



AUSTRIAN RED CROSS

**ACCORD**

Austrian Centre for Country of Origin  
& Asylum Research and Documentation

# Ethical considerations for the use of Artificial Intelligence in COI

January 2026



Bundesministerium  
Inneres



**UNHCR**  
The UN Refugee Agency

ACCORD is co-funded by the Asylum, Migration and Integration Fund, UNHCR and the Ministry of the Interior, Austria.

<http://accord.redcross.at>

ACCORD - Austrian Centre for Country of  
Origin & Asylum Research and Documentation

## Ethical considerations for the use of Artificial Intelligence in COI

January 2026

© Austrian Red Cross/ACCORD

An electronic version of this document is available on [www.ecoi.net](http://www.ecoi.net).

Austrian Red Cross/ACCORD

Wiedner Hauptstraße 32

A- 1040 Vienna, Austria

Phone: +43 1 58 900 – 582

E-Mail: [accord@redcross.at](mailto:accord@redcross.at)

Web: <http://accord.redcross.at/>



## TABLE OF CONTENTS

1	Introduction.....	3
1.1	The role of COI in international protection procedures .....	3
1.2	The potential of AI as “quick-fix”, relief mechanism or support option .....	4
1.3	Why ethical guidelines are needed: Understanding the risks and limits of generative AI.....	5
1.4	Aim of the paper.....	7
2	Overview of existing ethical guidelines and practical considerations.....	8
2.1	Ethic guidelines: discourse and consensus .....	8
2.2	Selected examples of established AI ethics guidelines .....	11
2.2.1	UNESCO: Recommendation on the Ethics of Artificial Intelligence.....	11
2.2.2	OHCHR: AI and Human Rights 101 .....	12
2.2.3	EU: Guidelines, legal obligations and its implications for the use of AI in the field of COI .....	12
2.3	Ethical considerations and practice in journalism, medicine and academia .....	14
3	The specific case of COI .....	17
3.1	Specific risks and responsibilities .....	17
3.2	(Potential) Applications of AI in COI.....	18
3.2.1	AI support for textual and linguistic tasks .....	18
3.2.2	AI support with research tasks .....	19
3.2.3	AI support to include more sources in COI databases.....	20
3.2.4	AI support for search optimisation in COI databases .....	21
3.2.5	AI support for end users’ access to COI.....	21
4	Discussion on the ethically sound use of AI in COI practice in accordance with COI Standards and Principles .....	23
4.1	COI quality standards.....	23
4.2	Principles for researching and using COI .....	26
5	Conclusion .....	28
6	Sources .....	30

# 1 Introduction

## 1.1 The role of COI in international protection procedures

Country of Origin Information (COI) plays a central role in decisions on applications for international protection. COI refers to information that provides insight into the situation in the countries of origin of asylum seekers and is therefore relevant for determining their protection status. In 2004, UNHCR defined accurate and reliable information as *conditio sine qua non* for determining the need for international protection and for developing strategies that may include voluntary return, but also termination or withdrawal of protection status (UNHCR, February 2004, pp. 3-4). According to the European Union Asylum Agency (EUAA, formerly EASO), COI should support fair decision-making but not dictate decisions in these procedures (EUAA, February 2023, p. 6). Therefore, COI is fundamentally different from so-called *country guidance*<sup>1</sup> and legal assessments of conditions in countries of origin but forms an essential basis for both.

In the context of international protection procedures, COI is used to help assess whether applicants would be at risk of persecution on the grounds set out in the 1951 Refugee Convention or of other inhumane treatment upon their return based on which subsidiary protection is granted. COI also supports the evaluation of the credibility of the asylum seeker's statements, in terms of their consistency with available knowledge about their country of origin.

It should be noted, however, that not every element of an applicant's claim presented in the procedure must be substantiated, provided that all available evidence has been examined, and the applicant's statements have been found to be plausible and consistent ("benefit of the doubt"). As ACCORD notes in its Training Manual on Researching Country of Origin Information, COI can substantiate certain statements made by asylum seekers, but even with comprehensive COI research, it is hardly possible to independently confirm all information. COI rarely provides definitive answers about an individual's credibility or need for international protection (ACCORD, January 2024, p. 29).

In international protection procedures, COI serves as supplementary evidence. It is not the sole determinant of a protection status and does not replace the legal assessment of the facts describing the circumstances in the country of origin.

Given its evidential but non-determinative nature, COI holds a distinct and critical position between factual research and legal decision-making. It requires a stringent adherence to various COI quality standards and research principles. However, upholding these is increasingly being tested in light of limited resources and automation trends. The question arises as to whether new AI-based tools can be used responsibly in COI processes – to increase efficiency

---

<sup>1</sup> Country Guidance produced by EUAA or UNHCR's International Protection Considerations are policy advice documents that provide country-specific guidance for consistency in asylum decision-making.

without compromising the quality and ethical foundation of sound COI research practice and outcomes.

## 1.2 The potential of AI as “quick-fix”, relief mechanism or support option

In many countries of asylum, responsibility for COI research lies with the public administration. As in many other areas of public administration, asylum and migration systems, e.g. within the European Union, face resource constraints. Several European asylum systems continue to be under pressure of high application numbers (see e.g. Secretary-General of the European Commission, 12 November 2025, p. 7). Intergovernmental organisations (IGOs) as well as non-governmental organisations (NGOs) active in the field of COI are similarly affected by limited resources, often due to their dependence on public funding. Across national contexts, COI research is frequently affected by limited time, human resources, and language capabilities, all of which influence the comprehensiveness and reliability of COI products. These structural challenges have a direct impact on the quality of COI and, consequently, on the fairness and efficiency of international protection procedures.

In response to resource limitations in fields such as public administration and following the widespread public attention surrounding generative AI tools such as ChatGPT, AI is often discussed as a potential ‘quick-fix’ solution. It is seen as a potential means of achieving cost savings, round-the-clock availability, speed, and flexibility, and as a way of providing solutions where human labour is scarce (see, for example, EY, 27 March 2024). A 2023 study identified “predictive analytics for decision-making” and “document reviews” among the most promising application areas of artificial intelligence in public administration (Madan & Ashok, 2023, pp. 1–2) – tasks sharing similarities to those relevant in COI research, which include information retrieval, document synthesis and analysis.

In a February 2025 paper on the potential of AI in the field of COI published in UNHCR’s Legal and Protection Policy Research Series, Evangelos Kanoulas noted that artificial intelligence can be considered a promising avenue for advancing COI research. As tools such as large language models (LLMs) can be used to support various COI research tasks, including cross-checking facts, summarising information, extracting key details, fill in report templates, detecting tone, bias and inappropriate language, translating languages and transcribing audio (Kanoulas, February 2025, pp. 11-13), these applications can improve efficacy, accuracy as well as the overall quality of COI reports, according to Kanoulas (Kanoulas, February 2025, p. 19).

