vera.ai: VERification Assisted by Artificial Intelligence

Handbook on the legal and ethical obligations to developers and deployers of AI-based fact-checking tools

Sophia Wistehube, François Lavoir, Ausra Semeniene, Michele Evangelista, Lalya Gaye – European Broadcasting Union (EBU)

05 September 2025

This document provides an accessible guide for developers and deployers of AI systems to applicable legal and ethical requirements, especially under the EU AI Act, and with a particular focus on the development of AI-based tools for fact-checking and investigative journalism.

Keywords: AI systems, Act AI, data protection, ethics, fact-checking, journalism.

vera.ai is a Horizon Europe Research and Innovation Project co-financed by the European Union under Grant Agreement ID: 101070093, an Innovate UK grant 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00245.

Glossary

Abbreviation	Meaning
AI	Artificial Intelligence
BBC	British Broadcasting Corporation
CCTV	Closed-circuit television
CE	Conformité Européenne (In English: European Conformity)
DPIA	Data Protection Impact Assessment
DW	Deutsche Welle (In English: German Wave)
EBU	European Broadcasting Union
EDPB	European Data Protection Board
EU	European Union
GDPR	General Data Protection Regulation
GPAI	General-Purpose AI models
LLM	Large Language Model
MFA	Multi-Factor Authentication
PSM	Public Service Media
ZDF	Zweites Deutsches Fernsehen (In English: Second German Television)

Table of Contents

Executive Summary

This handbook aims to guide AI developers and deployers working on tools for fact-checking and investigative journalism through the complex landscape of legal and ethical requirements imposed by the EU Artificial Intelligence Act (AI Act) and related regulations, such as the General Data Protection Regulation (GDPR), as well as the ethical requirements that media organisations generally impose on their process of procuring AI-based tools. The document outlines the requirements that could apply specifically to AI-based tools for fact-checking and investigative journalism, which AI developers and deployers must meet to create or deploy these AI systems in a manner that is compliant, safe, ethically responsible and respectful of fundamental rights and EU values.

The EU AI Act introduces a risk-based approach, categorising AI systems into four levels: unacceptable risk, high risk, limited risk, and minimal risk. In general, the higher the risk, the more stringent the regulations. Obligations for providers and deployers of AI systems that are relevant for the tools in question include risk management, data quality, documentation, human oversight, and compliance with transparency obligations, as well as ongoing monitoring and incident reporting throughout the AI system's lifecycle. The AI Act also outlines specific exemptions, delays, and rules for general-purpose AI models.

Data protection requirements under GDPR play a central role for such tools as well, focusing on principles such as purpose limitation, data minimisation, transparency, and safeguarding sensitive data. Developers must ensure compliance when processing personal data, including conducting Data Protection Impact Assessments (DPIAs) for high-risk cases and preserving data subject rights.

Beyond the legal framework, the report emphasises ethical considerations, particularly for media organisations and journalists in their procurement of AI-based tools and systems. These include ensuring human oversight, transparency, explainability, accuracy, and privacy. Developers are encouraged to design tools that empower journalists, protect their data and sources, and minimise risks of biases or harm. Ethical AI development fosters trust, audience accountability, and journalistic integrity while maintaining compliance with regulations.

As this handbook focuses specifically on the obligations that could be particularly relevant for AI tools supporting fact-checking and investigative journalism, it does not provide a comprehensive overview of obligations under the EU AI Act or ethical considerations that could apply to all AI systems in general. Importantly, this handbook does not constitute an exhaustive compliance guide, nor does it represent legal advice. It is however a useful handbook that can serve as a foundational resource for the targeted AI developers and deployers to navigate regulatory compliance and ethical best practices.

1 Introduction

This handbook provides an accessible guide to the legal and ethical requirements for developers and deployers of AI systems, especially under the EU AI Act, and with a particular focus on the development of AI-based tools for fact-checking and investigative journalism. It outlines key obligations regarding the development and use of AI tools, focusing mainly on high-risk and general-purpose AI systems. Topics include the legal obligations under the EU AI Act, data protection under the General Data Protection Regulation (GDPR), and ethical considerations such as accuracy, human oversight, transparency, and privacy. Designed for AI developers and deployers, it serves as a helpful first reference for building and deploying AI fact-checking tools responsibly.

General disclaimer

Please note that these explanations are provided to support your understanding of EU law when developing and deploying AI systems. Be aware that this document does not constitute an exhaustive compliance guide, nor does it represent legal advice. Always make sure that you fully comply with all the relevant national and EU laws, including those not covered in this document.

2 Legal Obligations for AI Developers of AI-Based Fact-Checking Tools Under the AI Act

In recent years, the fast-paced progress of AI technologies and their increasing impact on society has led to the development of laws and regulations around the development and deployment of AI systems, both at national and European levels. It is essential for all AI developers and deployers – including those of AI-based fact-checking tools – to familiarise themselves with these new rules and to adhere to the obligations that come with them.

In this handbook, we specifically pick out those obligations that could be relevant to developers and deployers of AI-based tools for fact-checking and investigative journalism, or that could become relevant to them in the near future. As AI technology progresses very fast, we have probably included more cases than would be currently relevant. But having this broader understanding will also help guide responsible innovation in the future.

2.1 Overview

The European Union Artificial Intelligence Act (AI Act)[1] is a regulatory framework proposed by the European Commission to govern the development and use of artificial intelligence within the EU. It aims to ensure that AI systems are safe, respect fundamental rights, and align with EU values.

[1] AI Act: https://artificialintelligenceact.eu/ai-act-explorer/ (last accessed on 26.08.2025)

The Act classifies AI systems into different categories based on their risk levels – unacceptable risk, high risk, limited risk, and minimal risk – and sets requirements accordingly. The vast majority of AI systems are expected to be of low or minimal risk and will remain unregulated. AI systems deemed to pose limited risks will merely need to comply with transparency obligations. By contrast, high-risk AI systems must ensure additional safeguards, including risk mitigation, high level of robustness, security and accuracy, accountability and human oversight.

