Too Focused on Accuracy to Notice the Fallout: Towards Socially Responsible Fake News Detection

Esma Aïmeur, Gilles Brassard, Dorsaf Sallami

Department of Computer Science and Operations Research (DIRO), University of Montreal, Montreal, Canada
aimeur@iro.umontreal.ca, brassard@iro.umontreal.ca, dorsaf.sallami@umontreal.ca

Abstract

The rise of fake news is one of the most pressing threats to the digital public sphere. Artificial intelligence (AI) systems promise to fight it — but at what cost? Unlike other machine learning applications designed to optimize efficiency in low-stake domains, fake news detection operates at the core of democratic discourse, public trust and epistemic integrity. This paper begins by unpacking the core challenges that make fake news detection uniquely demanding. In response, we argue for a shift toward Socially Responsible AI (SRAI) as a more appropriate framework for addressing these complexities. We map the identified challenges onto the SRAI pyramid—functional, legal, ethical and philanthropic. Finally, we review emerging initiatives, highlight current limitations and propose future directions for developing fake news detection systems that are not only accurate but also socially accountable and publicly trustworthy.

Introduction

The proliferation of fake news has emerged as a defining challenge of the digital age, undermining the integrity of global information ecosystems. Traditional interventions, such as manual fact-checking, have proven insufficient in addressing the scale, speed and evolving complexity of fake news (Sallami and Aïmeur 2025c). In response to these limitations, the research community has increasingly turned to automated solutions, with Artificial Intelligence (AI) and Machine Learning (ML) techniques playing a central role in the development of large-scale fake news detection systems.

This technological shift has spurred a competitive research environment focused heavily on benchmark performance. Models are routinely evaluated using metrics such as accuracy, precision and recall and then ranked on public leaderboards. In fake news detection, this culture of "leaderboardism" (Hutchinson et al. 2022) has led to a prioritization of marginal performance gains over real-world relevance. While benchmarking has undoubtedly propelled technical advancements, it often obscures empirical limitations. Models that excel on curated datasets may struggle to generalize to emerging or underrepresented forms of fake news.

Critically, even high-performing models are not immune to error and in the context of fake news, such errors can yield disproportionately harmful consequences. A weather forecast with 90% accuracy entails minimal risk: at worst, an individual unnecessarily brings an umbrella. By contrast, a fake news detection model that erroneously labels fake news as credible with similar confidence may facilitate the spread of falsehoods. These asymmetric consequences underscore the need to reassess error tolerance and consider the ethical implications of deploying such systems at scale.

As Hutchinson *et al.* (2022) and Bengio *et al.* (2021) argue, prevailing evaluation frameworks in ML do not always translate into real-world efficacy. Overreliance on narrowly defined metrics and static benchmark datasets risks fostering a false sense of robustness while concealing critical vulnerabilities. Hence, a fake news detector that performs well under controlled conditions may fail catastrophically in the wild.

More fundamentally, the field has adopted a reductive framing of fake news as a classification task, an approach that abstracts away its deeply social, political and epistemological dimensions. This reductionist view has led to the optimization of models for metrics rather than for outcomes that more meaningfully align with societal well-being. In this context, a narrow focus on accuracy or efficiency is not only inadequate, it is potentially harmful. As these tools are increasingly deployed in real-world settings, it is essential to critically engage with their potential risks and unintended consequences. This represents a significant gap in the current literature, one that this paper seeks to address through an ethically and socially grounded lens. Although similar concerns have been explored in adjacent domains, including public employment services (Dahlin 2021), judicial decision-making (Lakkaraju et al. 2017) and medical diagnosis (Lebovitz, Levina, and Lifshitz-Assaf 2021), they remain underexamined in the context of fake news detection.

We contend that before pursuing increasingly complex models to "solve" the fake news problem, we must first grapple with the harms these solutions may themselves produce. We argue that while AI detection methods have made significant technical strides, they also risk reinforcing existing inequities or introducing new forms of harm. Our goal is not to reject the use of AI in fake news detection outright, but rather to promote a more critical, reflective and socially responsible approach to its development and deployment. Through the lens of Socially Responsible AI (SRAI), we aim

to reframe the discourse from a focus on algorithmic performance to one centred on ethics and public accountability.

Rethinking AI for Fake News Detection

Addressing fake news with AI may seem inevitable in an age of information warfare. Yet as we turn to machines to judge truth, it becomes clear that the very tools we deploy carry risks of their own. Rethinking how AI engages with fake news is no longer optional; it is essential. To understand the uniqueness of this domain, we can examine it across three key dimensions (Figure 1): System–Society Conflict, Limits of Learning and Media-Specific Challenges.

System–Society Conflict

The first challenge lies in the value conflicts between what AI systems traditionally optimize for and what society demands from interventions in the information ecosystem.

Automation vs. judgement is at the core of this misalignment. Historically, ML has been optimized for efficiency, automating tedious or low-skill tasks like handwriting recognition. The primary concern was maximizing accuracy and minimizing human labour, with little attention to broader social consequences. In these contexts, the goal of automation was clear and ethically uncontroversial: improving efficiency. Even when errors occurred, such as misidentifying rare handwriting characters, they were seen as technical limitations rather than moral failings. However, as AI systems have expanded into domains that involve making judgements about truth, this framework has become inadequate. Fake news detection demands more than efficiency. This misalignment has a *social impact* (Narayanan 2019). Errors in early ML systems affected operational efficiency. For instance, consider product recommendation systems, an incorrect prediction here may simply result in a user being recommended vanilla ice cream instead of chocolate, a minor inconvenience with limited impact. In contrast, errors in fake news detection have a direct impact. Misclassifying legitimate news as fake can suppress minority or dissenting views. Conversely, failing to detect fake news can mislead about critical issues such as public health. Unlike product recommendations, these impacts ripple through social trust, fundamentally reshaping public opinion over time.

The tension between ***utility vs. fairness*** further complicates the problem. While it may seem sufficient to maximize the number of fake news articles detected (high utility), such an approach can result in distributional unfairness. For example, a model might focus heavily on detecting political fake news while consistently missing rare but critical medical news (Kato, Yang, and Ikeda 2022). Even if these failures represent a small fraction of cases, they disproportionately harm vulnerable groups who rely on accurate information the most (Sallami and Aïmeur 2024). Moreover, the data-driven nature of ML introduces risks of bias (Raj, Mukherjee, and Zhu 2023). If an algorithm correlates certain writing styles based on historical patterns, it may unjustly suppress legitimate voices, even when individuals evolve or change their behaviour (Murayama, Wakamiya, and Aramaki 2021). Without deliberate ethical design, detection systems risk entrenching the very inequalities they aim to dismantle (Deepak 2020).