**AI terminology.** “Artificial intelligence” currently serves as an umbrella term encompassing various systems and infrastructures, rather than representing a single, clearly definable technology (see, e.g., Möller, 6 June 2024). Following the hype surrounding artificial intelligence that began with the launch of ChatGPT in late 2022, when talking about “AI” the general public tends to refer to so-called foundation models (FMs), a form of generative AI, i.e., AI capable of generating content. As Kanoulas explained in his February 2025 paper, FMs are “advanced AI systems trained on massive amounts of data, allowing them to handle many different tasks using a single model”. These models typically rely on deep learning techniques, which use multi-layered neural networks to recognise complex patterns and generate outputs. As FMs are capable of adapting to various situations, they are deemed “incredibly useful in many areas” (Kanoulas, February 2025, p. 7). LLMs, such as the ones used by OpenAI’s ChatGPT,

are a specific type of FM designed to generate human-like text. These text-based generative models are trained on training data and use a statistical model to generate their output, predicting the next word, phrase, or sentence based solely on probability and context (see ACUTE, forthcoming, p. 6)

In the context of public service provision, however, the public sector’s commitment to public value entails a heightened ethical responsibility when adopting AI technologies (Madan & Ashok, 2023, pp. 1-2; van Noordt & Tangi, 2023, p. 2). Unlike the private sector, which does not face the same level of scrutiny in prioritising public values (van Noordt & Tangi, 2023, p. 2), for public institutions “ethical tensions [related to the use of AI] such as questions of fairness, transparency, privacy, and human rights” remain a major concern (Madan & Ashok, 2023, pp. 1–2). As these tensions highlight the need to carefully balance efficiency gains with the protection of fundamental rights, introducing AI tools into sensitive administrative areas raises not only technological questions, but also normative ones.

Consequently, for the field of COI we are faced with essential questions about the extent to which AI technologies can support COI research processes within relevant normative frameworks. Although AI applications have the potential to mitigate resource constraints and enhance the efficiency of COI production, the central issue is not merely whether AI can be applied, but whether it can be integrated in a responsible manner so that established COI principles and quality standards that ensure fair international protection procedures are upheld. As Radanliev and colleagues (2024, p. 3) argue in their paper on ethics and AI deployment, effectively addressing ethical concerns is essential to ensure that “the benefits of AI are maximised while minimising its potential negative impacts. By adopting an ethical framework that promotes transparency, accountability, and fairness, we can ensure that AI is employed ethically and in the best interests of society.”

### 1.3 Why ethical guidelines are needed: Understanding the risks and limits of generative AI

As decisions made in international protection procedures can have a profound impact on people’s lives, misuse or overreliance on automated tools may directly endanger individuals’ rights and safety. A closer examination of the inherent risks and limitations of generative AI is therefore essential for the development of ethical guidelines.

A key limitation of generative AI models is the phenomenon of **hallucinations**, which arises directly from these models’ **probabilistic output generation**. As large language models cannot distinguish between true and false information but generate text based on statistically likely word sequences, they may produce entirely fabricated content presented in a coherent and persuasive manner. Recent reporting indicates that newer models are even more prone to hallucinations. Some researchers therefore argue that hallucinations are not a temporary flaw that will eventually be eliminated, but an inherent feature of generative AI models’ underlying functionality (NYT, 6 May 2025).

According to Kanoulas (February 2025, p. 15), in the context of COI where “accuracy is paramount”, the main risks and limits of generative AI arise not only from their “probabilistic output generation, and vulnerability to manipulation or misalignment”, but also from these models’ **high dependency on training data** (Kanoulas, February 2025, pp.17-18). This reliance

is widely recognised across ethical frameworks as a core concern, since the functioning and outputs of generative AI models are shaped by data that often remain undisclosed to users and rarely undergo rigorous ethical scrutiny.

Moreover, these vast amounts of undisclosed training data, which are often based on human-generated content, are rarely controlled for potential bias. As a result, the models – and subsequently the decisions they inform – are inherently **susceptible to reproducing existing biases** (Möller, 6 June 2024). Accordingly, AI systems can adopt biases that impact the fairness, transparency, and accountability of their outputs (Lu, 2024, p. 2, Radanliev et al., 2024, p. 3). This means that AI applications may discriminate against certain groups. Even if parameters such as race and gender are not directly used, these systems can still derive such characteristics from other information, such as a combination of a person’s neighbourhood and occupation. Moreover, it is argued that it is often quite challenging to demonstrate that AI applications discriminate against specific groups because these systems can justify “any of its decisions in page-long statistical formulas, but these explanations are rarely transparent, especially not for its users” (Möller, 6 June 2024).

This leads to another layer of risk arising from the opacity and interpretability challenges inherent to deep learning models. AI systems employing such technologies are highly complex and **often function as ‘black boxes,’** making it difficult – even impossible in some cases – to explain how inferences are generated or how input data is processed into conclusions (Jarrahi et al., 2023, pp. 93-94; Radanliev et al., 2024, p. 1). The fact that spurious correlations can remain hidden within “opaque outputs” underlines the importance of validating AI-generated insights.

In light of the above-mentioned limitations and since AI-algorithms cannot contextualise or interpret their own outputs, **human oversight remains essential**, especially in evidence-based domains such as public administration, where compliance requirements demand clear explanations of how AI systems generate recommendations and conclusions based on their input data (Jarrahi et al., 2023, pp. 93-94).

Beyond these technical aspects, the increasing **datafication of social life** introduces broader ethical implications. According to current research, the increasing focus on data and algorithms on the part of practitioners and researchers also carries the risk that individual human needs are being “simplified into data points” and that “emotions, privacy, and societal roles risk becoming entrenched in the digital realm.” (Lu, 2024, p. 2)

It is also vital to acknowledge that there are **power asymmetries** inherent in current AI development, since the (moral) design of AI technologies are largely determined by commercial developers (Rudolphina Research Magazine, 6 June 2024) and the “demand for market knowledge is propelled by commercial competition across various industrial sectors” (Lu, 2024, p.1). This concentration of influence further underscores the need for public-sector frameworks and ethical governance mechanisms to ensure that AI adoption aligns with democratic values and human rights rather than purely market-driven objectives.

According to Radanliev et al. (2024, p. 1), aligning artificial intelligence with ethical norms poses a complex and multifaceted challenge. This involves balancing the societal benefits of AI with protecting individual rights and ensuring transparency. To address this, ethical considerations must be embedded in system design from the outset, guided by “principles such as justice,

transparency, and accountability to ensure that AI systems make fair decisions, are understandable to users, and have proper checks and balances” (Radanliev et al., 2024, p. 10).

### 1.4 Aim of the paper

This paper aims to provide an overview of key ethical considerations for the responsible application of AI in COI practice by taking into account the field’s specific legal, societal and contextual characteristics. While AI has demonstrated transformative potential in various sectors, COI is a particularly sensitive field, characterised by its qualitative, descriptive, and context-dependent nature. Unlike other fields, COI research does not seek to provide a definitive answer – it operates within a framework that values contextualisation and the coexistence of multiple perspectives. It is therefore essential to ensure that technological efficiency does not come at the expense of nuance, human scrutiny, or ethical integrity.

This tension between technological promise and the trustworthiness of “knowledge production” and its potential costs has been described by Lu (2024, p.2), depicting the ethical imperative that guides the present paper:

“This age poses intricate questions regarding knowledge oversight and selection. Artificial intelligence has become a catalyst for accelerated knowledge innovation. As we harness the power of AI, we simultaneously grapple with the unintended consequences of this newfound knowledge production tool, often overlooking the imperative need for an accountable system. Amidst the time of technological enthusiasm, there’s a growing inclination to place unwavering faith in models and data, forming a belief that machines outperform human intellect. Yet, amid this technological enthusiasm, we should not lose sight of our duty to uphold the dignity and diverse needs of individual human beings.”