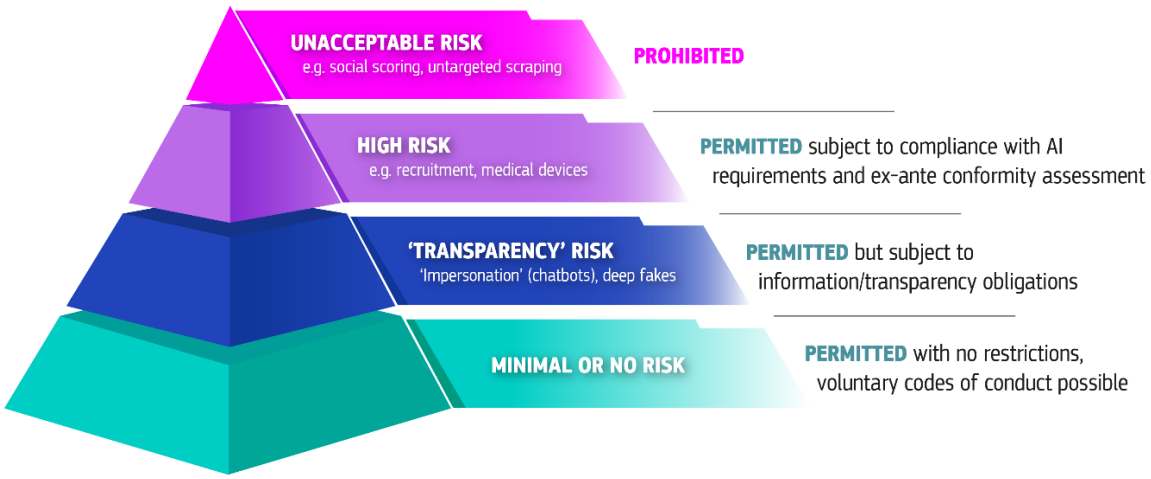

Figure 1 AI Act's risk-based approach[2].

AI-based tools for fact-checking and investigative journalism could fall into the high-risk category for instance if they use biometric data. In situations where AI-generated content could mislead individuals, these will likely need to comply with transparency obligations. Generative AI tools also need to make sure to respect copyright. The relevant obligations will be discussed in more detail in the following sections.

Other relevant EU rules besides the EU AI Act: Data Protection and Copyright

The AI Act is merely a product safety regulation, and it is merely one of the various EU legal instruments applicable to AI developers and AI tools. Other relevant laws for fact-checking tools developers include the GDPR[3] and the Directive on Copyright in the Digital Market (Copyright Directive[4]). Whenever you use personal data or copyrighted material, you must make sure to comply with the respective rules on data protection (more details in Section 3) and copyright (more details in Section 2.2.2).

[2] Source: European Commission https://digital-strategy.ec.europa.eu/en/factpages/ai-act/ (last accessed on 26.08.2025)
[3] GDPR: https://gdpr-info.eu/ (last accessed on 26.08.2025)
[4] https://eur-lex.europa.eu/eli/dir/2019/790/oj/eng (last accessed on 26.08.2025)

2.1.1 To whom does the EU AI Act apply?

The AI Act's regulations will apply to both public and private entities, regardless of whether they are located within or outside the EU, provided the AI system is made available on the EU market or its use has an impact on individuals within the EU.

The AI Act applies to different types of participants in the AI value chain, including providers, product manufacturers, deployers, authorised representatives, importers, and distributors.

However, the AI Act does not pertain to ordinary users, that is, deployers who are natural persons using AI systems solely in the context of personal, non-professional activities.

2.1.2 Who is a provider under the EU AI Act?

AI developers are generally referred to as *providers* in the EU AI Act. An AI provider is anyone, a person, company, or organisation, that creates an AI system or has one created for them and then offers it for use or sale, whether they charge for it or give it away for free. The term *provider* also includes organisations that develop AI systems for their own internal use. In these situations, the organisation acts as both the *provider* and the *deployer* of the AI system, so it must follow the rules for both roles.

> As developers, you are *providers* of AI systems, even if you only develop them for your own use.

2.1.3 Who is a deployer under the EU AI Act?

A *deployer* is anyone, e.g. a person, company, or organisation, that uses an AI system under their control, unless they are using it for personal, non-work-related activities. Here, this means news and fact-checking organisations and their employees working with these tools.

> If you use an AI system for your job, you are a *deployer*.

2.1.4 Can a deployer become a provider?

Deployers of AI systems can become providers in certain situations. They will be considered a provider of a high-risk AI system and must follow the relevant rules of the AI Act if they:

- Make major changes to a high-risk AI system already available on the market so that it still qualifies as high risk.
- Change the intended use of an AI system already on the market, including a general-purpose AI system, in a way that turns it into a high-risk AI system.
- Put their own name or trademark on a high-risk AI system that is already available on the market.

This rule ensures that if someone modifies a high-risk AI system or repurposes an AI system initially not meant for high-risk activities, they must still follow the stricter rules set for AI providers. It makes sure to close any loopholes, so all high-risk AI applications are fully regulated.

2.1.5 When do the rules of the EU AI Act start applying?

The AI Act will start applying two years after its entry into force, that is on 2 August 2026, with some exceptions for specific provisions:

- February 2025: Phasing out of prohibited systems.
- 2 August 2025: Obligations for general purpose AI become applicable.
- 2 August 2026: Transparency obligations become applicable.
- 2 August 2027: Obligations for high-risk systems become applicable.

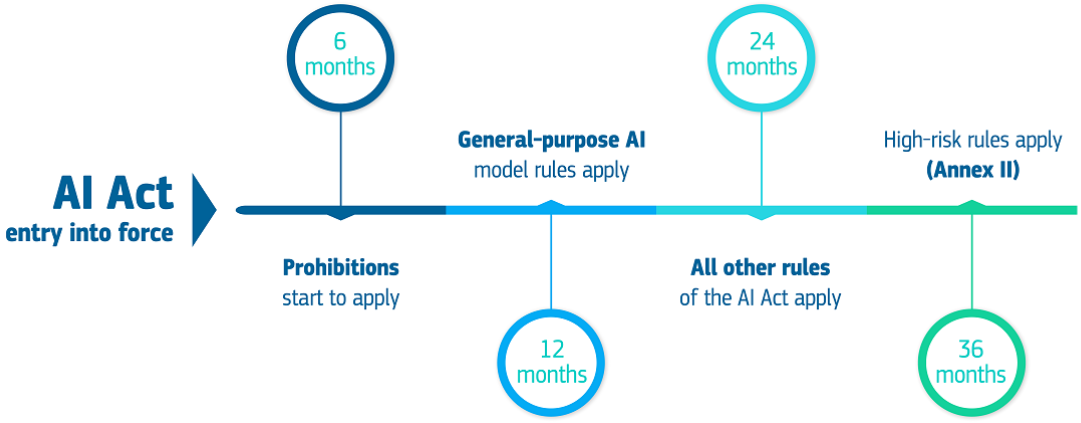

Figure 2 Timeline for the AI Act's entry into force[5].