The ***accountability gaps*** further compound the issue. In this context, the use of ML algorithms carries significantly greater societal responsibility. This responsibility manifests along at least two dimensions. First, there is an obligation to the media organizations whose content is being categorized as either fake or legitimate. Second, there is a responsibility toward the users who are expected to engage with and trust the algorithmic outputs (Deepak 2020). How these obligations can be effectively fulfilled remains an open question. One possible approach is to ensure that algorithmic decisions are accompanied by *explanations* or traceable justifications that can be made publicly available, thereby allowing for critical scrutiny and informed debate. A further facet of accountability gaps in this context concerns the *privacy and security* of the models themselves. Certain fake news detection techniques utilize user profile information or behavioural data (Sahoo and Gupta 2021), raising concerns about data protection and consent. Moreover, if such data is not adequately safeguarded, it may expose users to risks.

Limits of Learning

Beyond the broader system–society conflict, the task of fake news detection is further constrained by the fundamental limitations of current ML approaches.

To begin understanding the limits of learning, we have to start with the most basic challenge: the ***ground truth problem***. Ground truth refers to the labelled data used to train and evaluate ML models. In many cases, labels are produced by human experts. Sometimes, this labelling process is relatively clear-cut. For instance, in an image classification task where the goal is to distinguish between images of cats and dogs, the labels are typically obvious. Most people can reliably tell the difference between the two. However, labelling becomes far more uncertain in fake news detection. Experts may disagree on whether a piece of content is misleading or false. This ambiguity shapes the outcomes of ML models (Lebovitz, Levina, and Lifshitz-Assaf 2021). Nevertheless, in much of the academic literature, model evaluations rely heavily on labels provided by third-party organizations (Sallami, Gueddiche, and Aïmeur 2023; Moalla et al. 2025). While researchers often defer to these external sources to minimize the influence of personal bias on the labelling process, the methodologies employed by these organizations may themselves lack transparency and be subject to their biases (Kuntur et al. 2024). Notably, some of these credibility-rating entities can operate as for-profit businesses. When classifiers are trained on such labelled data, another layer of opaque decision-making is introduced, further complicating the determination of what is deemed credible or misleading information.

This leads to the second challenge, known as the ***accuracy trap***. AI systems do not "know" facts. They calculate likelihoods based on training data and generate outputs that are statistically coherent within a given context. This raises a critical challenge: when AI flags or verifies information, is it performing an act of truth discernment or merely simulating it? This concern leads to a deeper realization: *AI is*

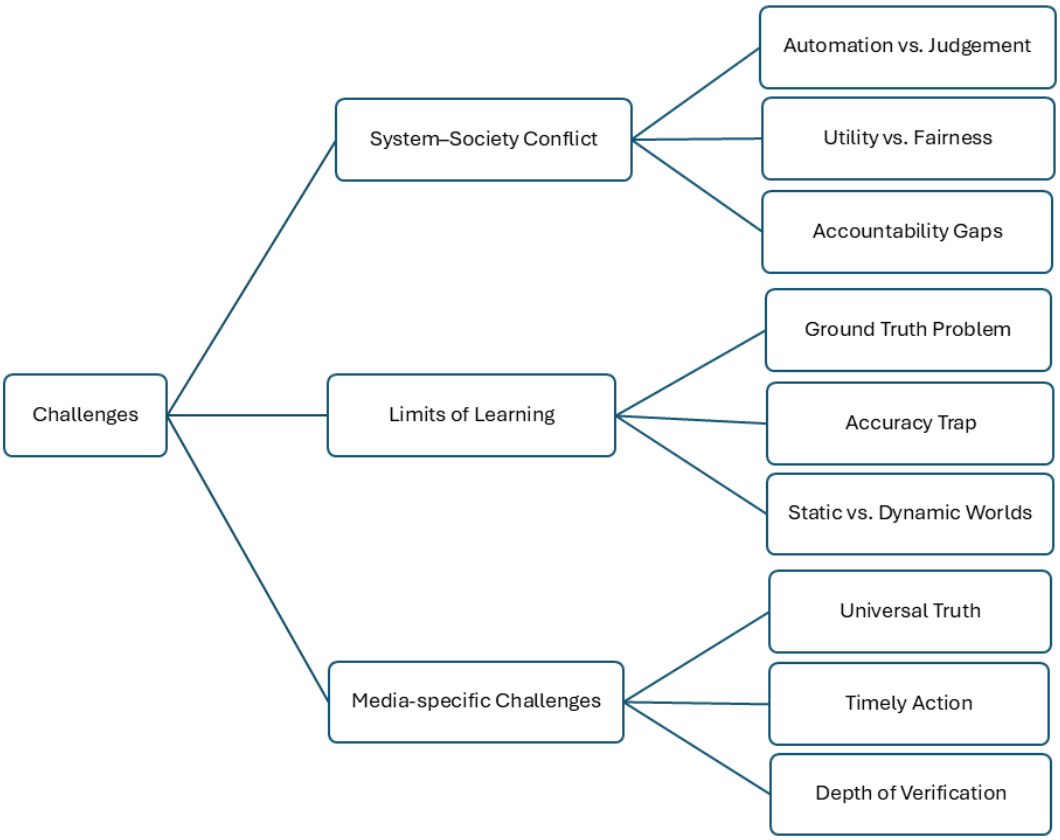

Figure 1: Key challenges in applying AI to fake news detection.

not a fact-checker, it is a mirror. Efforts to position AI as an autonomous fact-checker reflect a broader tendency to treat fake news as a technical glitch, one that can be "solved" by more sophisticated algorithms. But this presumes that AI is capable of independently discerning truth from falsehood, an assumption that fails to recognize AI's role as a mirror of its data and design. These systems reflect the biases, omissions and assumptions embedded in their training corpora. Consequently, when AI flags a piece of news as "fake" or "real," it is not operating on a universal standard of truth but on parameters defined by its training environments. Recognizing this, it becomes clear that *AI's accuracy isn't about truth, it's about coherence*. The notion of "accuracy" in AI-driven systems often rests on a misunderstanding of what these systems are designed to do. AI models generate outputs based not on verification, but on statistical likelihood. They produce what is most probable, not necessarily what is most true. As such, AI can be accurate in terms of internal consistency while being factually incorrect in substance. This distinction matters. When AI identifies or fails to identify fake news, the result may reflect the model's ability to align with patterns it has previously encountered, not its access to an objective reality.

