STRENGTHENING IMPARTIALITY FOR THE FACT-CHECKING

LATIF'S RECOMMENDATIONS

Custureri S., De Rosa S., Nicolai A.
T6 Ecosystems

Table of Content

European | MEDIA AND INFORMATION | Fund

Managed by
Calouste Gulbenkian Foundation

Executive Summary

The role of Artificial Intelligence (AI) in combating misinformation is evolving rapidly. AI systems have the capacity to streamline fact-checking operations and counter disinformation, but they also present risks, especially in terms of bias and transparency. As AI becomes increasingly integrated into the media sector, questions around governance and regulation become critical.

The LATIF project uses the Analysis of Competing Hypotheses (ACH) methodology to mitigate these issues, providing tools to enhance impartiality in fact-checking. This policy brief will explore the implications of AI governance, with a particular focus on the EU AI Act, global best practices, regulatory gaps, and the ethical responsibilities involved in the development and deployment of AI in media.

Project Overview and Policy Brief's Purpose

The LATIF (Leveraging Argument Technology for Impartial Fact-checking) project tackles one of the most pressing challenges in AI-assisted fact-checking: cognitive bias.

Cognitive biases, such as confirmation bias, occur when individuals - whether human fact-checkers- or AI systems (Gallimore & Lee, 2022) - favour information that supports their existing beliefs or hypotheses while dismissing contradictory evidence (Lewandowsky, Ecker, & Cook, 2017). In the context of fact-checking, these biases can undermine the accuracy and objectivity of verification processes, especially when dealing with disinformation campaigns where partial truths are strategically mixed with misleading information (Masotina, 2024; Park, Kang, & Cha, 2021). Such biases make it challenging to separate facts from falsehoods, thereby affecting both the detection and correction of misinformation (Santos, 2023; Karduni, 2019).

To mitigate these risks, **LATIF incorporates the Analysis of Competing Hypotheses (ACH) methodology**, a structured analytical technique initially developed for intelligence analysis. ACH encourages a systematic evaluation of multiple hypotheses or explanations for a given set of facts. By applying this method in fact-checking, the LATIF project promotes a more comprehensive analysis, ensuring that fact-checkers (both human and AI) are not prematurely dismissing alternative explanations in favour of familiar narratives or biases.

The ACH methodology is particularly effective when integrated with AI, as it leverages AI's capacity to process large amounts of information quickly and efficiently (Keding, 2021; Farah et al., 2023; Pereira et al., 2023).

In traditional fact-checking, human analysts may struggle with the cognitive load required to assess multiple competing explanations. AI tools can automate this process, enabling fact-checkers to consider a broader array of viewpoints without becoming overwhelmed by the data volume. In modern disinformation campaigns, adversaries often employ sophisticated techniques to blend factual information with false or misleading content, making it difficult for fact-checkers to discern the truth. Selective use of facts, combined with emotive framing, can easily mislead even experienced analysts. LATIF addresses this challenge by encouraging fact-checkers to remain open to alternative viewpoints, ensuring that disinformation strategies designed to exploit cognitive shortcuts are less effective.

For instance, disinformation narratives often exploit several biases, where early information (even if incorrect) disproportionately influences subsequent judgments. On the topic, the project has actively disseminated its research through various channels, including participation in sectoral events[1], blog posts, reports, and surveys. Key publications cover topics such as cognitive bias in fact-checking, disinformation threats to European democracy, and a cross-national analysis of fact-check dissemination on social media (Musi, Masotina, Federico, & Yates, 2023; Musi, Masotina, & Yates, 2024). LATIF has organised events like focus groups and surveys to gather input from professional and sectoral stakeholders and improve its tools, aiming for more reliable and transparent fact-checking processes, and focus its activity on academic and public outreach through talks and publications. Publications addressed critical issues such as the spread of fact-checking on Twitter and cognitive biases in the fact-checking process[2]. Recent key events include invited talks and presentations, such as a workshop on "Mis-Disinformation and Young People" and a lecture at the Arglab Research Colloquium in Lisbon on impartial fact-checking strategies. LATIF research has also been presented at international conferences, including Dubrovnik Media Days and the EDMO Scientific Conference. Additionally, a Chrome extension developed by LATIF will be demonstrated at the ESRC Festival of Social Sciences. These engagements reflect the project's dedication to enhancing critical media literacy and improving fact-checking through practical applications and public outreach.

