessments performed under these commitments are frequently conducted by the organisations developing them. This leads to concerns over independence similar to those that led to mandatory third-party auditing in finance.

Comparative performance on safety benchmarks (2020–2025)

Table 4 synthesises published safety evaluation results across major generative AI systems on the benchmarks discussed in Section 5. Where multiple evaluation versions exist, the most recent published results are used. All figures are from primary papers or official model cards; unavailable values are marked with '—'.

Analytical observations from benchmark data

The comparative data in Table 4 yield several insights. Most notably, Constitutional AI and DPO-based methods consistently outperformed standard RLHF-trained systems in TruthfulQA performance. Claude 2 and Claude 3 achieved the highest TruthfulQA scores. This scores reflects the explicit focus on factual calibration, due to the CAI training methodology. This implies that the frequently contended safety-capability trade-off is not invariant, but dependent on the particular alignment technique used.

Second, the attack success rate of HarmBench shows a generation-wise improvement, as it is greater than zero for all evaluated systems. A major finding for governance is that the lack of any system with zero ASR in extensive adversarial evaluation means that there is no currently deployed system that has shown robustness to adversarial attack to be unconditionally deployable. Treating the pass rates of safety evaluation as binary sufficient conditions within governance frameworks is methodologically unsound.

Table 4. Safety and trustworthiness benchmark performance across major generative AI systems

System	Release Year	TruthfulQA (%)	HarmBench ASR (%)	MMLU (%)	Decoding Trust Score	Alignment Method	Trans-parency
GPT-3 (base)	2020	21.4	High (~80%)	43.9	Low	Pre-training only	Closed
InstructGPT (RLHF)	2022	54.0	Moderate	52.0	Medium	RLHF	Partial
GPT-4	2023	59.0	Low–Medium	86.4	Medium–High	RLHF + SFT	Closed (system card)
Claude 2 (CAI)	2023	68.0	Low	78.5	High	Constitutional AI + RLHF	Partial (model card)
Llama 2 (Meta)	2023	57.4	Low–Medium	68.9	Medium	RLHF (open model)	Open weights
Gemini Ultra	2024	—	Low	90.0	Medium–High	RLHF + SFT	Closed (tech report)
Claude 3 Sonnet	2024	71.0	Very Low	90.2	High	CAI + DPO	Partial
GPT-4o	2024	—	Low	88.7	Medium–High	RLHF + SFT	Closed (system card)

The lack of visibility into most deployed systems -with only Llama 2 available with open weights, and most systems only partly documented in model cards -greatly limits independent safety audits. This puts governance in a principal-agent dilemma: regulators and users must trust developers' self-reported safety metrics, but they do not have the technical access to check them. The EU AI Act proposal involves a requirement for mandatory disclosure of safety evaluation methodologies and results to independent auditing bodies. This helps fill the accountability gap, albeit facing strong commercial push-back.

Critical synthesis: Gaps between research and practice

The technical–governance interface

A weak interface between technical AI safety research and governance can be observed across all four dimensions investigated in this review. Regulatory frameworks define trustworthiness properties: safety, robustness, and transparency in terms that make them insufficiently precise for operationalising in technical evaluation. Technical evaluation frameworks, on the other hand, are crafted with little regard for the governance environments where their outputs will be analysed and used. Given this bidirectional misalignment, governance mechanisms are unable to translate into technical requirements, while technical advances cannot be relied upon to improve governance outcomes.

NIST AI risk management framework, in a structured manner, that aims to address this gap. It provides a process for identification, measurement, and mitigation of the risk of AI in a technically informed and governance relevant way. Despite this, the fact that it is voluntary and process-oriented, a specification on how organisations should manage AI risk rather than what technical properties they must have, limits its effectiveness as a mechanism to guarantee

trustworthiness. All the major governance frameworks, at present, lack technically precise requirements that are compulsory in nature, along with standardised measurement methods, third-party verification, etc.