The risks and challenges associated with the use of generative AI make it clear that using AI in the sensitive field of international protection procedures raises technical, legal and ethical questions. Against this backdrop, **ethical guidelines and legal frameworks are becoming increasingly important, as they provide guidance for the design and use of AI systems and are essential for addressing and counteracting the inherent ethical pitfalls of generative AI applications.** These pitfalls include, as outlined above, their susceptibility to manipulation or misalignment, vulnerability to bias and discrimination, opaque and non-transparent results, the oversimplification of people’s lives into mere data without consideration of the human consequences, and the risk of reinforcing power asymmetries.

## 2 Overview of existing ethical guidelines and practical considerations

The ethical challenges surrounding the development and application of artificial intelligence have been the subject of extensive discussion, even before generative AI applications drew significant public attention (see, e.g., Fjeld et al., 2020). To assess whether and how AI can be used responsibly in COI research, it is helpful to examine existing ethical frameworks and regulatory mechanisms. This chapter gives an overview of discussions about ethical principles and compares various AI ethics guidelines, highlighting both their commonalities and differences. Furthermore, with the EU AI Act now in place, the chapter examines how ethical considerations have informed legal requirements and discusses considerations from related areas of professional practice.

### 2.1 Ethic guidelines: discourse and consensus

Over the past decade, the rapid evolution of artificial intelligence has prompted a surge in the number of ethical and governance frameworks proposed by a wide range of organisations. As stated by Fjeld and colleagues (2020, p. 4), “seemingly every organization with a connection to technology policy has authored or endorsed a set of principles for AI”. Similarly, a systematic literature review published in 2025 highlighted that the discourse on responsible approaches to the development and implementation of AI has gained significant attention, largely due to an increasing number of incidents in which the use of AI has had “unforeseen or undesirable repercussions” (Papagiannidis et al., 2025, p. 4).

However, despite governments, corporations and researchers increasingly emphasising their importance, key questions about the definition and implementation of AI-related ethical standards and principles still need to be addressed. In this regard, Papagiannidis and colleagues speak of “conceptual opacity”, as academic discourse, for example, still uses overlapping or synonymous terms such as “trustworthy AI” and “principled AI”, reflecting the field’s conceptual immaturity. Consequently, determining what constitutes responsible AI remains “a work in progress” (Papagiannidis et al., 2025, pp. 4-6).

Despite the continuing lack of clarity surrounding its fundamental definition, research indicates that a growing consensus is emerging regarding the key themes and pillars of responsible AI guidelines (see, Fjeld et al., 2020; Papagiannidis et al., 2025). Already in 2020, Fjeld and colleagues (2020, pp. 4-5) analysed a quite diverse set of 36 AI principles documents, including guidelines authored by actors ranging from governments to civil society and private companies, each tailoring their guidelines to strategic, advocacy, or organisational needs. Their study revealed eight recurring themes across these documents, with more recent publications tending to incorporate all eight of these themes:

**“Privacy.** Principles under this theme stand for the idea that AI systems should respect individuals’ privacy, both in the use of data for the development of technological systems and by providing impacted people with agency over their data and decisions made with it. [...]

**Accountability.** This theme includes principles concerning the importance of mechanisms to ensure that accountability for the impacts of AI systems is appropriately distributed, and that adequate remedies are provided. [...]

**Safety and Security.** These principles express requirements that AI systems be safe, performing as intended, and also secure, resistant to being compromised by unauthorized parties. [...]

**Transparency and Explainability.** Principles under this theme articulate requirements that AI systems be designed and implemented to allow for oversight, including through translation of their operations into intelligible outputs and the provision of information about where, when, and how they are being used. [...]

**Fairness and Non-discrimination.** With concerns about AI bias already impacting individuals globally, Fairness and Non-discrimination principles call for AI systems to be designed and used to maximize fairness and promote inclusivity. [...]

**Human Control of Technology.** The principles under this theme require that important decisions remain subject to human review. [...]

**Professional Responsibility.** These principles recognize the vital role that individuals involved in the development and deployment of AI systems play in the systems' impacts, and call on their professionalism and integrity in ensuring that the appropriate stakeholders are consulted and long-term effects are planned for. [...]

**Promotion of Human Values.** Finally, Human Values principles state that the ends to which AI is devoted, and the means by which it is implemented, should correspond with our core values and generally promote humanity's well-being." (Fjeld et al., 2020, pp. 4-5)

More recently, Papagiannidis and colleagues (2025) conducted a scoping review of 48 already published research papers on responsible AI frameworks, identifying the following seven core principles and 18 sub-dimensions:

#### **"Accountability**

- *Auditability*: Ability to assess AI applications concerning the algorithms, data, and design processes.
- *Responsibility*: Oversight of the various stages and activities involved in AI deployment and how it should be allocated to people, roles, or departments.

#### **Diversity, non-discrimination and fairness**

- *Accessibility*: Design of systems in a manner that makes them accessible and usable for everyone, regardless of age, gender, abilities, and characteristics.
- *No unfair bias*: Inclination of prejudice toward or against people, objects, or positions, as well as inherent biases in datasets, which can precipitate undesirable outcomes.

#### **Human agency and oversight**

- *Human review*: Right of a person to challenge a decision made by an AI.
- *Human well-being*: The notion that AI must include human well-being as a primary success factor for development.

### Privacy and data governance

- *Data quality*: Accuracy of values in a dataset, matching the true characteristics of the entities described by the dataset.
- *Data privacy*: AI systems' development and operation in a manner that considers data privacy throughout the data lifecycle.
- *Data Access*: National and international rights laws during the design of an AI for data access permissions.

### Technical robustness and safety

- *Accuracy*: AI system's ability to make correct judgements, such as correctly classifying information into the appropriate categories or being able to predict, recommend, or make intelligent decisions based on data or data models.
- *Reliability*: AI system's ability to work properly when subjected to a range of inputs or situational contexts.
- *General Safety*: Safety rules and fallback plans that should be established for AI systems in the event of problems.
- *Resilience*: AI systems that should be protected against vulnerabilities that adversaries can exploit, e.g., hacking.

### Transparency

- *Explainability*: Ability to explain the technical processes of an AI system and related human decisions (e.g., application areas of a system).
- *Communication*: Human right to be informed in advance when interacting with an AI agent.
- *Traceability*: Ability to track data and processes that yield the AI system's decision, including data gathering, labeling, and algorithms.

### Social and environmental well-being

- *Social well-being*: ubiquitous exposure to social AI systems in all areas of society, such as work and education.
- *Environmental well-being*: most pressing environmental and climate concerns facing the planet" (Papagiannidis et al., 2025, Table 4)

While the two above-mentioned scientific papers suggest an emerging consensus on the development of ethical AI standards, and their frameworks – one from 2020 and the other from 2025 – demonstrate substantial overlap in identified principles, it is essential to note that a comprehensive framework summarising commonly cited principles is not inherently superior to any individual principles document. This is because the **relevance of ethical principles depends largely on the context of the respective field**, as certain themes may hold greater significance in specific settings (see, Fjeld et al., 2020, p. 5). It is crucial to take into account the particular legal context, ethical factors, and social consequences involved in implementing AI across various domains (Radanliev et al., 2024, p. 14). Therefore, **principles and AI ethical guidelines should always be tailored to the context in which they aim to be applied.**

## 2.2 Selected examples of established AI ethics guidelines

The following section presents three illustrative examples of AI ethics guidelines. Examining these cases, which differ considerably in scope and form, will help us to better understand how principles are adapted to diverse contexts. These insights may provide a valuable foundation for developing appropriate guidelines for our specific case: the ethical use of AI in the field of COI.