In essence, the LATIF project seeks to create a more reliable, bias-resistant fact-checking ecosystem where both AI and human fact-checkers work in tandem to assess information from multiple angles. The ACH methodology ensures that AI systems remain impartial, transparent, and capable of considering a wide range of potential explanations, leading to more accurate and trustworthy fact-checking outcomes. In a world increasingly shaped by disinformation, these advances represent a critical step forward in the fight for media integrity and truth.

The rapid deployment of AI tools in the fact-checking ecosystem has raised significant questions regarding governance and regulation. AI systems can process vast amounts of data, often automating parts of the verification process that were previously time-consuming for human fact-checkers. However, as AI becomes more central to this process, the risk of bias - both in the data used to train AI and in the algorithms themselves - grows. This policy brief examines how policymakers can ensure that AI tools in the fact-checking space operate with transparency, fairness, and accountability.

1) Invited talks/presentations include: Workshop on "Mis-Disinformation and Young People: Developing Strategies for Critical Media Literacy" at Cumberland Lodge (18-19 March 2024); Lecture at the Arglab Research Colloquium, IFILNOVA (15 September 2023), titled "Leveraging Argumentation for Impartial Fact-checking"; Paper Presentation When fact-checks go viral: a cross-national analysis of the dissemination of European fact-checkers on Twitter at the ISSA conference (4-7 July 2023); the same paper presented at Dubrovnik Media Days 2023 (29-30 November 2023); LATIF project presented at the EDMO Scientific Conference 2024 (26-27 January 2024).
2) For more information visit: https://latifproject.eu/

The Eu AI Act and its Relevance to the Media Sector and Fact-Checking

The **EU AI Act**[3] marks a significant development in the governance of AI technologies, especially in high-risk applications. Within the media sector, **fact-checking activities themselves are not explicitly classified as high-risk under the EU AI Act**. Instead, the Act focuses on AI applications that have more direct consequences for fundamental rights and safety. The EU AI Act identifies high-risk AI systems based on specific sectors (like healthcare, transportation, education, law enforcement, etc.) or their significant impact on individual's rights. High-risk AI systems include those used in critical infrastructure, biometric identification, recruitment, and law enforcement, among others.

While fact-checking could fall under broader concerns about disinformation or AI systems affecting democratic processes, **it is not directly categorised as high-risk** in the same way that AI in biometric surveillance or hiring processes is. However, AI systems used in media, content moderation, and information verification (fact-checking) could still be subject to scrutiny if they have a substantial impact on fundamental rights, depending on their use and implementation.

In summary, **fact-checking AI systems are not automatically classified as high-risk under the EU AI Act**, but they could come under regulatory oversight if they are used in ways that significantly affect rights or public discourse.

The Act introduces stringent requirements aimed at mitigating risks of bias, ensuring transparency, and enforcing accountability in AI operations. A key pillar of the EU AI Act is the **requirement for human oversight**. AI tools used in media fact-checking cannot operate entirely independently; human fact-checkers must always have the ability to intervene, particularly when AI-generated decisions may be ambiguous, potentially biassed, or harmful. This provision ensures that AI systems do not make unchecked decisions that could propagate disinformation or influence public opinion in unintended ways.

The **Act's transparency provisions** go further, requiring developers of AI systems to provide detailed documentation explaining how the system operates, including the sources of data and the logic behind AI-driven decisions. This ensures that the decision-making processes are traceable and open to scrutiny, allowing users, fact-checkers, and regulatory authorities to understand how conclusions are drawn.

3) See: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. *Adopted text here: https://www.europarl.europa.eu/doceo/document/TA-9-2024-03-13-TOC_EN.html.*