However, as the next section will show, there are significant exceptions and delays for AI systems already on the market or that will be released within the first year of the AI Act's entry into force.

2.1.6 Do already-existing AI systems have to oblige with the rules?

The AI Act introduces significant exemptions and delays for AI systems that are already on the market or in use. Specifically, providers of general-purpose AI models (GPAIs) that are introduced until 2 August 2025 will only need to adhere to its regulations by 2 August 2027. This effectively means that for the next two years, current GPAIs will not be subject to regulation.

[5] Source: European Commission https://digital-strategy.ec.europa.eu/en/factpages/ai-act (last accessed on 26.08.2025)

Providers or deployers of high-risk AI systems placed on the market or in use until 2 August 2026 will only need to comply if those systems undergo significant design changes after that point. Otherwise, they will remain fully exempt.

A *significant change* is a change to an AI system that was not initially foreseen or planned, and thus was not accounted for in the conformity assessment of the AI system, or a change that results in a modification to the intended purpose of the system – such as e.g. a journalistic tool being used instead for law enforcement. However, changes occurring to the algorithm and the performance of AI systems which merely continue to 'learn' by automatically adapting how functions are carried out do not constitute a substantial modification.

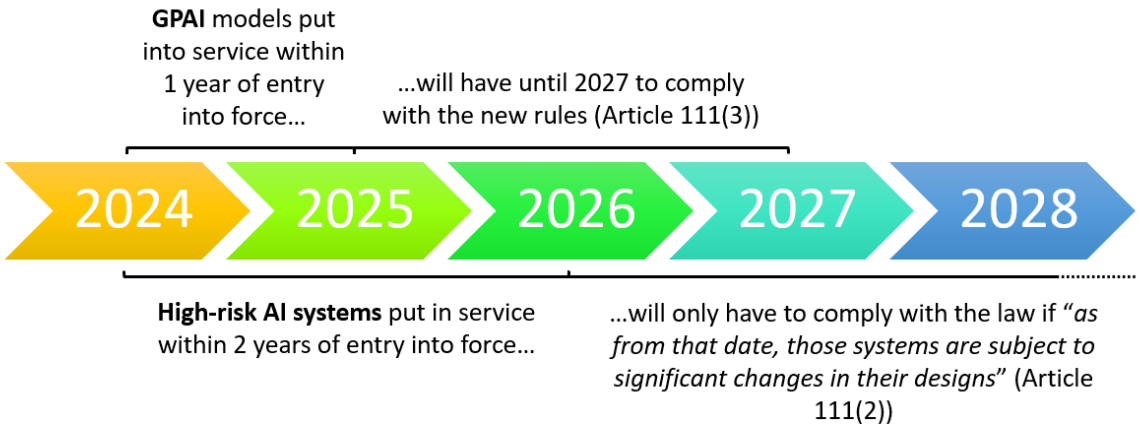

Figure 3 Exceptions to AI Act's timeline of entry into force.

2.1.7 What exemptions exist for free and open-source models?

Providers of free and open-source general-purpose AI models are exempt from certain documentation obligations (see Section 2.2.3 below). Despite these exemptions, GPAI providers must still respect copyright and publish a sufficiently detailed summary of their copyrighted training data.

2.1.8 Is research, testing or development exempted from the rules?

The AI Act also does not apply to research, testing or development activities that occur prior to the AI system being released on the market or put into service. This exception does not cover testing in real-world conditions.

2.2 Obligations for General-Purpose AI Models

The AI Act defines a number of obligations for providers and deployers of *general-purpose AI models*.

2.2.1 What are general purpose AI models?

General-purpose AI models (GPAIs) are models that are not intended for any particular purpose. Instead, they display significant generality and are capable of performing a wide range of distinct tasks – including

assisting journalistic with various tasks linked to their investigations. These models are usually trained with a large amount of data using self-supervision at scale. They can be integrated into a variety of downstream systems or applications. Prominent examples of GPAIs include large language models (LLMs) such as ChatGPT[6].

2.2.2 What obligations apply to GPAI providers?

GPAI providers will need to abide by the following basic rules:

Draw up the technical documentation of the model

Providers should outline key details about the model, including its purpose, acceptable uses, release methods, inputs/outputs, design, training data, and computational resources. They must also describe how the model can be integrated into other systems and provide transparency about design choices, efforts to address biases, and energy consumption. For high-risk models with systemic impacts, providers must include additional information on testing methods (e.g. adversarial testing such as "red teaming"), evaluation criteria, and system architecture. These obligations promote safety, transparency, and accountability, ensuring the model's proper use, reliability, and compliance with regulations. The depth of the documentation depends on how the model is used and its potential risks. for the purpose of providing it, upon request, to the AI Office and the national competent authorities.[7]

Make available sufficient information to downstream providers

Providers who intend to integrate a GPAI into their AI system, must have a good understanding of the capabilities and limitations of the GPAI to comply with their obligations under the AI Act. The information provided to them shall at a minimum describe what the model is designed to do, how it can be used, and any rules for its acceptable use. They must provide details such as when the model was released, how it is distributed, how it interacts with other software or hardware, and whether specific software versions are needed to use it. The explanation should also include the model's overall structure, its inputs (such as text or images), outputs, any size limits (such as the amount of text it can process), and the licensing terms for its use. Additionally, providers must share how the model was developed, including the tools required for integrating it into other systems, details about the formats it uses, and information about the data it was trained on, tested with, and validated, such as where the data came from and how it was prepared or checked.[8]

Put in place a policy to respect copyright law

GPAI providers use vast amounts of data for training purposes. The AI Act requires them to adopt and keep up to date an internal policy to respect copyright, in particular, to identify and respect the reservation of rights expressed by rightsholders (so-called *opt-outs*). The internal policy shall ensure that providers do not circumvent technical measures that restrict access to copyrighted content (e.g. paywalls)

[6] ChatGPT: https://chatgpt.com (last accessed on 26.08.2025)
[7] More details on technical documentation: https://artificialintelligenceact.eu/annex/11/ (last accessed on 26.08.2025)
[8] More details on technical documentation to downstream providers: https://artificialintelligenceact.eu/annex/12/ (last accessed on 26.08.2025)

and do not scrape content from websites that repeatedly and persistently violate copyright (i.e. pirate or counterfeit sites). AI providers are also required to employ web-crawlers that read and comply with the instructions expressed by the robots.txt and other machine-readable protocols. They must commit to inform rightsholders about their employed web-crawlers and to implement adequate measures that prevent their models from generating outputs that reproduce copyrighted content from the training data. Lastly, AI providers who also run online search engine are required to ensure that their compliance with rights reservations does not prejudice the indexing of rightsholders' websites.