### 2.2.1 UNESCO: Recommendation on the Ethics of Artificial Intelligence

UNESCO's *Recommendation on the Ethics of Artificial Intelligence* was published in November 2021. It was created with the aim of serving as the first-ever “globally accepted normative instrument” on AI ethics (UNESCO, November 2021, p. 14) and is now applicable to all 194 UNESCO member states (UNESCO, no date). The recommendation functions not only as a set of value- and principle-based guidelines but also aims to provide actionable measures through concrete policy recommendations (UNESCO, November 2021, p. 14). Central to its approach are the four core values of ensuring that AI development and deployment uphold **human rights and dignity**, foster **environmental sustainability**, embrace **diversity and inclusion** across cultural, linguistic, gender, and social dimensions, and contribute to **peaceful, just, and interconnected societies** (UNESCO, November 2021, pp. 18-20). While all its four core values and ten principles (see below) are beneficial, the recommendation acknowledges that it may be necessary in practice to carefully weigh up and assess the context in order to resolve potential conflicts (UNESCO, November 2021, p. 18).

#### “Core principles

**Proportionality and Do No Harm:** The use of AI systems must not go beyond what is necessary to achieve a legitimate aim. Risk assessment should be used to prevent harms which may result from such uses.

**Safety and Security:** Unwanted harms (safety risks) as well as vulnerabilities to attack (security risks) should be avoided and addressed by AI actors.

**Right to Privacy and Data Protection:** Privacy must be protected and promoted throughout the AI lifecycle. Adequate data protection frameworks should also be established.

**Multi-stakeholder and Adaptive Governance & Collaboration:** International law & national sovereignty must be respected in the use of data. Additionally, participation of diverse stakeholders is necessary for inclusive approaches to AI governance.

**Responsibility and Accountability:** AI systems should be auditable and traceable. There should be oversight, impact assessment, audit and due diligence mechanisms in place to avoid conflicts with human rights norms and threats to environmental wellbeing.

**Transparency and Explainability:** The ethical deployment of AI systems depends on their transparency & explainability (T&E). The level of T&E should be appropriate to the context, as there may be tensions between T&E and other principles such as privacy, safety and security.

**Human Oversight and Determination:** Member States should ensure that AI systems do not displace ultimate human responsibility and accountability.

**Sustainability:** AI technologies should be assessed against their impacts on ‘sustainability’, understood as a set of constantly evolving goals including those set out in the UN’s Sustainable Development Goals.

**Awareness & Literacy:** Public understanding of AI and data should be promoted through open & accessible education, civic engagement, digital skills & AI ethics training, media & information literacy.

**Fairness and Non-Discrimination** [sic]: AI actors should promote social justice, fairness, and non-discrimination while taking an inclusive approach to ensure AI’s benefits are accessible to all.” (UNESCO, no date)

### 2.2.2 OHCHR: *AI and Human Rights 101*

In contrast to the 43-page UNESCO recommendation, which is primarily addressed to states and public policy makers, the *AI and Human Rights 101* of the Office of the High Commissioner for Human Rights (OHCHR) offers a compact, practical four-page guide for companies, civil society, and states. Based on the three pillars of the UN Guiding Principles on Business and Human Rights (UNGPs)<sup>2</sup> – the state’s duty to protect, corporate responsibility, and access to remedy – the approach emphasises the Human Rights Due Diligence (HRDD) process, centred on the four steps of “human rights risk identification and prioritization, integration and action, tracking effectiveness, and communication”, as a key mechanism. The document acknowledges that implementing AI can pose risks, including increasing bias, discrimination and a lack of transparency in decision-making processes (OHCHR, 23 October 2025, p.1).

### 2.2.3 EU: *Guidelines, legal obligations and its implications for the use of AI in the field of COI*

In 2019, the High-Level Expert Group on Artificial Intelligence (AI HLEG) of the European Commission published the AI Ethics Guidelines under the title *Ethics Guidelines for Trustworthy AI*. These guidelines set out seven core principles or “key requirements” for the ethical use of artificial intelligence, designed to ensure that AI systems are not only robust but also lawful and ethical (European Commission, 8 April 2019). The EU’s ethical guidelines are intended as a dynamic working document, to be regularly updated in line with technological, social, and knowledge developments. As outlined in the guidelines, they are not designed to replace or prevent current or future policy decisions or regulations, but rather to serve as a starting point

---

<sup>2</sup> More information on the UN Guiding Principles on Business and Human Rights can be found here: OHCHR – Office of the High Commissioner for Human Rights: Guiding Principles on Business and Human rights. Implementing the United Nations “Protect, Respect and Remedy” Framework, 2011  
[https://www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR\\_EN.pdf](https://www.ohchr.org/sites/default/files/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf)

for discussions on trustworthy AI in Europe (AI HLEG, 8 April 2019, p. 4). The guidelines provide orientation and recommendations and are legally non-binding in nature.

Five years later, in 2024, the EU AI Act entered into force, becoming the world's first transnational set of rules regulating the use of artificial intelligence. Unlike the ethical guidelines, the AI Act has legally binding force. While the ethical guidelines are not formally part of the binding regulation, they have significantly influenced its development (Recital 27 AI Act). Many of the ethical recommendations have been incorporated into the act (see, for example, Art. 10 AI Act on Data and Data Governance, Art. 13 AI Act on Transparency and Provision of Information to Deployers, Art. 14 AI Act on Human Oversight). Recital 27 of the AI Act explicitly refers to the guidelines and acknowledges them, along with their seven principles, as a foundation for the legislative framework:

“According to the guidelines of the AI HLEG, **human agency and oversight** means that AI systems are developed and used as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans.

**Technical robustness and safety** means that AI systems are developed and used in a way that allows robustness in the case of problems and resilience against attempts to alter the use or performance of the AI system so as to allow unlawful use by third parties, and minimise unintended harm.

**Privacy and data governance** means that AI systems are developed and used in accordance with privacy and data protection rules, while processing data that meets high standards in terms of quality and integrity.

**Transparency** means that AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights.

**Diversity, non-discrimination and fairness** means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law.

**Social and environmental well-being** means that AI systems are developed and used in a sustainable and environmentally friendly manner as well as in a way to benefit all human beings, while monitoring and assessing the longterm impacts on the individual, society and democracy.” [emphasis and paragraph breaks added] (Recital 27 AI Act)

The EU AI Act seeks to provide legal certainty for stakeholders such as providers, deployers, and affected individuals. One of its core components is the classification of AI systems into distinct risk categories: prohibited, high-risk, limited risk, and minimal risk. Currently, the wording of the EU AI Act leaves room for interpretation, particularly regarding the vaguely

defined category of “high-risk AI”<sup>3</sup> (see e.g., FRA, 2025, pp. 1; 23-24), which, however, could have significant implications for AI use in the context of COI.

Taken together, these three examples illustrate how ethical principles for AI are presented in markedly different formats depending on their scope and regulatory intent. While UNESCO offers a comprehensive, value-driven framework for states, OHCHR translates human rights obligations into a concise, practice-oriented tool for organisations, and the EU links ethical considerations to binding legal obligations. Despite these variations, they share a common normative core centred on human rights, accountability, fairness/non-discrimination, transparency, and the need for contextual assessment. The comparative examination thus highlights both the versatility and the recurring normative foundations of AI ethics. These insights form an essential basis for deriving context-specific guidance for the ethical use of AI in the field of COI, where similar principles must be interpreted with regard to the sector’s particular tasks and responsibilities.