This is vital in maintaining public trust in AI, as decisions made by opaque systems could otherwise exacerbate the spread of false and misleading contents, prone to pollute the public discourse on important topics. In addition, **AI-generated content**, such as deepfakes or text produced by generative models, must be labelled clearly to indicate its artificial origin. The labelling is crucial to avoid confusing AI-generated or manipulated content with authentic information, particularly in contexts where public interest is at stake. This mandatory labelling of manipulated content aligns with broader efforts within the AI Act to safeguard the media landscape against the misuse of generative AI. The EU AI Act introduces a comprehensive framework for **post-market monitoring**, placing a responsibility on providers to continuously track the performance of AI systems after they have been deployed. This includes identifying and addressing biases, errors, or other risks that may emerge over time, ensuring that systems remain compliant with the stringent standards set forth in the regulation. For AI systems used in fact-checking, this kind of ongoing scrutiny is essential to ensure that the tool continues to provide accurate and unbiased outputs, even as the information landscape evolves. Data governance is another critical component of the Act, particularly regarding the quality and representativeness of the datasets used to train and test AI systems. To prevent the introduction of bias or error, the Act requires that datasets be free from significant flaws and must be continuously audited to ensure their suitability for high-risk applications. This provision directly addresses one of the major challenges in AI development for media and fact-checking - ensuring that training data accurately reflects the diverse social, cultural, and political environments that the system operates within.

The relevance of the EU AI Act to the **media sector and fact-checking** cannot be overstated. AI tools have become indispensable in managing the sheer volume of content that needs verification in today's information landscape. These technologies, while powerful, come with the inherent risk of being exploited for the creation of false or misleading content, which can easily go viral in digital environments. The EU AI Act's provisions ensure that such tools are used responsibly. The requirement for human oversight ensures that AI tools enhance, rather than replace, the role of human judgement in verifying facts and countering disinformation.

Moreover, the Act recognizes **the importance of ethical AI development**, in this way encouraging media organisations to adopt AI technologies that not only meet regulatory standards but also align with journalistic ethics. These measures reflect a broader European strategy to foster innovation while protecting the fundamental rights of individuals, ensuring that AI tools do not compromise the quality and trustworthiness of information.

One of the most forward-thinking aspects of the EU AI Act is its adaptability to **future technological developments**. As AI systems, particularly general-purpose AI models, continue to evolve, the Act sets up mechanisms for regular updates and evaluations to ensure that its regulations keep pace with technological advances.

This future-proofing approach ensures that the legal framework remains flexible enough to address emerging risks, particularly as AI models grow more complex and capable[4]. The AI Act creates a flexible legal framework that can quickly adapt to technological advances. It sets general requirements, leaving the specifics to industry standards, allowing for customization across use cases. The Act can be updated through delegated acts, such as revising high-risk AI use cases. Regular evaluations ensure it remains up-to-date and responsive to new challenges.

Overall, the EU AI Act provides a robust framework for regulating AI, with a possibility of extensive application to the media sector, ensuring that these technologies are used responsibly to enhance the accuracy and transparency of fact-checking processes. By enforcing human oversight, transparency, and ongoing monitoring, the Act positions Europe as a global leader in creating a safe, ethical AI ecosystem that fosters public trust while mitigating the risks of disinformation

4) See "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (AI ACT)", namely the following articles: Clause (12); Clause (101); Clause (174); Clause (179); Article 6; Article 51; Article 53; Article 56.

Regulatory Challenges in AI for Fact-Checking and Tackling Disinformation

AI systems used for fact-checking face significant regulatory challenges, particularly when it comes to mitigating bias and addressing disinformation. These systems must navigate not only factual inconsistencies but also **cultural and linguistic nuances**, which adds complexity to their governance (NIST, 2021; Santos, 2023).

One of the major challenges in the regulatory landscape is ensuring that biases embedded in AI models - often a result of the datasets they are trained on - do not perpetuate harmful stereotypes or disinformation (RAND, 2021; NIST, 2021).

Generative AI, for instance, has made significant advancements in fact-checking, but performance discrepancies remain between languages. A study by the Reuters Institute (2023) shows that AI systems perform far better with well-documented languages like English but tend to struggle with smaller or less-represented languages such as Swahili or Georgian. This poses a serious challenge for global fact-checking efforts, particularly in regions where misinformation spreads through these underrepresented languages.

The difficulty in navigating these complexities highlights the importance of diverse and representative training datasets. However, the regulatory frameworks in place often focus on transparency and accountability, while neglecting to enforce measures that tackle bias at its root - within the data itself (RAND, 2021; NIST, 2021).