Demographic and cultural representativeness

Bias occurs in safety and alignment evaluations that favor the developers' demographic and cultural contexts. Annotating data for RLHF takes place largely in North America and Europe, along with a few other wealthy countries. This reflects systematic differences in cultures, languages, and notions of harmfulness, which do not generalise. The generative AI systems that have been deployed have been found to show a higher rate of refusal to answer questions posed in minority languages. They were less factually accurate regarding non-Western geography and history. Moreover, there was a systemic diecrepancy in how values were expressed in diverse languages.

The disparity between races and ethnicities creates not only an equity issue, but also a trustworthiness issue. The distributional failures caused by systems misaligned to a representative population are genuine safety risks in global deployment. The existing frameworks of governance have an inadequate focus on demographic representativeness as a dimension of trustworthiness.

Several priority research challenges in advancing trustworthy generative AI during the 2025–2030 period include the development of formal verification methods for alignment properties to ensure that model behavior can be systematically constrained beyond reliance on purely empirical evaluation. Additional challenges involve the establishment of robust safety benchmarks that are resilient to adversarial optimization and manipulation of evaluation frameworks, as well as the development of cross-cultural alignment methodologies capable of adequately representing the diversity of global cultural values and perspectives. Furthermore, there is a pressing need for scalable automated auditing infrastructures that enable independent third-party oversight while remaining compatible with commercial deployment environments. Equally important is the advancement of uncertainty quantification techniques that allow generative AI systems to communicate well-calibrated confidence levels in their outputs. Another critical area of research concerns long-horizon alignment approaches aimed at preserving alignment consistency across extended multi-turn interactions and agentic deployment settings. Beyond these technical dimensions, a significant challenge also lies in the formulation of standardized governance–technical interface frameworks capable of translating regulatory requirements related to trustworthiness into precise and operational technical specifications..

DISCUSSION

The conclusions supported by the research discussed in this paper contradict widely-held views. To begin with, the present situation in the alignment is one wherein techniques can lower the rate of alarmingly toxic outputs, but do not offer principled guarantees against any distributional failures, or adversarial manipulations, or subtle misalignments. The approaches of RLHF (Ouyang et al., 2022), CAI (Bai et al., 2022), and DPO (Rafailov et al., 2023) do impose changes over pre-training, which was not aligned. However, it is not a complete solution for full alignment to use in high-stakes deployment contexts (Rudin, 2019).

Furthermore, the most popular evaluation frameworks used to certify AI safety TruthfulQA, (Lin et al., 2021), MMLU (Hendrycks et al., 2021), and HarmBench (Mazeika et al., 2024) can be shown to be gamed and have limited distributional coverage. Safety assurances that mislead developed countries are created by governance frameworks that treat performance on these benchmarks as certification criteria. It is of utmost research priority to create adversarially robust and comprehensive evaluation standards that cannot be gamed.

Governance mechanisms across all major jurisdictions suffer from structural misalignment with the speed and complexity of generative AI deployment. The currently dominant governance

model based on voluntary commitment lacks enforcement, independence, and technical specificity. Instituting compulsory, accurately specified, independently assessed governance mechanisms (Schuett et al., 2025a) that are similar to financial auditing or pharmaceutical approval will bring about genuine trustworthiness improvements that will likely arise from governance reforms (Schuett et al., 2025a). The literature increasingly proposes institutional mechanisms analogous to financial auditing, aviation certification, or pharmaceutical approval processes, involving mandatory third-party model evaluation and pre-deployment stress testing (Shevlane et al., 2023; Anderljung et al., 2023). Instituting compulsory, technically rigorous, and independently assessed governance infrastructures would likely generate more substantive improvements in trustworthiness than reliance on voluntary compliance mechanisms.

None of these conclusions suggests that no progress has been made or that today's systems are not significantly safer than yesterday's. Claude 3 or GPT-4's impact from GPT-3 is proof of significant trustworthiness improvements (Kasneci et al., 2023). Nonetheless, this trajectory has primarily been realised via scaling and empirical iteration, rather than through theoretical understanding, while the implicit extrapolation –that unceasing scaling and iteration will one day yield truly trustworthy systems –is not backed by rigorous justification (Kaplan et al., 2020; Hoffmann et al., 2022). It remains a fundamental open challenge to develop complementary theoretical foundations for trustworthiness, similar to those for cryptographic security or fault-tolerant computing.