Beneath these issues lies a deeper limitation of ***static vs. dynamic worlds***. ML traditionally relies on the assumption that the distribution of training data matches that of the real-world data the system will encounter. This assumption holds in relatively stable domains like handwriting recognition, where individual styles remain relatively stable. In contrast, a model trained on yesterday's fake news patterns may be blind to today's. Indeed, historical data rapidly loses relevance and models trained even a few months prior can become ineffective. This stems from the phenomenon of *adversarial evolution*, in which fake news creators actively adjust their tactics to circumvent detection. This challenge is further intensified by the emergence of powerful Large Language Models (LLMs), whose unprecedented capabilities in reasoning and generative tasks (Berti, Giorgi, and Kasneci 2025) have significantly narrowed the distinction between machine-generated and human-authored content. As a result, malicious actors are now equipped with tools that enable them to closely imitate the linguistic style of credible news outlets, thereby enhancing their ability to circumvent automated detection systems (Wu, Guo, and Hooi 2024). Indeed, current detection systems struggle to maintain accuracy and robustness in the face of evolving adversarial tactics (Sallami, Gueddiche, and Aïmeur 2023).

Media-Specific Challenges

Compounding these difficulties are the special challenges posed by the media domain itself.

A primary issue is the ***universal truth***. In many ML ap-

plications, personalization is a legitimate and even desirable goal (Rafieian and Yoganarasimhan 2023). Returning to our earlier example of a recommender system, offering one user vanilla and another chocolate highlights how personalization can be both effective and ethically straightforward. Fake news detection, however, cannot accommodate personalized truth. A false claim about vaccines cannot be legitimate for one group and fake for another. Truth demands a normative standard that is applied universally, regardless of personal beliefs or satisfaction metrics. In this context, even seemingly useful features, such as a user's profile information, can blur the line between universal evaluation and personalization. Some studies (Dou et al. 2021) explore user preference-aware fake news detection, but this personalization risks aligning judgements with user beliefs rather than objective facts. Classifiers that rely on such data risk basing judgements on audience profiles rather than content, thereby introducing a form of personalization incompatible with the pursuit of objective truth (Allein, Moens, and Perrotta 2023).

Timely action is also crucial. Research shows that exposure to fake news can have lasting effects, even after retractions (Kemp, Alexander, and Wahlheim 2022). Therefore, detection systems must intervene quickly to minimize the window during which fake news spreads. Moreover, if an initial classification is later reversed, identifying a story as fake after it was initially considered true, ethical responsibility demands informing users who encountered the original error. Otherwise, fake news continues to influence public opinion long after the system's judgement has changed.

The final subtle issue concerns the ***depth of verification***. Many detection systems focus on superficial verification, such as checking whether a person made a certain statement (Amri, Boleilanga, and Aïmeur 2023). For example, confirming that a celebrity said "turmeric cures COVID-19" satisfies surface-level verification. Yet deeper truthfulness matters. Even if the celebrity genuinely made the statement, the underlying claim is false. Detection systems that stop at surface-level verification risk legitimizing harmful falsehoods under the guise of journalistic fidelity. A genuinely responsible system must distinguish between accurate reporting of statements and endorsement of factual truth.

Toward Socially Responsible AI in Fake News Detection

Tackling the challenges of AI in fake news detection requires more than incremental improvements, it demands a fundamental redesign of AI systems. We argue that a Socially Responsible AI framework is suited to meet these challenges.

Socially Responsible AI

Definition The notion of *Socially Responsible AI*, as introduced by (Cheng, Varshney, and Liu 2021), offers a comprehensive framework for guiding AI development. Cheng *et al.* define SRAI as a process in which concepts such as fairness, transparency, or reliability are *principles-by-design*, AI algorithms are the *means* and addressing social challenges to both generate positive impacts on society and improve AI is the *goal*. SRAI is structured across four distinct levels of AI responsibility (Figure 2).

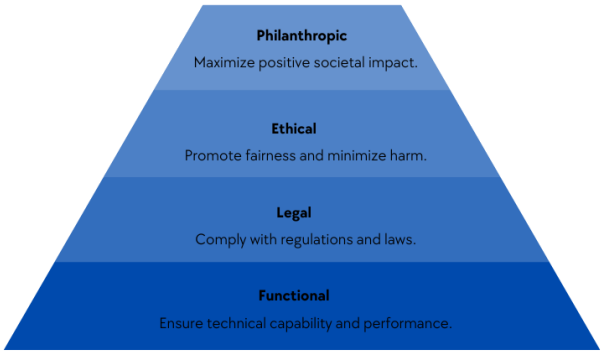

Figure 2: Pyramid of the Social Responsibility of AI, adapted from (Cheng, Varshney, and Liu 2021).

At the base is *functional responsibility*, which refers to an AI system's ability to solve tasks effectively. This level concerns performance, ensuring the AI accomplishes the task it was built for, using metrics like accuracy or error rates. Building on this foundation is *legal responsibility*. This level asserts that certain practices, even if effective, may be legally or ethically unacceptable. AI systems must operate within the boundaries of applicable laws. Beyond legal compliance lies *ethical responsibility*. This dimension emphasizes aligning AI behaviour with moral and ethical expectations. Unlike legal constraints, ethical principles guide AI agents to foster trust, reliability and security among all stakeholders. The highest level of this hierarchy is *philanthropic responsibility*. Even when an AI system satisfies functional, legal and ethical standards, its broader objective should be to contribute positively to society.