### 2.3 Ethical considerations and practice in journalism, medicine and academia

Although the field of COI is highly specialised, it shares important similarities with other fields, particularly with regard to ethical considerations surrounding AI applications. Having outlined three different AI ethics guidelines from various intergovernmental/supranational actors, it is also worth examining other sectors to identify the challenges they face when implementing AI and assess whether best practices have already been established. These cross-sector insights – focusing on the field of journalism, medicine and academia – can highlight potential pitfalls and best practices, thereby ensuring that ethical standards in the field of COI benefit from a broader pool of knowledge and experience.

#### About responsibility

In some academic disciplines – where textual analysis and knowledge production are as central as in the COI field – the use of AI is already being discussed, with studies showing its significant potential to shape how knowledge is produced and disseminated (Lu, 2024, p. 1). AI has been described as a “game-changer” (Krüger, 25 September 2024) that is expected to accelerate scientific discovery (University of Copenhagen, 3 October 2023) and enhance various aspects of knowledge management, encompassing the “creation, storage, retrieval, sharing, and application of knowledge” (Jarrahi et al., 2023, p. 87). However, given the high standards set in

---

<sup>3</sup> High-risk AI, defined in Article 6, generally includes systems listed in Annex III (Article 6 (2)), unless they pose no significant risk to health, safety, or fundamental rights. AI is not considered high-risk if it performs narrowly defined procedural tasks, supports rather than replaces human judgement, or only prepares assessments (Article 6 (3)). Annex III specifies AI systems used by authorities in asylum, visa, and residence permit procedures with the aim of determining applicants’ eligibility, including assessing the reliability of evidence are considered high-risk AI. This means AI applications in the field of COI may as well fall under this category. Whether the exception in Article 6 is applicable in this context is uncertain, as fundamental rights – such as the right to respect for private and family life (Article 8 ECHR) – may be affected during asylum proceedings. Clarification on what constitutes “material influence” and how this aligns with Annex III provisions is pending; the Commission will issue guidelines including practical examples of what is considered high-risk and not high-risk. The requirements for high-risk AI will reportedly take effect in December 2027 at the latest.

academic settings, it is crucial that AI systems are used in a responsible and transparent manner to avoid undermining the legitimacy of scientific knowledge within wider society (University of Copenhagen, 3 October 2023).

As with academic research, the responsible use of AI is of paramount importance in medicine and healthcare. While the emphasis in an academic context is on preserving the legitimacy of scientific findings in terms of their societal role, medicine focuses on ensuring that the well-being of the individual patient always takes priority when using AI to reduce diagnostic errors and improve treatment outcomes, for example. As medicine can significantly impact a patient's life, it is a particularly sensitive area. This creates a great responsibility to ensure the safe and ethically sound implementation of AI technologies (see, e.g., Samios, 12 January 2024).

The rise of AI in journalism has sparked debate about whether current ethical standards are adequate. In November 2023, for example, Reporters Without Borders (RSF) and 16 partner organisations introduced the Paris Charter on AI and Journalism<sup>4</sup>, setting out ten dedicated principles to guide the use of AI in journalism, emphasising the importance of human oversight in all editorial decisions and requiring media organisations to ensure transparency, traceability and clear differentiation between authentic and AI-generated content (RSF, 10 November 2023; euroactive, 29 October 2024). Other press and media councils, however, argue that existing ethical frameworks already provide adequate guidance. From this perspective, AI does not fundamentally change journalists' core obligations, as succinctly put by the Chairman of the Dutch Journalistic Council (Raad voor de Journalistiek):

“Journalistic responsibility is paramount and then it does not matter whether a publication is the work of the editor-in-chief, an intern or a chatbot. And whoever bears responsibility must, in our opinion, also be prepared to account to the public. The use of AI is therefore subject to the same principles as any other journalistic conduct.” (Chairman of the Dutch Raad voor de Journalistiek, quoted in Press Councils.eu, no date)

### AI application and its ethical implications

Ethical considerations surrounding AI extend beyond questions of responsibility, encompassing broader concerns about the impacts of AI deployment within various disciplines. In medicine, for example, scientific literature increasingly addresses the “black box” problem posed by generative AI systems. As already discussed in chapter 1, the internal processes of these systems generally remain undisclosed, meaning that only their inputs and outputs are observable (Xu & Shuttleworth, February 2024). In this context, it is stressed that transparency and explainability are essential for integrating AI into clinical practice, as clinicians must be able to understand why AI systems make particular recommendations. Such clarity is crucial for thoroughly reviewing AI-generated suggestions and identifying deviations that may be

---

<sup>4</sup> More detailed information on the Paris Charter on AI and Journalism can be found here: RSF – Reporter Sans Frontières: Paris Charter on AI and Journalism, 10 November 2023

<https://rsf.org/sites/default/files/medias/file/2023/11/Paris%20charter%20on%20AI%20in%20Journalism.pdf>

significant for individual cases (Lauritsen et al., 31 July 2020). Without adequate transparency, there is a risk that medical expertise may become less accessible and clear to patients, increasing the risk of opacity in medical decision-making. This could violate patients' rights to full information and diminish their ability to participate meaningfully in decisions that affect their lives profoundly. This can result in a return to a paternalistic physician-patient relationship, where medical decisions are made without fully informing patients. The lack of explainability can also strongly increase patient mistrust and suspicion (Xu & Shuttleworth, February 2024).

Concerns about the broader implications of generative AI are also emerging within scientific research and knowledge production. Recent research has discussed the potential negative impact of using generative AI in scientific knowledge production. Since generative AI is based on statistical probabilities, there are concerns that its use could “inadvertently narrow perspective”. This could occur even if the number of publications increases, and there are concerns that these technologies could create scientific “monocultures”. In this scenario, certain forms of knowledge production would dominate, “potentially leading to bias in content and a loss of innovation” (Krüger, 25 September 2024).

The parallels with journalism, medicine, and academic research highlight that fields requiring contextual judgement, dealing with human vulnerability, or involved in knowledge production all encounter similar risks when adopting AI. For the field of COI, these insights underscore the essential need to **ensure transparency** in AI usage and the **traceability of its outcomes**. This is not only to preserve the legitimacy of COI products and the decisions based on them, but – perhaps most importantly – to guarantee that applicants whose lives are profoundly affected by these decisions are not subject to opaque or even AI-hallucinated decisions, thereby safeguarding their rights. Additionally, one must consider whether the use of AI might **unintentionally reduce the variety** of sources of information accessible over time. And, as in journalism, it is worth considering **whether the field of COI requires specific guidelines for AI use, or whether established COI standards and principles are sufficient** and potentially just require clarification or adaptation. These lessons from related professions thus provide important guidance for shaping ethical AI use in COI practice.

### 3 The specific case of COI

Having broadened our perspective on AI and ethical guidelines by looking beyond the immediate field of COI, we have discovered similarities but also conclude that the question of ethically appropriate AI usage can best be answered by considering the context and internal structures of a specific domain. This chapter examines how the specific characteristics of COI shape both the opportunities and the limitations of integrating AI tools into COI practice.

#### 3.1 Specific risks and responsibilities

For the field of COI, three key aspects stand out in terms of the specific roles of COI and the responsibilities of COI researchers, and how these must be considered when implementing AI applications.