CONCLUSION

This article presents a systematic, critical study of trustworthy research across the themes of safety and alignment, robustness, evaluation methodology, and governance for generative AI published from 2015 to 2025. The field has matured from toy models to deployed, regulated, and monitored systems. There is already a very substantial body of empirical knowledge piling up, and yet foundational problems remain unaddressed.

This review makes an essential contribution: a critical synthesis of technical safety. It identifies three structural gaps: (1) the alignment gap, referring to current techniques as not satisfying or offering principled safety guarantees; (2) the evaluation gap, which describes benchmark performance not being trustworthy in a wider, real-world (or non-benchmarked) context; and (3) the governance gap, illustrating that regulatory or governance ambitions are not matched by technical capacity. Bridging these gaps calls for continuing investment in foundational research, linking this to coordinated technical-governance interface development. However, no single research community or institutional actor can deliver this alone.

The consequences of this challenge are not hypothetical. At scale, generative AI systems are already influencing medical diagnoses, legal verdicts, financial transactions, and educational assessments. The systems' trustworthiness properties, or lack thereof, have real-world consequences for the people affected by the systems. It is thus not only a research challenge but also a social challenge to close those gaps.

Funding

This reseaarch did not receive any specific funding from public, private, or non profit organisations.

Data availability statement

All data are available from the author.

Conflict of interest

The authors declare no conflicts of interest related to this work.

REFERENCES

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., ... & Wolf, K. (2023). Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint.* *https://doi.org/10.48550/arXiv.2307.03718*.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., & Henighan, T. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv Preprint. https://doi.org/10.48550/arXiv.2204.05862*.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Cabello, L., Jørgensen, A. K., & Søgaard, A. (2023, June). On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 370-378). https://doi.org/10.1145/3593013.3594004

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and machines*, *30*(4), 681-694. https://doi.org/10.1007/s11023-020-09548-1

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *ArXiv Preprint. https://doi.org/10.48550/arXiv.1412.6572*

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., & Steinhardt, J. (2020). Aligning AI With Shared Human Values. *CoRR*, *abs/2008.0*. https://arxiv.org/abs/2008.02275

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, *10*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Laurie E, L. (2023). *Artificial Intelligence Risk Management NIST AI 100-1 Artificial Intelligence Risk Management*.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., & Li, B. (2024). Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *ArXiv Preprint ArXiv:2402.04249*.

Nick, B. (2014). Superintelligence: Paths, dangers, strategies. *Strategies*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). Red teaming language models with language models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3419–3448. 10.18653/v1/2022.emnlp-main.225

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*, 53728–53741.

Regulation 2024/1689 of the Eur. Parl. & Council of June 13, 2024 (EU Artificial Intelligence Act). (2024). In *Official Journal of the European Union* (2025/03/10). Cambridge University Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206-215. https://doi.org/10.1038/s42256-021-00421-3

Russell, S. (2019). *Human compatible: AI and the problem of control*. Penguin Uk.

Schuett, J. (2025). Frontier AI developers need an internal audit function. *Risk Analysis*, *45*(6), 1332-1352. https://doi.org/10.1111/risa.17665

Schuett, J. (2025). Three lines of defense against risks from AI. *AI & SOCIETY*, *40*(2), 493-507.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., ... & Dafoe, A. (2023). Model evaluation for extreme risks. *arXiv preprint. https://doi.org/10.48550/arXiv.2305.15324*.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). *Trustworthy artificial intelligence*. 447–464. https://doi.org/10.1007/s12525-020-00441-4

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., ... & Li, B. (2023). DecodingTrust: A Comprehensive Assessment of Trustworthiness in {GPT} Models. https://par.nsf.gov/servlets/purl/10616577.