> Note that under copyright law, it is prohibited to use content behind a paywall to train a generative AI model even if you paid for accessing the content. This is because you only paid for accessing the content, not for using it for other purposes, such as developing AI models.

Draw up and making publicly available a summary of the content used for training

To enable rightsholder to retain control over their copyrighted material, GPAI providers must publish *sufficiently detailed* summaries of the content they used for training their general-purpose AI model. The EU AI Office has recently published a template[9] that should be used for this purpose.

2.2.3 What exemptions exist for open-source GPAIs?

Providers of free and open-source general-purpose AI models are exempt from the first two of the four documentation obligation:

- They need not draw up the technical documentation of the model.
- They need not make available sufficient information for downstream providers who intend to integrate the GPAI into their AI system.

However, they are still required to put in place a policy to respect copyright law and to making publicly available a sufficiently detailed summary of the training data.

2.2.4 What obligations apply to GPAI deployers?

No obligations apply to deployers of GPAI as such. If you integrate GPAIs into your AI system you must only follow the rules applicable to your own AI system if any.

Some GPAIs that are used to interact with humans (e.g. journalists or fact-checkers) or that generate synthetic content will need to comply with certain transparency obligations (see Section 2.3 below).

[9] Explanatory notice and template for the public summary of training content for general-purpose AI models: https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models (last accessed on 26.08.2025)

2.3 Transparency Obligations

The AI Act introduces transparency obligations for limited risk AI systems to make sure that individuals know that they are interacting with an AI system or exposed to an AI-generated or AI-manipulated output.

2.3.1 Which AI systems trigger transparency obligations under the EU AI Act?

The EU AI Act distinguishes four types of limited risk AI systems that trigger transparency obligations due to their potentially misleading nature:

- AI systems intended to interact directly with natural persons (e.g. journalists or fact-checkers).
- AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, including deepfakes.
- Emotion recognition systems.
- Biometric categorisation systems (including facial detection and/or recognition).

2.3.2 What obligations apply to AI systems that interact with individuals?

Providers of AI systems intended to interact directly with individuals (e.g. chatbots) must design and develop these systems in such a way that the individual be informed that they are interacting with an AI system, unless this is obvious from the point of view of an individual who is reasonably well-informed, observant and circumspect, considering the circumstances and the context of use.

They need to clearly and noticeably inform the individuals involved about this at the very latest during their first interaction with it. The information should also meet applicable accessibility standards.

This obligation does not apply to AI systems authorised by law to detect, prevent, investigate or prosecute criminal offences. However, it is unlikely that an investigative journalism tool would receive such authorisation.

2.3.3 What obligations apply to generative AI outputs?

Providers of AI systems, including general-purpose AI systems, generating synthetic audio, image, video or text content, must mark the outputs of the AI system in a machine-readable format so they are detectable as artificially generated or manipulated.

Providers need to make sure their technical solutions are effective, interoperable, robust and reliable. They should consider the unique features and limitations of different types of content, the cost of implementing the solution, and the commonly accepted best practices in the field, usually found in relevant technical standards.

However, this requirement does not apply if the AI systems are just helping with basic editing tasks or if they do not significantly change the input data or its meaning.

2.3.4 What obligations apply to deepfakes?

Deepfakes are defined as AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful. Notably, under the AI Act, deepfakes must resemble an existing person, object, place, entity or event. This means that images, audio or video content does not constitute a deepfake if it does not resemble anything real even if it seems real.

If deployers use an AI system to create or change image, audio, or video content that results in a deepfake, they need to clearly state that the content has been artificially made or altered.

Again, they need to clearly and noticeably inform the individuals exposed to the deepfake about this at the very latest at the moment of exposure. The information should also meet applicable accessibility standards.

2.3.5 What obligations apply to AI-generated or AI-modified texts?

If a deployer uses an AI system to create or alter text intended to inform the public about important issues – such as during a journalistic investigation – they need to clearly indicate that the text was made or changed by AI. However, this obligation is generally not relevant for journalists because of a specific exemption for editorial content: AI-generated text that has been reviewed or edited by a human, and where someone holds editorial responsibility for its publication, does not need to be indicated as AI-generated or AI-modified.

2.3.6 What obligations apply to emotion recognition systems and biometric categorisation systems?

If someone is using a system that recognises emotions or categorises people based on their biometric data, they need to tell the individuals involved that the system is being used at the very latest during their first interaction or exposure to it. The information should also meet applicable accessibility standards.

2.4 Obligations for High-Risk AI Systems

While not prohibited, certain AI systems and practices still pose a high level of risks that need to be addressed by developers and deployers, by following a number of obligations.

2.4.1 What are high-risk AI systems under the EU AI Act?

The EU AI Act singles out certain AI systems that it considers posing significant risks to the health, safety, or fundamental rights of individuals. These so-called *high-risk AI systems* are subject to stricter requirements and oversight to mitigate potential harms.

Specifically, high-risk AI systems include the following uses that could be relevant for investigative journalists:

- **Remote biometric identification**: Using an AI system for the purpose of identifying individuals, without their active involvement, typically at a distance through the comparison of a person's biometric data with the biometric data contained in a reference database. Biometric data includes physical, physiological, behavioural, or psychological human features.
- **Emotion recognition**: Using an AI system for the purpose of identifying or inferring emotions or intentions of natural persons based on their biometric data. The AI Act refers to emotions or intentions such as happiness, sadness, anger, surprise, disgust, embarrassment, excitement, shame, contempt, satisfaction and amusement. It contrasts them with physical states, such as pain or fatigue, the detection of which would not be considered a high-risk practice.
- **Biometric categorisation** according to sensitive or protected attributes or characteristics: Using an AI system for the purpose of assigning natural persons to a certain category based on one of the following types of information: Racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, health data, sex life or sexual orientation.

2.4.2 What AI systems used in high-risk contexts are exempted?

The AI Act recognises that even if the context in which an AI system is used is highly sensitive, the system's actual use might be so limited that it does not significantly increase the related risk. The following four kinds of uses of AI systems, even if they occur in high-risk contexts, are exempted:

- If the AI system performs a simple, narrow task.
- If the AI system is intended to enhance the outcomes of a task that was previously completed by a human.
- If the AI system is used to identify decision-making patterns or deviations, without replacing or influencing previous human assessments unless properly reviewed by a human.
- If the AI system is used to carry out preliminary tasks for evaluations relevant to high-risk cases.