Comparison of Related Concepts Based on the definition, we compare SRAI with a range of related concepts that have emerged within the AI discourse. Several well-known concepts are often discussed in the AI literature: (1) *Robust AI* refers to AI systems with the ability "to cope with errors during execution and cope with erroneous input" (Cheng, Varshney, and Liu 2021). (2) *Ethical AI* involves systems that do what is right, fair and just, aiming to prevent harm (Lepri, Oliver, and Pentland 2021). (3) *Trustworthy AI* describes systems that achieve their full potential if trust can be established in their development, deployment and use (Thiebes, Lins, and Sunyaev 2021). (4) *Fair AI* focuses on systems that are absent from "any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics" (Mehrabi et al. 2021). (5) *Safe AI* pertains to AI systems deployed in ways that do not harm humanity (Morales-Forero, Bassetto, and Coatanea 2023). (6) *Dependable AI* concerns systems that focus on reliability, verifiability, explainability and security (Singh, Vatsa, and Ratha 2021). Finally, (7) *Human-centred AI* is defined as AI systems that are continuously improving because of human input while providing an effective experience between human and robot (Capel and Brereton 2023).

While these concepts contribute valuable perspectives, SRAI encompasses a broader vision that unites these elements within a holistic structure. It encompasses existing concepts such as ethical and trustworthy AI while emphasizing philanthropic duties. Whereas ethical AI often focuses on compliance to prevent harm, SRAI emphasizes adherence to societal norms. While closely related, SRAI centres on the AI system's obligations to society, distinguishing it from ethics, which covers the responsibilities of all stakeholders, including researchers, engineers and users. Similarly, trustworthy AI focuses on the unidirectional trust relationship from users to AI systems. In contrast, SRAI stresses collective accountability (Cheng and Liu 2023).

Fake News Detection Through the SRAI Lens

The need for SRAI framework is urgent in the context of fake news detection. To illustrate the stakes, consider a platform that deploys a detection model with over 95% accuracy (functionally effective). Despite its technical performance, the system is trained on copyrighted content without authorization and collects user data without consent, violating privacy laws (legally non-compliant). It offers no transparency to users and exhibits measurable bias against non-Western media outlets (ethical flaws). In certain regions, the system is even repurposed to censor political opposition and suppress protest-related content, with no effort to promote media literacy (philanthropically irresponsible).

To mitigate such risks, we propose an adaptation of the SRAI framework. In the original pyramid, responsibilities are structured hierarchically from functional capabilities to the philanthropic goals, suggesting a progression from technical feasibility toward broader societal impact. However, in the context of fake news detection, where the societal consequences are profound and far-reaching, we propose an inversion of this pyramid (as illustrated in Figure 3). Although the stages remain functionally defined, this inversion shifts design attention to begin with philanthropic responsibility rather than technical feasibility, emphasizing that AI development should be guided from the top by societal values. This reorientation reflects our belief that philanthropic responsibility should not be treated as a final consideration but rather as the primary driver of system design and evaluation.

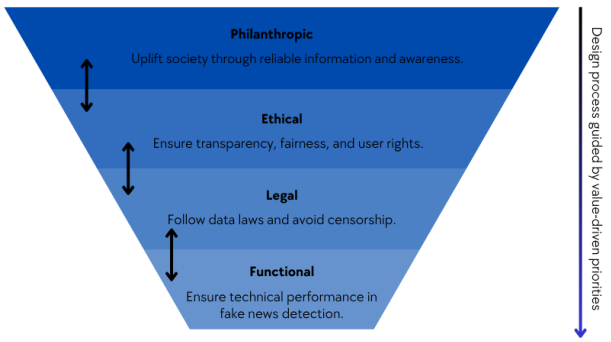

Figure 3: Inverted SRAI pyramid for fake news detection.

Notably, fake news detection systems are frequently well-optimized, demonstrating high performance (Sallami and Aïmeur 2025c). As a result, technical capability is no longer the primary issue. However, high performance alone does not guarantee that such systems serve the public good. This inversion signals that questions of societal benefit must be the starting point, not the end point. In this framework, capabilities are not judged solely by their technical merits but by how well they serve philanthropic goals.

Moreover, we recognize that these responsibilities are interrelated, engaging in an ongoing process of feedback and adaptation. Rather than viewing dimensions as independent or sequential, we argue that each layer should inform and evolve in response to the others. For example, advances in functionality, such as improved detection of multilingual fake news, may broaden the scope of ethical and philanthropic commitments, encouraging goals such as linguistic equity or global media inclusion.

After we justified our shift toward the SRAI framework, we now aim to map the previously discussed challenges onto its core principles (Table 1).

SRAI level	Fake News Challenges
Philanthropic	- Universal Truth - Automation vs. judgement - Accuracy Trap
Ethical	- Utility vs. Fairness - Accountability gaps (Explainability) - Ground Truth Problem
Legal	- Accountability gaps (Privacy and security)
Functional	- Static vs. Dynamic Worlds - Timely Action - Depth of Verification

Table 1: Mapping fake-news challenges onto the four tiers of Socially Responsible AI.

At the philanthropic level, issues such as universal truth, automation vs. judgement and the accuracy trap reflect the need for the necessity of establishing shared societal standards. The ethical layer encompasses challenges of accountability gaps, the ground truth problem and utility vs. fairness, emphasizing the need for transparency in algorithmic decision-making and addressing biases in training data, issues central to fostering fairness and public trust. The legal layer reinforces accountability gaps through regulatory mandates, ensuring compliance with transparency laws and redress mechanisms that align technical systems with democratic norms. Finally, the functional layer tackles static vs. dynamic worlds, depth of verification and timely action, necessitating adaptive technical solutions to counter evolving adversarial tactics, verify contextual accuracy and minimize harm through rapid response.

Current Initiatives

Figure 4 presents an overview of existing initiatives. Most research focuses on explainability, while efforts on bias mit-

igation and other ethical or legal aspects remain limited.

Functional Initiatives

Generalizability Cross-domain fake news detection remains a significant challenge, with many studies exploring how to adapt models to diverse and evolving content. Han *et al.* (2020) treat the task as continual learning, where a model learns incrementally across domains based on dissemination patterns. However, their approach assumes a sequential domain arrival and requires prior domain knowledge, both of which are unrealistic in practice. Nan *et al.* (2021) address domain variation through expert-informed domain gating, but their method relies heavily on large annotated datasets. Expanding on this, Liang *et al.* (2022) and Wang *et al.* (2023) introduce models that apply multi- or soft-domain labels, though each news item is ultimately treated as belonging to a single domain. Ma *et al.* (2024) follow a similar line, assigning rigid domain-specific labels. Recent studies (Li et al. 2023; Lin et al. 2022) collect and annotate data from emerging domains and fine-tune pre-trained models, but they remain constrained by the high cost and limited availability of labelled data. To overcome this, Rastogi *et al.* (2021) propose an adaptive method that dynamically selects the most appropriate model per domain, reducing dependency on labelled examples while enhancing generalizability across varied content.