One of the most pressing concerns stemming from this issue is cognitive bias in AI-assisted decision-making, particularly in fact-checking. Cognitive bias remains a critical issue in AI-assisted decision-making, especially in fact-checking. AI systems learn from the datasets they are trained on, and these datasets often reflect societal biases, leading to skewed decision-making. This becomes problematic when AI systems are deployed in fact-checking, where neutrality and accuracy are essential. For example, **confirmation bias** - the tendency to prioritise information that aligns with pre-existing beliefs - can easily be replicated by AI if it is not carefully designed to avoid such pitfalls. In addition, **contextual understanding** is another major challenge for AI. AI systems often lack the ability to grasp the subtleties of language and culture, making it difficult for them to fully interpret the context of disinformation. This failure can lead to inaccurate fact-checking, especially in linguistically and culturally diverse regions. As a result, human oversight is crucial to ensure that AI systems do not exacerbate existing biases or produce misleading results.

However, while cognitive bias in AI remains a critical issue, another key challenge lies in regulatory gaps that fail to fully ensure AI impartiality. Despite advancements in AI governance, significant **regulatory gaps** persist, particularly in addressing **impartiality**. While current regulations often prioritise transparency and data governance, they fall short when it comes to addressing the biases embedded in AI datasets. Many AI tools still rely on datasets that are inherently biassed, which means that even when the decision-making process is transparent, the system may still produce biassed outcomes. A more **nuanced regulatory approach** is needed - one that not only mandates transparency but - also ensures that AI systems are trained on **diverse datasets** and are regularly audited to address biases. Moreover, regulations must require AI developers to continuously update and improve their systems to reflect the dynamic nature of cultural a

LATIF'S main recommendations

To conclude, the LATIF project has revised the current regulatory approach and has derived the main recommendation to allow a fair and transparent adoption of AI in relation to fact-checking tools. LATIF main recommendations are reported here.

01 Strengthening AI Governance for Fact-Checking Tools

To ensure the effectiveness and impartiality of AI tools in fact-checking, current regulatory frameworks need to extend beyond basic transparency and accountability measures. A primary recommendation is to mandate diversity in training datasets. This is crucial because many biases in AI systems arise from imbalanced datasets that fail to adequately represent different demographics or cultural contexts. When data from marginalised groups is underrepresented, AI systems can reinforce existing inequalities and produce biassed outputs. Diverse training data allows AI models to perform more equitably, ensuring that these systems can handle content from various regions and cultural backgrounds without defaulting to stereotypes or inaccuracies. However, achieving diversity in datasets is not without challenges. For instance, it can be difficult to gather representative data from certain groups or regions. Data augmentation and targeted data collection are methods that can help overcome this issue by artificially expanding datasets or focusing on under-represented groups during data collection processes. This approach mitigates the risks of underfitting, where AI models might fail when exposed to previously unseen data points, as highlighted in research from MIT and other studies on AI diversity.

Beyond dataset diversity, continuous auditing of AI systems is essential. Regular audits help identify and correct emerging biases, ensuring that fact-checking tools remain up to date with societal changes and do not perpetuate outdated or harmful assumptions. Auditing tools can be designed to track the fairness of AI systems by examining their outputs in real-world applications, such as monitoring for biassed decision-making against particular groups. Auditing processes should be integrated into AI governance as a mandatory practice, ensuring that biases are flagged and rectified before they affect end-users.

02 Enhancing Transparency and Accountability in AI Applications

Transparency and accountability remain two of the most critical challenges in AI fact-checking tools. Policymakers must implement strict guidelines that require full disclosure of how AI systems are used in fact-checking, including detailed explanations of the training datasets, algorithms, and decision-making processes. This level of transparency ensures that human fact-and the public can understand how an AI tool reaches its conclusions, preventing the spread of disinformation.

Human oversight should be integrated at all stages of the fact-checking process to ensure that AI outputs are reviewed and validated by experienced fact-checkers before they are disseminated. This human-in-the-loop approach is crucial for handling complex or ambiguous cases that require cultural sensitivity or deep contextual understanding that AI might not possess.

03 Supporting Ethical AI Development Aligned with EU Standards

Policymakers should also emphasise the ethical development of AI tools, aligning with frameworks like the EU AI Act. While the EU Act sets strong foundations for transparency and accountability, more needs to be done to ensure that AI systems in fact-checking adhere to strict ethical guidelines, focusing on non-discrimination, fairness, and respect for human rights. This includes setting ethical standards that govern the AI development process itself, ensuring that diverse voices are included in the creation and deployment of AI models.