It remains unclear for now how these exemptions will be interpreted as they could be potentially very broad or very narrow. The AI Act mandates the European Commission to provide guidelines on how to implement these exceptions in practice by 2 February 2026 of the AI Act entering into force, along with detailed examples.

For instance, using biometric identification for the purpose of retrieving footage from public service media (PSM) archives might fall under one of these exemptions. However, if the biometric identification is considered to constitute profiling[10], then the exception will not apply. AI-based fact-checking tools may or may not fall under any of these exemptions, depending on its task and how it completes it.

[10] Article 3(52) of the AI Act refers to the GDPR definition of profiling (Article 4(4) GDPR): "*any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements*". In 2018,

2.4.3 What obligations apply to providers of high-risk AI systems?

AI systems used for these high-risk purposes must be more rigorously evaluated and monitored to protect individuals and society from adverse impacts. Providers and deployers of these systems are required to adhere to specific standards for data quality, transparency, documentation, and human oversight. Coincidently, several of these standards overlap with journalistic standards (e.g. risk management, data protection), and design requirements for fact-checkers and journalists using AI-based fact-checking systems (e.g. clear instructions, human oversight)[11].

Here is what providers need to do:

Establish a Risk Management System

Providers need to create a system that continually monitors and updates risk management throughout the AI's lifecycle. This system should identify and assess potential risks to health, safety, or fundamental rights and implement strategies to address these risks. It should aim to minimise risks while meeting essential requirements and ensure that any remaining risks are acceptable. For AI systems classified as high-risk, they should conduct thorough testing to find the best risk management strategies and verify their effectiveness. The system should also consider how the AI might adversely affect minors or other vulnerable groups.[12]

Employ Best Practices for Managing AI Training and Testing Data

For high-risk AI systems, it is essential to use high-quality data sets for training, validation, and testing. Proper management of these data sets is crucial, including taking into account aspects such as data collection methods, preparation processes, potential biases, and any data gaps. The data should be relevant, representative, as error-free as possible, and comprehensive. It must also be tailored to the specific context in which the AI will be deployed. In certain instances, developers may need to process sensitive personal data to identify and correct biases, but they must follow strict conditions to protect individuals' rights and freedoms.[13]

Develop and Maintain Technical Documentation for at Least 10 Years

Before introducing a high-risk AI system, providers should prepare and keep comprehensive technical documentation up to date. This documentation should demonstrate compliance with legal requirements and provide clear information for authorities to verify adherence. It must include specific key elements. Small businesses and start-ups can submit this information in a more straightforward format, and the EU will provide a simplified form for this purpose. If the high-risk AI system is connected to a product

the EDPB published https://ec.europa.eu/newsroom/article29/items/612053/en/ (last accessed on 26.08.2025) on automated individual decision-making which provide additional details as to the interpretation of the definition.

[11] For more details on these needs, see conference paper: Gaye, L., Schild, A., Lopez, E. (2025). Designing for trustworthiness in AI-based fact-checking services. In Proceedings of HCI International Conference 2025.

[12] More details on risk management: https://artificialintelligenceact.eu/article/9/ (last accessed on 26.08.2025)

[13] More details on data and data governance: https://artificialintelligenceact.eu/article/10/ (last accessed on 26.08.2025)

regulated by other EU laws, a single set of documentation should encompass all necessary details. The EU may update these requirements as technology evolves.[14]

Log AI System Activities for a Minimum of 6 Months

High-risk AI systems need to be equipped with an automatic logging feature to document events throughout their entire lifecycle. This ensures that the system's actions can be traced, particularly if the AI poses a risk or undergoes major changes. Logs should include details such as usage times, the database used for data checks, any data matches, and who validated the results. This is crucial to maintain accountability and ensure the safe operation of high-risk AI systems.[15]

Provide Clear User Instructions for Operating the AI and Interpreting Its Results

High-risk AI systems should be designed with transparency in mind, enabling users to understand and operate them effectively. These systems must include clear instructions that cover details about the provider, the system's capabilities and limitations, as well as any potential risks. The guidance should also explain how to interpret the system's outputs, any preset modifications, and maintenance procedures. Additionally, if applicable, the instructions should describe how to collect, store, and interpret data logs.[16]

Ensure Human Oversight of the AI System

High-risk AI systems need to be designed to allow for effective human supervision. The purpose of this oversight is to prevent or reduce potential risks to health, safety, or fundamental rights that could result from the system's use. The oversight methods should be appropriate to the risks and context of the AI's application. These methods can either be integrated into the system by the provider or applied by the user. The AI system should be presented such that the overseer can comprehend its capabilities and limitations, identify and address issues, avoid excessive dependency on the system, interpret its outputs, choose not to use it, or halt its operation. For some high-risk AI systems, any action or decision based on the system's conclusions must be confirmed by at least two qualified individuals.[17]

Ensure the AI System is Reliable, Accurate, and Secure Against Cybersecurity Threats

High-risk AI systems must be engineered to be reliable, accurate, and secure. They should maintain consistent performance throughout their lifecycle. The European Commission will collaborate with stakeholders to establish methods for assessing these qualities. The accuracy of the AI systems should be clearly stated in their documentation. They should be able to withstand errors and faults and have contingency plans ready. Additionally, these systems should be designed to minimise the risk of biased outputs. Finally, they must be protected against attempts by unauthorised parties to exploit any vulnerabilities.[18]

---

[14] More details on technical documentation: https://artificialintelligenceact.eu/article/11/ (last accessed on 26.08.2025)
[15] More details on record keeping: https://artificialintelligenceact.eu/article/12/ (last accessed on 26.08.2025)
[16] More details on transparency and provision of information to deployers: https://artificialintelligenceact.eu/article/13/ (last accessed on 26.08.2025)
[17] More details on human oversight: https://artificialintelligenceact.eu/article/14/ (last accessed on 26.08.2025)
[18] More details on accuracy, robustness and cybersecurity: https://artificialintelligenceact.eu/article/15/ (last accessed on 26.08.2025)

Conduct the relevant conformity assessment[19] before placing the AI system on the market or putting into service

For high-risk AI listed in Annex III of the AI Act, including biometric identification, categorisation and emotion recognition, providers can either follow an internal review process[20] or a stricter procedure[21]. This stricter procedure involves a notifying body which the provider is free to choose. If no recognised standards exist, or certain standards are not fully followed, the more stringent process of Annex VII[16] must be used. Note that compliance rules may be updated by the European Commission to reflect technical progress or improve safeguards, so providers need to make sure to always align implementations with the most current legal guidance.