Early Detection Recent research on early fake news detection focuses on minimizing their spread by leveraging limited initial propagation data. Silva *et al.* (2021) reconstruct full propagation trajectories from early-stage signals, using a hierarchy-based attention mechanism to emphasize informative nodes and cascades. A common limitation in early detection approaches is their reliance on fixed, manually designed labelling functions (Li et al. 2021; Leite et al. 2023), which are labour-intensive to produce and limited in diversity. In response, more recent initiatives (Akdag and Cicekli 2024) propose a hybrid method that combines content- and model-based techniques to automatically generate labelling functions. This dual approach not only reduces manual effort but also enables more robust early detection.

Legal Initiatives

Privacy and Security: Privacy and security in fake news detection have received some attention. Rani *et al.* (2024) introduced a blockchain-based deep learning model to focus more on securing data records than protecting individual user privacy. Khullar *et al.* (2023) propose a federated learning framewok to enhance privacy and computational efficiency, performing well across both single- and multi-client environments. However, it falls short of addressing cross-platform deployment challenges across various online social networks. To more directly secure user data, Ali *et al.* (2022) combine federated learning with homomorphic encryption to enable encrypted fake news detection. Englander *et al.* (2024) take a different approach by employing lightweight cryptographic hash functions to preserve privacy, offering a more scalable alternative.

Ethical Initiatives

Explainability Explainability in fake news detection has evolved through multiple approaches. Traditional methods often use attention-based models (Yang et al. 2019; Sharma and Sharma 2021; Silva et al. 2021), which highlight relevant text features but offer only local, model-specific insights and struggle to provide holistic or multimodal explanations. Some of these models rely on social media comments (Shu et al. 2019) or incorporate user metadata and posting sequences (Ni, Li, and Kao 2021; Lu and Li 2020), though the reliability of attention as an explanation method is still debated (Jain and Wallace 2019). Ablation-based techniques (Qiao, Wiechmann, and Kerz 2020; Zhou and Zafarani 2019) assess feature importance by removing inputs but are often language-dependent. More interpretable alternatives like the Tsetlin Machine (Bhattarai, Granmo, and Jiao 2021) offer transparent, rule-based decisions but sacrifice deep contextual understanding. Model-agnostic post-hoc methods such as LIME (Ribeiro, Singh, and Guestrin 2016; Sallami and Aïmeur 2025b) and SHAP (Lundberg and Lee 2017) have gained popularity for their applicability. LIME has been used to generate instance-level explanations (Dua et al. 2023; Moe, Kundu, and Nguyen 2023; Joshi et al. 2023; Amri, Sallami, and Aïmeur 2021), while SHAP provides fine-grained feature attributions (Fu et al. 2023; Purificato et al. 2023; Upadhyay, Pasi, and Viviani 2023). However, SHAP's complexity and interpretability can pose challenges for non-expert users (Epstein et al. 2022; Kaur et al. 2020).

Fairness and Bias Mitigation Bias mitigation in fake news detection remains a central concern. To reduce political bias, Park *et al.* (2023) promote article-level over source-level labelling, while Raj *et al.* (2023) apply interpretability to mitigate bias. Park *et al.* (2022) propose a Reject Option Classifier to improve fairness without sacrificing accuracy. For entity bias, Murayama *et al.* (2021) mask names and locations using Wikidata and Zhu *et al.* (2022) use a dual adversarial learning framework to debias label-content interactions. Domain bias initiatives include Kato *et al.* (2022), who mask noun phrases with limited success and Liu *et al.* (2024), who eliminate domain bias via adversarial learning. Dataset bias is addressed through counterfactual reasoning (Wu et al. 2022), data augmentation methods like CrossAug (Lee et al. 2021) and the ReW model (Schuster et al. 2019), which minimizes bias from skewed linguistic patterns. Finally, Sallami and Aïmeur (2025a) focus on gender bias, highlighting its impact on model performance and proposing targeted mitigation strategies to improve gender fairness.

Philanthropic Initiatives

Even though existing AI models are ostensibly designed to aid society in combating fake news, most remain confined to lab environments and are not accessible to the general public. While some applications—such as browser extensions—have been developed to assist everyday users (Sallami and Aïmeur 2025c), they still inherit the core limitations discussed earlier. There is a notable absence of frame-

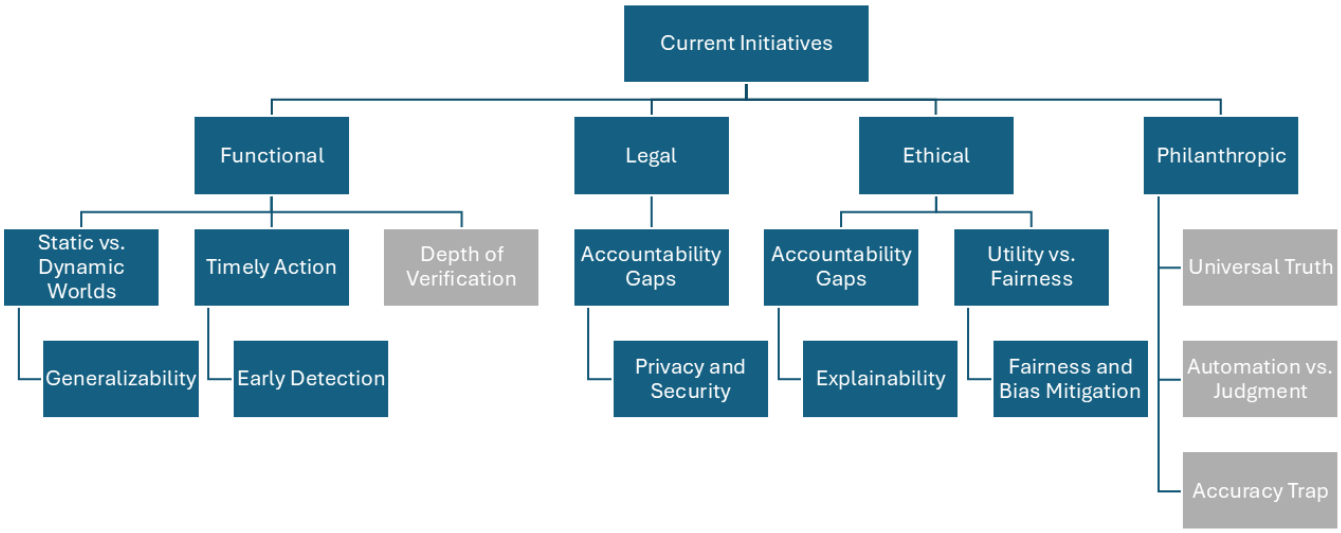

Figure 4: Current initiatives in fake news detection, structured according to the four tiers of the SRAI framework. Each path illustrates a key challenge (middle row) and the associated computational initiatives (bottom row). Gray shaded boxes indicate areas where significant research gaps remain.

works centred on universal truth and a careful balance between automation and human judgement. These omissions reveal a significant gap: the lack of philanthropic approaches that prioritize public trust and long-term societal resilience.