Collaboration between AI developers, fact-checkers, and civil society organisations is necessary to design AI systems that serve the public interest. Involving a broad range of stakeholders will help ensure that these systems are not co-opted for corporate or political gain, but instead remain dedicated to supporting democratic discourse and accurate information dissemination.

04 The Role of Policy in Advancing Ethical AI for Media and Fact-Checking

Policymakers are crucial in shaping ethical AI, particularly in media and fact-checking, by setting regulations that promote transparency, fairness, and accountability. As AI increasingly impacts public discourse and media integrity, robust regulations are required to ensure responsible use. Cross-border cooperation is vital due to the global nature of disinformation and aligning AI policies across regions like the EU, North America and Asia can create unified standards. Overcoming mistrust between major AI powers, such as the US and China, is essential for cohesive governance.

Governments should also support AI research focused on ethical development, ensuring diverse training data, bias audits and algorithmic transparency. This reduces the risk of biassed systems and improves public trust by making AI decision-making processes clearer, particularly in fact-checking.

Furthermore, a multi-stakeholder approach is needed, involving tech companies, civil society, and international organisations. Initiatives like the AI Governance Alliance show how collaboration can create frameworks that respect privacy and human rights while combating disinformation. Policymakers must prioritise ethical AI in media through clear standards, global cooperation, and support for innovation to counter disinformation and enhance public discourse.

References

- Farah, J., et al. (2023). Explainable and interpretable artificial intelligence in medicine: A systematic review. Discover Artificial Intelligence. https://link.springer.com/article/10.1007/s00198-020-05333-1.
- Gallimore, R., & Lee, C. Z. (2022). Rolling in the deep of cognitive and AI biases. arXiv preprint. https://ar5iv.org/2407.21202.
- Karduni, A. (2019). Human-misinformation interaction: Understanding the interdisciplinary approach needed to computationally combat false information. arXiv preprint. https://arxiv.org/abs/1903.07136v1.
- Keding, C. (2021). Understanding the interplay of artificial intelligence and strategic management. Management Review Quarterly, 71(1), 91-134. https://link.springer.com/article/10.1007/s11301-021-00222-2.
- Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the post-truth era. Journal of Applied Research in Memory and Cognition, 6(4), 353-369. https://doi.org/10.1016/j.jarmac.2017.07.008.
- Masotina, M. (2024). Cognitive biases in news-making and fact-checking: A mixed methods approach. EDMO Policy Analysis. https://edmo.eu/publications/fact-checking-analysis.
- Musi, E., Masotina, M., Federico, L., & Yates, S. (2023). Impartiality and cognitive bias in the fact-checking process: an overview (ISBN: 978-88-947920-0-3).
- Musi, E., Masotina, M., & Yates, S. (2024). An Argumentative Approach to Map the Issue of Impartiality in Newsmaking and Factchecking. In Tenth Conference of the International Society for the Study of Argumentation (pp. 677-698). Leiden.
- NIST. (2021). NIST Proposes Approach for Reducing Risk of Bias in Artificial Intelligence. National Institute of Standards and Technology. https://www.nist.gov/news-events/news/2021/06/nist-proposes-approach-reducing-risk-bias-artificial-intelligence.
- Park, S., Kang, J. H., & Cha, M. (2021). Unexpected biases in online fact-checking: Evidence from politically diverse groups. Misinformation Review. https://misinforeview.hks.harvard.edu/wp-content/uploads/2021/01/park_unexpected_biases_online_fact_checking_20210127.pdf.
- Pereira, V., Hadjielias, E., Christofi, M., & Vrontis, D. (2023). A systematic literature review on the impact of artificial intelligence on workplace outcomes. Human Resource Management Review. https://link.springer.com/article/10.1007/s10098-023-00133-0.
- RAND. (2021). Towards an AI-Based Counter-Disinformation Framework. RAND Corporation. https://www.rand.org.
- Reuters Institute. (2023). AI and Language Bias: Discrepancies in AI Performance Across Languages. https://reutersinstitute.politics.ox.ac.uk.
- Santos, F. C. (2023). Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis. Journal of Media, 4(2), 679-687. https://doi.org/10.3390/journalmedia4020043.

LATIF

https://latifproject.eu

Contact Details:
info@t-6.it
www.t-6.it