Firstly, as COI plays a supplementary yet potentially decisive role in international protection procedures, its application can have profound implications for applicants' lives. In this sense, **safeguarding the rights of applicants for international protection and upholding a do-no-harm approach** when considering the application of AI is essential at all stages. In this regard, in a September 2024 publication on the application of AI in migration, the European Network of National Human Rights Institutions (ENNHRI) expresses concerns about a lack of transparency surrounding the use of various technologies in determining international protection claims, directly impairing migrants' right to an effective remedy (ENNHRI, September 2024).

Secondly, **diversity of sources and consideration of the relevant context are central** to COI research, as is **presenting information in a balanced way**. Rather than providing a definitive answer, COI research aims to offer a well-contextualised account of complex situations by drawing on multiple perspectives, capturing developments over time and making uncertainties and ambiguity transparent. This approach helps to provide the most comprehensive and balanced picture possible. However, this approach conflicts with the way generative AI processes information, including the prevalent risk of oversimplification, and hidden biases in generative AI output. As emphasized in an article on AI's challenges arising from contextual differences in meaning across cultures and societies, even seemingly objective terms and statistical indicators are defined differently across legal, temporal, and geographic settings, leading to competing definitions. For probabilistic models that select a version without understanding the underlying context, such variability is a vulnerability. This problem is exacerbated by the uneven geographic origin of the training data fed into these models (Janowicz, 3 July 2024). For COI, which strives for an accurate depiction of local conditions, these opaque structural distortions pose a particular risk. Moreover, when dealing with contested or politically sensitive issues, such as competing claims over national borders, AI models cannot rely on statistical likelihoods to decide which version is valid. Such questions require contextual knowledge rather than statistical refinement. Therefore, understanding how AI systems represent the world and the contextual assumptions embedded in their outputs is essential when assessing their suitability for COI tasks.

Thirdly, **all information cited in a COI product must be clearly referenced and traceable, and its factual accuracy must be assessed** using well-established COI standards. When sources are of uncertain credibility, their reliability must be systematically and transparently evaluated at the meta-level. However, these requirements conflict with two inherent limitations of generative

AI: probabilistic output generation, which can result in hallucinations, and the black box-like nature of generative AI, which often makes it impossible to determine why certain conclusions are reached in the AI's output. These concerns are exacerbated by the fact that the amount of AI-generated content on the internet is currently growing significantly (OeAW, 14 February 2025). Consequently, the underlying sources of specific information are becoming increasingly difficult to identify, which in turn makes it much more difficult to assess the accuracy of information when conducting online research.

## 3.2 (Potential) Applications of AI in COI

The reasons for using generative AI applications in COI research include improving COI products in terms of their presentation and thoroughness, and making them easier for users to access and apply. AI applications are therefore expected to optimise the information gathering process, including in relation to information that was previously difficult to access; improve how information is presented and formulated; and make it easier for users to retrieve COI. Ideally, these advancements could lead to more accurate, detailed and timely information, greater reliance on reliable sources, and clearer presentation (see ACUTE, forthcoming, p. 11). The following section categorises potential AI applications by function and user group, highlighting both their benefits and the risks posed by the inherent limitations of generative AI.

### 3.2.1 AI support for textual and linguistic tasks

AI tools can provide substantial assistance in handling and refining information collected during COI research. LLMs, such as the ones used by OpenAI's ChatGPT for example, are designed to generate text that mimics human communication. Widely used LLMs have already demonstrated impressive performance in various linguistic tasks, including formulating, paraphrasing, summarising and translating text. In the context of COI, these capabilities can be applied in various ways. For example, they can be used to improve the clarity and coherence of COI products, paraphrase source excerpts, condense or synthesise lengthy passages or translate information from local languages into languages that are accessible to a wider range of COI practitioners. Other applications include transcribing interviews or audio material (see ACUTE, forthcoming, p. 12). These text-focused applications are particularly promising, given that COI practitioners often have to process large volumes of material within tight time constraints and sometimes lack the necessary language skills. After all, these types of tasks represent the key function of such generative AI models.

However, even when used in its core area, the benefits of AI are constrained by its inherent limitations. Even in supposedly 'safe' tasks such as summarising, translating or paraphrasing, LLMs may produce errors involving invented, distorted or omitted information due to their probabilistic output generation. This is particularly problematic in the area of COI, where such inaccuracies can have serious consequences for applicants. Furthermore, the internal workings of generative AI models are opaque, providing no insight into why certain textual decisions are made or how conflicting information is weighed. This lack of transparency goes against the methodological requirements of COI. Moreover, generative AI lacks an understanding of context, therefore, it is unable to accurately interpret political sensitivities, ambiguities, or nuances in the source text. This can result in meaning being lost or distorted, particularly in sensitive or complex situations.

The severity of these risks can vary considerably depending on the specific use case. For example, there is a substantial difference between asking an AI model to paraphrase a short passage of text that has been fully read and understood, and relying on it to summarise a multi-page document that has not yet been reviewed. These risks increase further when translating material from languages that the COI researcher does not speak, as there is no straightforward way to verify whether important nuances have been preserved, or whether omissions or distortions have occurred or context-dependent meanings have been lost. It should be noted, however, that the practice of using machine-generated translations (e.g. DeepL or Google Translate) accompanied by a disclaimer is already established in COI products.

Taken together, these factors show that while AI applications for textual and linguistic tasks can enhance efficiency and assist COI practitioners in managing already researched material, their outputs must always be subjected to careful human verification.

### *3.2.2 AI support with research tasks*

In addition to assisting with information that has already been gathered, generative AI tools are also being considered for their potential benefits at the research stage of COI work. These technologies have the potential to efficiently retrieve and organise vast amounts of data, making them quite useful in an environment where time is limited and tasks span over various languages and often quite specific subject areas.

In practice, a range of tools already provides such research-oriented assistance. Services like Perplexity.ai can be used for online research queries that combine web search with LLM-generated explanations. Some generative models, such as OpenAI's "deep research" mode, aim to produce a structured exploration of complex topics, while Google's integration of Gemini summaries at the top of search results illustrates the trend towards automatically generated overviews embedded directly in search interfaces. Such tools may help in quickly identifying relevant sources or provide an initial overview of the information landscape. This development is particularly noteworthy given that search engine ranking algorithms have long been unintelligible to their users. This makes it increasingly difficult to defend the idea that a traditional Google results page should, by default, be considered more reliable than an LLM-generated list of suggested sources.

At the same time, the influence of generative AI on shaping research paths should not be underestimated. While relying on AI to produce an initial overview may be considered as merely the first step in a broader research process, the persuasive formulation style of LLMs can create priming effects that subtly influence the COI researchers' subsequent information-gathering. These risks become even more pronounced when generative AI is used not only at the outset but repeatedly throughout the research process.

Moreover, in the case of these tasks as well, the constraints of generative AI models have to be considered when assessing generative AI's usefulness. For one, there is a clear lack of transparency, which prevents information from being traceable and hinders the assessment of their accuracy, since generative AI systems do not disclose how they select information, which sources specific claims originate from, or why certain elements are prioritised over others. Furthermore, LLM-based research tools remain vulnerable to hallucinations, such as invented references or unsubstantiated claims presented convincingly. Also, the internal optimisation

processes of LLMs tend to favour statistically common patterns over contextually accurate representations, making them ill-suited for research in politically sensitive, contested or highly nuanced subject areas. And finally, there is the issue of the growing volume of AI-generated material online. When AI tools retrieve or summarise text that is itself machine-generated, distortions can be reinforced through cycles of unverifiable reproduction, making it difficult to ensure that information stems from authentic, independent and verifiable sources and significantly complicating corroboration. This heightens the risk of problems well known to COI researchers, such as “round-tripping” and “false-corroboration” (see ACCORD, January 2024, p. 149).