Challenges and Limitations

Despite the breadth of ongoing research, several limitations persist. Functionally, in terms of generalizability, most models assume that each news item belongs to a single domain, which fails to capture the semantic complexity of real news articles and often results in information loss. A gap remains in the lack of real-time model updating to adapt to emergent content. Similarly, approaches to early detection aim to identify fake news at the early stages of dissemination, yet most of these systems are evaluated only in controlled environments, making their real-world effectiveness uncertain.

From a legal standpoint, existing initiatives focus predominantly on privacy and data protection, often through federated learning and encryption-based methods. While these approaches contribute to securing user data, they do little to address the broader legal challenges of accountability. Legal infrastructures for fake news detection remain underdeveloped and the absence of enforceable accountability mechanisms presents a significant gap.

Ethically, most explainability methods are limited to local, model-specific outputs rather than global, user-accessible reasoning. Although model-agnostic techniques attempt to offer broader applicability, they are often too complex for non-expert users to interpret (Epstein et al. 2022). This highlights a major gap in general user-level interpretability. Furthermore, research on bias mitigation has centred on political biases, with insufficient attention given to other forms, such as those related to race or socioeconomic status. As a result, fairness interventions remain fragmented and do not adequately address the full spectrum of sociocultural inequalities. Most critically, the philanthropic dimension of fake news detection, concerned with civic trust, remains underdeveloped.

Future Directions

Beyond addressing the challenges presented in the previous section, we outline several future directions:

(1) Beyond Isolated Trade-offs A critical direction for future research is to move beyond the common assumption that SRAI principles—such as fairness, interpretability and privacy—must inherently conflict with utility. While these trade-offs have often been treated as fixed and unavoidable, recent findings suggest they are context-dependent and reconcilable. For instance, fairness and utility can align when datasets are balanced appropriately (Cheng and Liu 2023) and interpretability need not severely compromise performance if achieved through thoughtful model design (Assis, Dantas, and Andrade 2025). Moreover, current work addresses these trade-offs in isolation—optimizing for fairness or interpretability or privacy, but rarely all at once. A promising future direction is to explore frameworks that support multi-objective optimization, enabling models to be fair, explainable and secure simultaneously.

(2) Human in the loop A key future direction in fake news detection is the integration of human-in-the-loop (HITL) methods to enhance the contextual, ethical and adaptive capabilities of AI systems. Automated models often lack the flexibility to handle high-stakes or ambiguous cases, particularly when fairness thresholds vary across subgroups or when emerging fake news bypasses known detection patterns. Incorporating human oversight allows for more nuanced decision-making, enabling systems to better calibrate

fairness cutoffs and apply ethical criteria that are context-specific. In fast-changing information environments, HITL is also crucial for supporting active learning pipelines, where human annotators can rapidly label novel or complex fake news cases. Beyond annotation, future systems should emphasize controllability, the capacity to guide model behaviour through direct human input. Reinforcement learning from human feedback offers a promising mechanism for aligning system behaviour with social expectations.

(3) Embracing Interdisciplinary Collaboration A critical future direction for advancing fake news detection lies in fostering interdisciplinary collaboration. The challenges posed by fake news are multifaceted and demand a mosaic of knowledge. Addressing these challenges requires the combined expertise of technologists, ethicists, social scientists, legal scholars, journalists and media researchers. For instance, understanding how fake news spreads, how it influences public trust and how regulatory frameworks apply across platforms cannot be tackled by technical solutions alone. Interdisciplinary collaboration enables the design of detection systems that are not only accurate but also socially grounded, legally compliant and culturally sensitive. Future research should prioritize collaborative models that integrate diverse forms of expertise early in the development pipeline to ensure that fake news detection tools are aligned with the broader values and realities of the societies they serve.

(4) It Is Not "Accuracy vs. SRAI" A key future direction in fake news detection is to move beyond the false dichotomy of accuracy versus social responsibility. Future systems must aim to optimize both. Accuracy remains critical, particularly in real-world scenarios. For instance, models must be able to perform reliably on emerging fake news. At the same time, we must acknowledge that automation alone may not always be the most effective solution.

Conclusion

In this paper, we have argued that asking AI to merely detect fake news is both insufficient and short-sighted. To address the deeper complexities of this task, we examined the unique challenges it presents and proposed a shift toward SRAI as a guiding framework. Our contribution includes mapping these challenges onto the SRAI pyramid, offering a structured lens for addressing them. We also reviewed emerging initiatives, noted key limitations and suggested directions for building detection systems that are not only accurate but also socially accountable and publicly trustworthy. We hope this work contributes to shifting the field toward confronting these broader challenges and inspires future research to integrate ethical, legal and societal considerations alongside technical innovation. What we have outlined here represents only the tip of the iceberg; a wide range of urgent and compelling research questions remain for the community to explore, understand and address in pursuit of a more responsible and beneficial AI future. Ultimately, AI must earn public trust, not just pass technical tests.

Acknowledgments

The authors are listed in alphabetical order, following the convention in the second author's field. Dorsaf Sallami conducted the literature review, synthesized the findings, and prepared the initial manuscript under the supervision and guidance of Esma Aïmeur. Gilles Brassard reviewed and refined the work into its final form.

References

Akdag, S. H.; and Cicekli, N. K. 2024. Early detection of fake news on emerging topics through weak supervision. *Journal of Intelligent Information Systems*, 62(5): 1263–1284.

Ali, H.; Javed, R. T.; Qayyum, A.; AlGhadhban, A.; Alazmi, M.; Alzamil, A.; AlUtaibi, K.; and Qadir, J. 2022. SPAMDaS: Secure and privacy-aware misinformation detection as a service. *Authorea Preprints*.