Note on standards

High-risk AI systems and general-purpose AI models conforming to recognised harmonised standards are presumed to meet EU requirements. The European Commission is tasked with issuing requests to create such standards to enhance AI system efficiency and compliance. The Commission may develop common specifications if harmonised standards are inadequate, unavailable, or fail to address essential rights. Compliance with these specifications would also ensure conformity with regulatory standards.

Make an EU declaration of conformity

Providers of high-risk AI systems must prepare a written declaration confirming the system meets required standards. This declaration, which can be physical or digital, must be kept for 10 years after the system is launched and be accessible to national authorities in an understandable language. If the system is also covered by other EU laws, a single declaration covering all requirements is needed. Providers are responsible for keeping the declaration up-to-date, and the EU can adjust its content if necessary.

Mark the AI system with a CE label

The CE marking, which shows a product meets EU safety standards, must be clearly visible on high-risk AI systems. If it cannot be physically placed on the system, it should be on the packaging or documentation. For digital AI systems, a digital CE marking should be easily accessible. If a notified body (an organisation that checks the product meets the standards) is involved, their identification number should be included next to the CE marking. If the AI system is also subject to other EU laws requiring a CE marking, the marking indicates it meets those requirements too.

Register the AI system into the EU database for high-risk AI systems

Before launching or using a high-risk AI system, providers or their representatives must register both themselves and the system in the EU database. This also applies to AI systems deemed not high-risk by the provider. Public authorities using high-risk AI systems must also register the system and its use. For

[19] More details on conformity assessment: https://artificialintelligenceact.eu/article/43/ (last accessed on 26.08.2025)
[20] See Annex VI: https://artificialintelligenceact.eu/annex/6/ (last accessed on 26.08.2025)
[21] See Annex VII: https://artificialintelligenceact.eu/annex/7/ (last accessed on 26.08.2025)

certain high-risk AI systems used in law enforcement, migration, or border control, registration must be in a secure, private section of the database, accessible only to the EU Commission and national authorities.

Establish a quality management system

Providers should establish a quality management system that should be thoroughly documented and incorporate strategies for regulatory compliance, design and development protocols, testing and validation processes, technical specifications, data management, risk management, post-market surveillance, incident reporting, communication procedures, record-keeping, resource management, and an accountability framework. The implementation of this system should be scaled to the size of the provider's organisation. Providers already under quality management requirements due to relevant European Union laws can integrate these elements into their current systems. Financial institutions can fulfil these obligations by adhering to EU financial services regulations.[22]

Set up a post-market monitoring system

Providers should create a system to gather and analyse data on the performance of AI systems throughout their lifespan, ensuring ongoing compliance with regulations. This monitoring system should follow a plan included in the technical documentation, with the European Commission providing a template for the plan. If a similar monitoring system already exists under other laws, providers can incorporate the necessary components from this new requirement, as long as it offers the same level of protection.[23]

Report Unacceptable Risks and Take Corrective Actions

If a company providing a high-risk AI system discovers that their system is not complying with EU regulations, they must promptly address the issue, cease its use, or recall it. They must notify everyone involved in distributing or using the system. If the system presents a risk, the company needs to investigate the cause and inform the authorities monitoring the AI market. Additionally, they should notify any organisation that certified the system, explaining the problem and the corrective measures taken.[24]

Report Serious Incidents

Providers of high-risk AI systems need to inform local authorities about any major incidents linked to their AI system as soon as they become aware of a possible connection. Reports should be made within 15 days of discovering the incident. If the incident is very severe or widespread, the report must be submitted within 2 days. In cases of fatalities, the report is required within 10 days. Providers can initially submit an incomplete report, if necessary, but they must follow up with a complete report. They are also responsible for investigating the incident and cooperating with authorities, who will respond within 7 days.[25]

[22] More details on quality management systems: https://artificialintelligenceact.eu/article/17/ (last accessed on 26.08.2025)
[23] More details on post-marker monitoring: https://artificialintelligenceact.eu/article/72/ (last accessed on 26.08.2025)
[24] More details on corrective actions and duty of information: https://artificialintelligenceact.eu/article/20/ (last accessed on 26.08.2025)
[25] More details on reporting serious incidents: https://artificialintelligenceact.eu/article/20/ (last accessed on 26.08.2025)

2.4.4 What obligations apply to deployers of high-risk AI systems?

Deployers of high-risk AI systems, as opposed to providers, have their own set of responsibilities, which they too overlap to some degree with journalistic responsibility:

- **Assign human oversight:** Ensure that the people overseeing the AI system have the right skills, training, authority, and support. Deployers are free to organise their own resources and activities for the purpose of implementing these human oversight measures.
- **Ensure data quality:** Make sure the data fed into the AI system is relevant and representative for its intended use. Deployers must retain control over which data is fed into the system.
- **Monitor system's operations:** Where deployers have reason to consider that the use of the high-risk AI system in accordance with the instructions may result in that AI system presenting a risk to the health or safety, or to fundamental rights, of persons, they shall, without undue delay, inform the provider or distributor and the relevant market surveillance authority, and shall suspend the use of that system. Where deployers have identified a serious incident, they shall also immediately inform first the provider, and then the importer or distributor and the relevant market surveillance authorities of that incident.
- **Maintain logs** that have been automatically generated by the AI system for at least 6 months.

⚠ If a deployer is a public body or a private group providing public services, they must assess the impact on fundamental rights. This includes describing how and when the system will be used, who it might affect, and what risks it might pose. They also need to outline how humans will oversee the system and what steps will be taken if risks materialise. This assessment must be done for the first use of the system, but can be updated. The results must be reported to the market surveillance authority. The AI Office will provide a template to help with this process.

All PSM – whether public or private entities – will have to conduct a fundamental rights impact assessment[26] if they deploy a high-risk AI system. However, this impact assessment will not need to be published; PSM and other entities providing public services are only required to inform the national market surveillance authority of its results. It is not clear whether an organisation working closely with or for a PSM would be covered by this provision. Still, as partners of PSM such organisations will likely be required by the PSM to ensure they respect human rights. In the context of high-risk AI systems, the main concern is with privacy rights, and in the context of the EU this means adherence to the (GDPR)[27].

[26] More details on the fundamental rights impact assessment: https://artificialintelligenceact.eu/article/27/ (last accessed on 26.08.2025)

[27] For more details on how to adhere to the GDPR, see Section 3

2.5 Prohibited AI Practices

Due to the risk that they may pose to society and individuals, a certain number of AI practices are prohibited by the AI Act.