Taken together, these factors demonstrate that, although generative AI tools can be useful for providing initial orientation and support when navigating large amounts of information, there is still a significant risk in using their results for anything other than exploratory cues.

### *3.2.3 AI support to include more sources in COI databases*

Beyond its application in COI production and research, AI is also being explored within COI databases. Given AI’s capacity to process and structure large volumes of data, such applications may appear particularly promising in efforts to broaden the range of COI documents made accessible to users. Examples of such applications include AI-assisted metadata extraction, automated document descriptions, relevance filtering, and automated data entry into databases that store COI-relevant information.

In addition, generative AI makes it technically feasible to integrate new types of sources that were previously difficult or impossible to include on a large scale. This includes the automated retrieval of local-language sources across different media formats (text, audio, or mixed media), as well as the provision of machine-generated translations of such material. The inclusion of selected social media content may also be considered in this context.

From the perspective of improving the thoroughness of COI products, the potential added value of these applications is evident. Expanding source diversity by including, for example, AI-generated transcriptions and/or translations of local language sources may help surface information that would otherwise remain inaccessible. However, the limitations of generative AI discussed in the previous sections apply equally in this context. AI-assisted expansion of source material inevitably requires an additional layer of human oversight. Where such oversight is not feasible on a large scale, transparency measures, such as the practice of disclaimers, which are already common in the field of COI, become necessary.

However, this raises another fundamental issue: the possible shift of responsibility from providers of COI databases to its users. Until now, users of established COI databases could assume that the sources provided had undergone a certain quality assessment and that the documents covered had already been checked for relevance. If AI-generated translations or only sparsely checked local sources were provided, including a disclaimer, this assumption would no longer apply in the same way. The responsibility for assessing reliability, contextual relevance, as well as possible biases would increasingly be transferred to the user, which would have a significant impact on the perception and use of COI databases.

### *3.2.4 AI support for search optimisation in COI databases*

AI-based tools also offer new possibilities for optimising search functions within COI databases, most notably through so-called semantic search based on vector embeddings. Unlike keyword-based search, semantic search aims to retrieve documents based on their overall content rather than the frequency of specific search terms. In principle, such systems can improve access to large and complex COI databases by enabling more flexible retrieval of relevant information.

At the same time, the integration of semantic search raises several methodological concerns. These systems rely on numeric embeddings derived from large training data and are therefore shaped by the same probabilistic logic that underpins generative AI models in general. Relevance rankings are influenced by patterns present in the training data. As a result, biases embedded in training data, such as the overrepresentation of dominant narratives, may influence which documents are prioritised or deprioritised in search results. Local or minority sources may therefore be systematically ranked lower, even when they are highly relevant from a COI perspective. Moreover, relevant sources that use local terminology or non-dominant linguistic patterns may be ranked lower because their phrasing diverges from the dominant patterns.

These risks are further amplified by the fact that COI itself constitutes a relatively specialised domain. If a model used for vectorisation has been trained primarily on general news content, it may lack the contextual sensitivity required to distinguish between superficially similar but substantively unrelated material. In such cases, search results may be influenced by implicit assumptions embedded in the training data. For example, a search for terms related to “honour killings” could disproportionately surface documents associated with certain regions based on stereotypical correlations, even when the content of those documents is unrelated to the topic (but related to the region). In extreme cases, this could introduce discriminatory bias into the search functionality itself.

In COI practice, these dynamics risk reinforcing mainstream narratives while marginalising minority perspectives or locally specific accounts. While semantic search can enhance usability and efficiency, its deployment in COI databases requires careful model selection, continuous monitoring, and transparency towards users regarding how search results are generated and ranked.

### *3.2.5 AI support for end users’ access to COI*

AI-supported applications, such as chat-based information tools or other types of interactive retrieval systems, are being explored as a means of supporting COI users, including decision-makers and legal representatives, in their work. Such systems promise more immediate access to information and may assist users in navigating complex documents or datasets or identifying relevant COI passages quickly.

However, these tools exacerbate the concerns about AI implementation outlined in the previous section, as the potential negative consequences of undermining applicants’ rights, as well as the fairness and soundness of international protection procedures, seem more imminent in these contexts. Chat-based interfaces, in particular, generate probabilistic summaries that condense complex realities into succinct outputs. This condensation process

obscures uncertainty and conceals conflicting perspectives. Because such systems present answers in confident language, they may be perceived as authoritative even when incomplete or factually distorted. Furthermore, users are often unable to reconstruct how an answer was generated, the source of the output or whether the response was influenced by discriminatory bias embedded in the model's training data. Consequently, users exclusively relying on such COI quick-access chatbots cannot properly assess the reliability of the information, and applicants are hindered in meaningfully challenging such AI-generated content, which directly or implicitly affects the subsequent course and/or outcome of the international protection procedure.

At the same time, as outlined in section 3.2.1, the severity of these risks can vary considerably depending on how such AI-supported chat systems are deployed. One comparatively limited-risk application would be "chatting" with individual documents. In such cases, the underlying content of the generated summaries remains more easily traceable, and the assessment of reliability is therefore less problematic. By contrast, systems designed to provide quick, highly condensed outputs, such as one-sentence summaries across multiple documents (for example during asylum interviews or during the court proceeding), potentially drawing on sources unfamiliar to the user, raise far more serious concerns. In such scenarios, AI-generated outputs move increasingly away from the core methodological principles of COI, which rely on contextualisation and if necessary multiple perspectives or even the transparent presentation of uncertainty. Since COI findings can rarely be reduced to a simple answer, the use of highly condensed, chatbot-style summaries risks oversimplifying complex realities.

## 4 Discussion on the ethically sound use of AI in COI practice in accordance with COI Standards and Principles

As COI research and its use in international protection procedures can have a significant impact on the lives of applicants, it is essential that all actors involved in the production and use of COI rely on a shared framework for assessing the quality of COI. COI quality standards and principles have been developed to support fair and efficient procedures and to safeguard objectivity within these procedures as far as possible. According to ACCORD’s Training Manual on Researching Country of Origin Information<sup>5</sup>, these COI quality standards are *relevance, reliability and balance, accuracy and currency, and transparency* and are based on the four fundamental research principles of *neutrality and impartiality, equality of arms as regards access to information, using public information, and data protection* (ACCORD, January 2024, p. 36).

This paper is concerned with determining what constitutes an ethically sound approach to the use of AI in COI research. In order to do this, established COI standards and principles should serve as a primary reference. The following discussion therefore analyses the extent to which the use of generative AI models is compatible with these standards and principles and highlights potential areas of conflict. This in turn serves as the basis for the final conclusion that outlines which adjustments may be necessary to maintain COI quality standards and principles in an AI-supported context.

### 4.1 COI quality standards

#### Relevance

“COI is relevant when it is based on questions rooted in legal concepts of refugee and human rights law or on questions derived from an applicant’s statements.” (ACCORD, January 2024, p. 37)

To meet the standard of relevance, COI must be directly pertinent to the specific circumstances of the applicant and to the issues raised within their case. Thereby, relevance is determined by whether the information assists in evaluating aspects that are important for the assessment of the applicant’s eligibility for international protection under refugee or human rights law, based on the facts and claims in the individual case (ACCORD, January 2024, p. 37).