Allein, L.; Moens, M.-F.; and Perrotta, D. 2023. Preventing Profiling For Ethical Fake News Detection. *Information Processing & Management*, 60(2): 103206.

Amri, S.; Boleilanga, H.-C. M.; and Aïmeur, E. 2023. ExFake: Towards an Explainable Fake News Detection Based on Content and Social Context Information. In *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*, 01–08. IEEE.

Amri, S.; Sallami, D.; and Aïmeur, E. 2021. Exmulf: an explainable multimodal content-based fake news detection system. In *International Symposium on Foundations and Practice of Security*, 177–187. Springer.

Assis, A.; Dantas, J.; and Andrade, E. 2025. The performance-interpretability trade-off: a comparative study of machine learning models. *Journal of Reliable Intelligent Environments*, 11(1): 1.

Bengio, Y.; Lecun, Y.; and Hinton, G. 2021. Deep learning for AI. *Communications of the ACM*, 64(7): 58–65.

Berti, L.; Giorgi, F.; and Kasneci, G. 2025. Emergent Abilities in Large Language Models: A Survey. *arXiv preprint arXiv:2503.05788*.

Bhattarai, B.; Granmo, O.-C.; and Jiao, L. 2021. Explainable tsetlin machine framework for fake news detection with credibility score assessment. *arXiv preprint arXiv:2105.09114*.

Capel, T.; and Brereton, M. 2023. What is human-centered about human-centered AI? A map of the research landscape. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–23.

Cheng, L.; and Liu, H. 2023. *Socially Responsible AI: Theories and Practices*. World Scientific.

Cheng, L.; Varshney, K. R.; and Liu, H. 2021. Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71: 1137–1181.

Dahlin, E. 2021. Mind the gap! On the future of AI research. *Humanities and Social Sciences Communications*, 8(1): 1–4.

Deepak, P. 2020. Ethical considerations in data-driven fake news detection. *Journal: Data Science for Fake News The Information Retrieval Series*, 205–232.

Dou, Y.; Shu, K.; Xia, C.; Yu, P. S.; and Sun, L. 2021. User preference-aware fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2051–2055.

Dua, V.; Rajpal, A.; Rajpal, S.; Agarwal, M.; and Kumar, N. 2023. I-flash: Interpretable fake news detector using lime and shap. *Wireless Personal Communications*, 131(4): 2841–2874.

Englander, K.; Samanthula, B. K.; and Dong, B. 2024. Towards a Privacy-Aware and Outsourced Fake News Detection Framework. In *2024 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 1–5. IEEE.

Epstein, Z.; Foppiani, N.; Hilgard, S.; Sharma, S.; Glassman, E.; and Rand, D. 2022. Do explanations increase the effectiveness of AI-crowd generated fake news warnings? In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 183–193.

Fu, C.; Pan, X.; Liang, X.; Yu, S.; Xu, X.; and Min, Y. 2023. Feature drift in fake news detection: An interpretable analysis. *Applied Sciences*, 13(1): 592.

Han, Y.; Karunasekera, S.; and Leckie, C. 2020. Graph neural networks with continual learning for fake news detection from social media. *arXiv preprint arXiv:2007.03316*.

Hutchinson, B.; Rostamzadeh, N.; Greer, C.; Heller, K.; and Prabhakaran, V. 2022. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 1859–1876.

Jain, S.; and Wallace, B. C. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Joshi, G.; Srivastava, A.; Yagnik, B.; Hasan, M.; Saiyed, Z.; Gabralla, L. A.; Abraham, A.; Walambe, R.; and Kotecha, K. 2023. Explainable misinformation detection across multiple social media platforms. *IEEE Access*, 11: 23634–23646.

Kato, S.; Yang, L.; and Ikeda, D. 2022. Domain bias in fake news datasets consisting of fake and real news pairs. In *2022 12th International Congress on Advanced Applied Informatics (IIAI-AAI)*, 101–106. IEEE.

Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; and Wortman Vaughan, J. 2020. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–14.

Kemp, P. L.; Alexander, T. R.; and Wahlheim, C. N. 2022. Recalling fake news during real news corrections can impair or enhance memory updating: The role of recollection-based retrieval. *Cognitive Research: Principles and Implications*, 7(1): 85.

Khullar, V.; and Singh, H. P. 2023. f-FNC: Privacy concerned efficient federated approach for fake news classification. *Information Sciences*, 639: 119017.

Kuntur, S.; WrĂłblewska, A.; Paprzycki, M.; and Ganzha, M. 2024. Fake News Detection: It's All in the Data! *arXiv preprint arXiv:2407.02122*.

Lakkaraju, H.; Kleinberg, J.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2017. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 275–284.

Lebovitz, S.; Levina, N.; and Lifshitz-Assaf, H. 2021. IS AI GROUND TRUTH REALLY TRUE? THE DANGERS OF TRAINING AND EVALUATING AI TOOLS BASED ON EXPERTS'KNOW-WHAT. *MIS quarterly*, 45(3).

Lee, M.; Won, S.; Kim, J.; Lee, H.; Park, C.; and Jung, K. 2021. Crossaug: A contrastive data augmentation method for debiasing fact verification models. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3181–3185.

Leite, J. A.; Razuvayevskaya, O.; Bontcheva, K.; and Scarton, C. 2023. Detecting misinformation with LLM-predicted credibility signals and weak supervision. *arXiv preprint arXiv:2309.07601*.

Lepri, B.; Oliver, N.; and Pentland, A. 2021. Ethical machines: The human-centric use of artificial intelligence. *IScience*, 24(3).

Li, J.; Wang, L.; He, J.; Zhang, Y.; and Liu, A. 2023. Improving rumor detection by class-based adversarial domain adaptation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6634–6642.

Li, Y.; Lee, K.; Kordzadeh, N.; Faber, B.; Fiddes, C.; Chen, E.; and Shu, K. 2021. Multi-source domain adaptation with weak supervision for early fake news detection. In *2021 IEEE International Conference on Big Data (Big Data)*, 668–676. IEEE.

Liang, C.; Zhang, Y.; Li, X.; Zhang, J.; and Yu, Y. 2022. FuDFEND: fuzzy-domain for multi-domain fake news detection. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 45–57. Springer.

Lin, H.; Ma, J.; Chen, L.; Yang, Z.; Cheng, M.; and Chen, G. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. *arXiv preprint arXiv:2204.08143*.