2.5.1 What AI uses are forbidden under the EU AI Act?

AI systems that are considered too risky to people's safety and rights will be banned according to the AI Act. These bans mostly concern AI systems used by public authorities such as law enforcement. But there are some types of AI systems that could also be developed or used by private actors such as journalists and that are prohibited:

- AI systems that gather facial images from the internet or CCTV without targeting specific individuals.
- AI systems that use manipulative or deceptive techniques, or exploit people's vulnerabilities, such as age or disability, to change their behaviour in a harmful way.
- AI systems that use biometric data to categorise individuals based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation.

These bans are put in place to protect individuals from potentially harmful or invasive uses of AI technology[28], especially by the state.

3 Data Protection Obligations for AI Developers and Deployers of AI-Based Fact-Checking Tools

AI developers and deployers must ensure that they fully comply with data protection and privacy laws that apply in relevant jurisdictions. If data processing involves personal data processing, i.e. processing of any information relating to an identified or identifiable natural person (*data subject*), in the EU they need to ensure compliance with the General Data Protection Regulation (GDPR)[29].

The text below sets out a number of GDPR principles and provisions that are likely to be relevant when developing and deploying AI-based fact-checking systems.

Defining a purpose[30]

An AI system using personal data must always have a clearly defined, legitimate, and understandable purpose established at the design stage. The goal must be specified upfront. A broad aim such as "develop AI" is insufficient: the type of system or use case should be specified.

[28] Full list of prohibited AI systems: https://artificialintelligenceact.eu/article/5/ (last accessed on 26.08.2025)
[29] GDPR: https://www.consilium.europa.eu/en/policies/data-protection-regulation/ (last accessed on 26.08. 2025)
[30] Article 5(1)(b) GDPR

Establishing a legal basis[31]

All personal data processing must be based on one of the six GDPR legal grounds: consent, contract performance, legal obligation, vital interests' protection, public interest task, legitimate interest. The legal basis needs to be determined before the start of processing personal data and the choice and reasoning should be documented. Amongst these:

- Consent must be freely given, specific, informed and unambiguous. NB: collecting consent for large-scale or scraped data is often impractical.
- Legitimate Interest requires[32] a three-step test:
    - The interest must be lawful and clearly defined.
    - Data use must be strictly necessary for that interest.
    - The impact on individuals' privacy must not be disproportionate.

Special categories of data[33]

Processing sensitive data demands particular caution. Sensitive data includes for example biometric data, or data revealing racial or ethnic origin. The processing of these categories of data is in general prohibited, except when specific exemptions apply, such as in the case of the data subject's explicit consent, or where the processing relates to personal data that are manifestly made public by the data subject. In the latter case, a legal basis (e.g. legitimate interest) is still required.

Data minimisation[34]

This principle requires that AI systems collect and use only data that are adequate, relevant and non-excessive, i.e. limited to what is necessary for the determined purpose.

Safeguards: pseudonymisation / anonymisation

To comply with the GDPR, personal data should be removed or pseudonymised wherever possible. However, it is important to note that the threshold for claiming that an AI model is anonymous is very high. According to the European Data Protection Board (EDPB), an AI model can only be considered anonymous if the likelihood of re-identifying individuals from the data used to train the model or from its outputs is insignificant.[35] This means that while pseudonymisation and anonymisation are valuable tools for reducing privacy risks, they must be applied carefully. Simply removing direct identifiers does not

[31] Article 6 GDPR
[32] Data protection aspects related to the processing of personal data in the context of AI models: https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en (last accessed on 26.08.2025)
[33] Article 9 GDPR
[34] Article 5(1)c GDPR
[35] EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, p.16: https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en (last accessed on 26.08.2025)

automatically exempt an AI system from GDPR obligations, unless the risk of re-identification is minimal to the point of being effectively impossible.

Defining a retention period

Under the GDPR, personal data must not be kept indefinitely.[36] To comply with this, a specific retention period must be set, after which the data should either be deleted or, where appropriate, securely archived. It is the responsibility of the data controller to define this retention period in accordance with the original purpose for which the data was collected.

Data protection impact assessment[37]

The GDPR requires to perform a data protection impact assessment (DPIA) in certain instances, e.g. when processing on a large scale of special categories of data. This can be relevant for AI systems that process biometric data.

In order to write a DPIA, developers and deployers can contact their local Data Protection Authority for guidance. As an example, the Irish Data Protection Authority offers a sample DPIA[38] on its website, for guidance purposes.

Transparency: providing information and explicability

The transparency principle of the GDPR requires any information or communication relating to the processing of personal data to be concise, transparent, understandable and easily accessible, using clear and plain language – i.e. privacy notices need to explain in clear language what personal data the system uses, for what purposes, and how outputs will be used.

Data subject rights management

When personal data is processed, it is essential to uphold the rights granted to individuals under the GDPR. These include the rights of access (Article 15), rectification (Article 16), erasure (Article 17), restriction of processing (Article 18), data portability (Article 20), and objection (Article 21) to certain personal data processing, e.g. processing based on legitimate interest grounds (including profiling) and processing for direct marketing purposes. These protections are fundamental in ensuring individuals are not subject to the outcomes of automated systems without having the ability to understand how their data is used and, if necessary, challenge or oppose such processing.

In practice, these rights apply at every stage of the AI system's life cycle – from the personal data within training datasets to the information processed and generated during the system's operational phase. As such, data controllers should account for these obligations from the design phase, incorporating suitable procedures and tools to address any data subject requests effectively.

[36] Article 5(1)e GDPR
[37] Article 35 GDPR
[38] Sample DPIA template from the Irish Data Protection Authority: https://www.dataprotection.ie/en/organisations/know-your-obligations/data-protection-impact-assessments#sample-dpia-template (last accessed on 26.08.2025)

Supervising automated decisions[39]

Under the GDPR, individuals have the right not to be subject to fully automated decisions, particularly those based on profiling, that produces legal effects or similarly significantly impact them. However, such decisions can be automated if the individual consents, if it is necessary for a contract, or if it is legally authorised. In these cases, individuals must be informed, able to access the logic behind the decision, challenge it, and request human review.

4 Ethical Considerations for Developers and Deployers of AI-Based Fact-Checking Tools

While legal obligations are clear-cut requirements set by laws and regulations, ethical obligations encompass broader considerations of principles or values that guide individuals and organisations.

Ethics require developers and deployers to think holistically about the societal, cultural, and human impacts of their tools, beyond what legal frameworks can predict or enforce. Laws such as the GDPR or the AI Act offer critical protections, however they often cannot account for all contexts, complexities, or the rapidly evolving nature of AI in journalism.

Ethical considerations, on the other hand, ensure developers create tools that not only comply with standards, but actively promote public trust, safeguard democratic values, and empower journalists in their mission to report truthfully and responsibly.

Due to their crucial role in providing accurate and truthful information to citizens, media organisations have been particularly active in developing ethical frameworks around the procurement and use of AI tools in professional settings. Some of them developed internal guidance even before the emergence of popular AI tools such as ChatGPT.[40]

More recently, ethical guidelines have flourished in the media industry, both around editorial and non-editorial issues related to AI. These cover questions such as how to integrate tools into internal workflows while maintaining a high level of oversight and responsibility, how to minimise bias in the output of AI systems, or how to inform audiences about certain uses of AI and what level of information is considered useful. Examples of guidelines from public service broadcasters include the BBC's editorial guidance on the use of AI[41], DW's approach to generative AI[42], and ZDF's guiding principles for using generative AI[43].

It is important for the developers and deployers of AI-based tools for fact-checking and journalism to become aware of these ethical considerations and to include them into their development process. Only

[39] Article 22 of the GDPR
[40] See for example the BBC's 2021 Machine Learning Engine Principles: https://www.bbc.co.uk/rd/projects/bbc-machine-learning-engine-principles (last accessed on 26.08.2025)
[41] BBC guidance for the use of AI from, last updated on 24.06.2025: https://www.bbc.co.uk/editorialguidelines/guidance/use-of-artificial-intelligence (last accessed on 26.08.2025)
[42] DW's ethical guidelines on AI: https://www.dw.com/en/what-is-deutsche-welles-approach-to-generative-ai/a-66868035 (last accessed on 26.08.2025)
[43] ZDF's guiding principles for using generative AI: https://www.zdf.de/unternehmen/guiding-principles-generative-ai-100.html (last accessed on 26.08.2025)

then will these tools meet the requirements set by the news community, and be adopted by them as end-users.

As part of its work with its Member organisations (public service media broadcasters from Europe and beyond) on the ethical aspects of AI in public service media, the European Broadcasting Union has gathered and published a regularly updated repository of Members ethical guidelines which is available on its AI Ethics Group's webpage[44]. In 2024, this group conducted a workshop with representatives of several member organisations, to determine the commonalities in terms of ethical requirements for AI systems used in their organisations, and compile them under thematic values and principles that are of importance for the news and media community.

Below, we list these values and principles, and provide practical recommendations in how to approach them in order to best meet the ethical requirements of these organisations.

Human oversight

Human-in-the-loop is an important ethical principle across PSM organisations. Ensuring that humans are involved is critical both to monitor the performance of the AI system and to enable journalists to take full responsibility for their work.

Practical recommendations:

- Design tools that support journalists in making informed editorial decisions rather than over-automating processes. Ensure that the AI functions as a collaborative assistant helps uphold journalistic integrity and autonomy.
- Design interfaces that clearly set out at which stage of the use of an AI tool human intervention can take place.
- Provide accessibility features using fonts, colours, redundant information visualisation, contrast, and usability.

Transparency and explainability

Journalists and audiences must understand how AI tools reach specific conclusions, such as why a fact was flagged as false or why a deepfake was identified. This demands explainable AI that provides actionable insights, enabling journalists to remain accountable to their audiences.

[44] EBU AI Ethics Group's guideline repository (some with EBU Members-only access): https://www.ebu.ch/groups/ai-ethics#relatedPresentations-ec5be566-a02c-4cd1-b3d2-5c0692175cbc (last accessed on 26.08.2025)

Practical recommendations:

- Provide a clear technical documentation of the AI system, including its design, training data, the methodologies it uses.
- Provide a clear explanation of how an output was produced, e.g. by using visual aids such as confidence scores, decision trees, or data provenance graphs to explain algorithmic decisions.
- Communicate in a transparent manner about tool limitations (e.g. error rates, potential bias), so that journalists and audiences can understand the scope and boundaries of the AI's capabilities, use it responsibly and trust its outputs.

Privacy and security

Developers must ensure that they comply with data protection and privacy laws, but they also need to accommodate the unique privacy needs of journalists, such as safeguarding sources or sensitive research data. This may include specialised measures well beyond general data security requirements.

Practical recommendations:

- Add a robust end-to-end encryption for all data processed by the AI, whether in transit or at rest.
- Provide options for local offline processing, allowing journalists to use AI tools without uploading sensitive data to the cloud.
- Make use of multi-factor authentication (MFA) to limit who can access AI platforms and sensitive data.
- Make use of audited and secure APIs, databases, and cloud infrastructure to ensure they are resilient against cyberattacks.

Accuracy and bias mitigation

As even minor errors in fact-checking or misinformation detection can propagate false information on a large scale and erode public trust, developers should have an ethical responsibility to test tools rigorously in real-world journalistic scenarios and to continually update them to ensure reliability.

Developer should also consider whether any inherent biases can affect the deployment of AI systems. For example, an AI fact-checking tool may exhibit cultural or geographic bias by disproportionately flagging

claims from certain regions of the world or from smaller independent outlets as "unverified", or trusting inaccurate automatic translations of rare languages from bigger companies over local expert ones.[45]

Practical recommendations:

- Audit AI outputs and their impacts post-deployment to ensure that the system's real-world performance aligns with ethical and regulatory requirements.
- Build transparent feedback systems allowing users or affected parties to highlight unintended harms the AI system may be causing, creating an iterative cycle of improvement.
- Support interdisciplinary and multicultural collaboration, with input from domain experts (ethicists, journalists, civil society) – not just data scientists or engineers – who can bring awareness of broader societal and cultural implications.
- Use diverse training datasets that include content from a variety of cultures, geographical regions, and media outlets. Paying attention to underrepresented communities or minority languages is of particular importance, especially for multilingual or translation tools.

Conclusion

This handbook has provided an accessible guide to the legal and ethical requirements for developers and deployers of AI systems – especially under the EU AI Act and under related regulations, such as the General Data Protection Regulation (GDPR) – that are relevant to the development of AI-based tools for fact-checking and investigative journalism. We hope that this document will prove useful to this community, as the fight against disinformation continues.

[45] The EBU and DW organised a workshop in 2024 on western bias in datasets and their impact on journalism (video talk accessible to EBU Members only): https://www.ebu.ch/events/aidi-microworkshops?date=20240422#previous-edition (last accessed on 26.08.2025)

vera.ai

vera.ai is a Horizon Europe Research and Innovation Project co-financed by the European Union under Grant Agreement ID: 101070093, an Innovate UK grant 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00245.