Liu, Q.; Wu, J.; Wu, S.; and Wang, L. 2024. Out-of-distribution Evidence-aware Fake News Detection via Dual Adversarial Debiasing. *IEEE Transactions on Knowledge and Data Engineering*.

Lu, Y.-J.; and Li, C.-T. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Ma, X.; Zhang, Y.; Ding, K.; Yang, J.; Wu, J.; and Fan, H. 2024. On Fake News Detection with LLM Enhanced Semantics Mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 508–521.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Moalla, H.; Abid, H.; Sallami, D.; Aïmeur, E.; and Hamed, B. B. 2025. Exploring the Power of Dual Deep Learning for Fake News Detection. *Informatica*, 48(4).

Moe, L.; Kundu, A.; and Nguyen, U. T. 2023. A BERT-Based Explainable System for COVID-19 Misinformation Identification.

Morales-Forero, A.; Bassetto, S.; and Coatanea, E. 2023. Toward safe AI. *AI & SOCIETY*, 38(2): 685–696.

Murayama, T.; Wakamiya, S.; and Aramaki, E. 2021. Mitigation of diachronic bias in fake news detection dataset. *arXiv preprint arXiv:2108.12601*.

Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; and Li, J. 2021. MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3343–3347.

Narayanan, A. 2019. How to recognize AI snake oil. *Arthur Miller lecture on science and ethics*.

Ni, S.; Li, J.; and Kao, H.-Y. 2021. MVAN: Multi-view attention networks for fake news detection on social media. *IEEE Access*, 9: 106907–106917.

Park, J.; Ellezhuthil, R.; Arunachalam, R.; Feldman, L.; and Singh, V. 2022. Toward fairness in misinformation detection algorithms. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*, volume 16.

Park, J.; Ellezhuthil, R. D.; Isaac, J.; Mergerson, C.; Feldman, L.; and Singh, V. 2023. Misinformation detection algorithms and fairness across political ideologies: The impact of article level labeling. In *Proceedings of the 15th ACM Web Science Conference 2023*, 107–116.

Purificato, E.; Shahania, S.; Thiel, M.; and De Luca, E. W. 2023. FACADE: Fake Articles Classification and Decision Explanation. In *European Conference on Information Retrieval*, 294–299. Springer.

Qiao, Y.; Wiechmann, D.; and Kerz, E. 2020. A language-based approach to fake news detection through interpretable features and BRNN. In *Proceedings of the 3rd international workshop on rumours and deception in social media (RDSM)*, 14–31.

Rafieian, O.; and Yoganarasimhan, H. 2023. AI and personalization. *Artificial Intelligence in Marketing*, 77–102.

Raj, C.; Mukherjee, A.; and Zhu, Z. 2023. True and Fair: Robust and Unbiased Fake News Detection via Interpretable Machine Learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 962–963.

Rani, P.; and Shokeen, J. 2024. FNNet: A secure ensemble-based approach for fake news detection using blockchain. *The Journal of Supercomputing*, 80(14): 20042–20079.

Rastogi, S.; Gill, S. S.; and Bansal, D. 2021. An adaptive approach for fake news detection in social media: single vs cross domain. In *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, 1401–1405. IEEE.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Sahoo, S. R.; and Gupta, B. B. 2021. Multiple features based approach for automatic fake news detection on social networks using deep learning. *Applied Soft Computing*, 100: 106983.

Sallami, D.; and Aïmeur, E. 2024. Fairframe: a fairness framework for bias detection and mitigation in news. *AI and Ethics*, 1–17.

Sallami, D.; and Aïmeur, E. 2025a. Does Gender Matter? Examining and Mitigating Gender Bias in Fake News Detection. In *International Symposium on Foundations and Practice of Security*, 249–266. Springer.

Sallami, D.; and Aïmeur, E. 2025b. Explain further: multi-level explanations for fake news detection using large language models. *International Journal of Information Technology*, 1–8.

Sallami, D.; and Aïmeur, E. 2025c. Exploring beyond detection: a review on fake news prevention and mitigation techniques. *Journal of Computational Social Science*, 8(1): 1–38.

Sallami, D.; Gueddiche, A.; and Aïmeur, E. 2023. From hype to reality: Revealing the accuracy and robustness of transformer-based models for fake news detection.

Schuster, T.; Shah, D. J.; Yeo, Y. J. S.; Filizzola, D.; Santus, E.; and Barzilay, R. 2019. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*.

Sharma, D. K.; and Sharma, S. 2021. Comment filtering based explainable fake news detection. In *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020*, 447–458. Springer.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 395–405.

Silva, A.; Han, Y.; Luo, L.; Karunasekera, S.; and Leckie, C. 2021. Propagation2Vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58(5): 102618.

Singh, R.; Vatsa, M.; and Ratha, N. 2021. Trustworthy AI. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data (8th ACM IKDD CODS & 26th COMAD)*, 449–453.

Thiebes, S.; Lins, S.; and Sunyaev, A. 2021. Trustworthy artificial intelligence. *Electronic Markets*, 31: 447–464.

Upadhyay, R.; Pasi, G.; and Viviani, M. 2023. Leveraging Socio-contextual Information in BERT for Fake Health News Detection in Social Media. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*, 38–46.

Wang, D.; Zhang, W.; Wu, W.; and Guo, X. 2023. Soft-label for multi-domain fake news detection. *IEEE Access*.

Wu, J.; Guo, J.; and Hooi, B. 2024. Fake News in Sheep's Clothing: Robust Fake News Detection Against LLM-Empowered Style Attacks. In *Proceedings of the 30th*

ACM SIGKDD conference on knowledge discovery and data mining, 3367–3378.

Wu, J.; Liu, Q.; Xu, W.; and Wu, S. 2022. Bias mitigation for evidence-aware fake news detection by causal intervention. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2308–2313.

Yang, F.; Pentyala, S. K.; Mohseni, S.; Du, M.; Yuan, H.; Linder, R.; Ragan, E. D.; Ji, S.; and Hu, X. 2019. Xfake: Explainable fake news detector with visualizations. In *The world wide web conference*, 3600–3604.

Zhou, X.; and Zafarani, R. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter*, 21(2): 48–60.

Zhu, Y.; Sheng, Q.; Cao, J.; Li, S.; Wang, D.; and Zhuang, F. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2120–2125.