As emphasised in ACCORD’s training manual, the relevance of information depends significantly on the formulation of meaningful questions that COI can answer, and on the relevance of these answers in the eligibility assessment. In the context of AI-assisted COI work, relevance can directly be tied to the formulation of so-called prompts: the specific task instructions that a user gives to a generative AI model. It should be noted that general-purpose generative AI models by default do not have an “understanding” of the legal concepts that underpin

---

<sup>5</sup> ACCORD first published its COI standards and principles in 2004 in its Training Manual. The 2024 edition of the Training Manual on Researching Country of Origin Information was revised with the participation of a broad range of experts from the international COI community.

international protection procedures, such as refugee status or subsidiary protection. The model will not readily access this contextual knowledge if it is merely asked to provide information on the situation of unmarried women in Uzbekistan, for example. However, it is important to bear in mind not only the lack of contextual knowledge specific to the field, but also that generative AI models lack genuine understanding altogether. They do not “understand” the concept of relevance, let alone relevance within a highly specialised, legally embedded field such as COI. Consequently, AI-generated outputs may include information that is irrelevant for international protection procedures or omit information that is highly relevant for the assessment of a specific case.

This risk is present regardless of the specific use case. Irrelevant content or the omission of relevant information may constitute a problem when AI is used as support in research tasks, search optimisation, but also when it is employed for seemingly limited textual or linguistic tasks. For instance, AI-generated summaries or paraphrases may omit relevant aspects of COI, while AI-assisted search or ranking tools may fail to surface relevant but non-mainstream sources just because these represent niche perspectives statistically less prominent in the training data.

While users can try to mitigate these risks by carefully contextualising prompts or by providing background information specific to the field within the model’s context window, such measures cannot ensure that irrelevant or even hallucinated content will be excluded, nor that all relevant aspects will be consistently included; as niche perspectives, in particular, remain structurally disadvantaged in probabilistic text generation processes.

### **Reliability and balance**

“Decisions on international protection should be based on COI from reliable sources, taking into account the source’s political and ideological context as well as its mandate, reporting methodology and motivation.

As each source has its own perspective and focus, different sources and different types of sources should be consulted to achieve the most comprehensive and balanced picture possible.” (ACCORD, January 2024, p. 38)

The quality of COI research products overall is determined by the reliability of the underlying sources. Therefore, it is crucial for COI researchers and users to understand how to evaluate sources against the established criteria. Bias plays a specific role, since bias does not automatically render a source unusable in the context of COI. Information from sources displaying a certain bias can still be useful, provided the bias is acknowledged and efforts are made to achieve balance by consulting additional sources (ACCORD, January 2024, pp. 38-39).

A key challenge in AI-supported COI processes is that generative AI models cannot independently assess the reliability of sources or contextualise their perspectives. While human COI experts can critically evaluate sources by reflecting on their mandate, intention and credibility, AI systems lack the capacity to make such nuanced judgements, even though their outputs may give the impression of highly analytical depth. Therefore, users must critically review any information suggested or summarised by AI tools, considering the quality and appropriateness of the underlying sources.

Furthermore, it is important to note that source evaluation in the COI context often follows criteria that differ from general journalistic or academic quality standards. For example, a local news website may publish articles in non-standard formats or imperfect English, and would likely fail to meet conventional Western quality benchmarks. Nevertheless, such a source may be highly relevant in a COI context if it is the only outlet reporting on a security incident in a specific village, for example. In such cases, this source may be crucial for answering a specific COI query, highlighting the need for a contextual rather than generic assessment of sources.

### Accuracy and currency

“Only information that is correct and valid at the time a decision is made should be used. Accuracy and currency can be achieved by cross-checking and corroborating information.”  
(ACCORD, January 2024, p. 39)

Accuracy and currency describe the degree to which information reflects the actual situation at the relevant point in time. While accuracy may be relatively easy to assess in some cases, such as the verification of names or dates, it is considerably more challenging in others, for example when assessing societal attitudes or the treatment of minorities. Information is considered current if it accurately reflects the situation at the time of research. Notably, older reports may still be valid in certain contexts, while very recent information may already be outdated in other situations due to rapid developments. (ACCORD, January 2024, pp. 39-40)

Among the COI quality standards, it is accuracy that most directly collides with the phenomenon of AI hallucinations discussed above. Hallucinations are not a marginal issue, but are inherent in the core functionality of generative AI models, which generate outputs based on probabilistic patterns rather than an understanding of truth or falsity. As with relevance, this issue arises independently of the specific AI use case. While the consequences of inaccurate information can be particularly severe when AI is used directly to support research tasks, factual inaccuracies can also be introduced or reinforced when AI is employed for merely textual or linguistic tasks, such as paraphrasing or summarising.

In terms of currency, it should be noted that generative AI models do not consistently prioritise recent information. Even when explicitly instructed to do so, AI-generated outputs may still incorporate sources or information of unclear or indeterminate temporal relevance. Moreover, in the COI context, the publication date alone is often insufficient to determine whether information still remains current. Assessing currency therefore requires contextual judgement that goes beyond what AI systems can provide by default.

### Transparency: clarity and traceability

“To ensure transparency, COI should be clearly presented, and its meaning must not be distorted. Every piece of information should be traceable to its source. Therefore, information should be fully referenced to enable readers to independently verify and assess the information.” (ACCORD, January 2024, p. 41)

Clarity, understood as the clear structuring of COI research results and the use of concise language, is an area in which generative AI can, in fact, provide clear benefits. Large language models are designed to produce fluent, coherent text, and have demonstrated strong

performance in a variety of linguistic tasks.<sup>6</sup> In this respect, AI tools may assist in improving the presentation of COI findings and thereby support compliance with this quality standard.

However, it is important to remember that generative AI models do not “comprehend” the content they generate. The high linguistic quality of AI-generated text can therefore be misleading; well-formulated sentences may mask a lack of substantive content or obscure factual inaccuracies. Furthermore, the persuasive nature of fluent language may reduce users’ tendency to critically question the output, which is particularly problematic in the case of hallucinations.

Similar risks arise with regard to traceability. Although many widely used LLMs have improved their ability to reference sources correctly, citation practices remain inconsistent and often incomplete. Since hallucinations cannot be entirely eliminated, every AI-generated reference and its content must be verified against the original source. Without systematic human cross-checking, the standard of traceability cannot be reliably upheld in AI-assisted COI products.

## 4.2 Principles for researching and using COI

### Neutrality and impartiality

“COI research should be conducted and presented in a neutral manner and not favour a particular outcome. COI service providers should be impartial with regard to their clients.” (ACCORD, January 2024, p. 42)

Generative AI models are not neutral. The data used to train them shapes their outputs, embedding implicit biases that are often neither transparent nor easily identifiable to users. This may still be an issue even when no overtly discriminatory patterns are observable. As mainstream viewpoints appear more frequently in large-scale datasets, these models tend to reproduce those perspectives by default.

Since generative models produce text by estimating probabilistic likelihoods rather than applying conceptual or normative reasoning (although some providers claim to offer something they call “reasoning”), neutrality and impartiality cannot be assumed. Unlike human COI experts, who can deliberately assess how information is selected and framed, and who remain accountable for those judgements, generative models cannot critically reflect on their own outputs.

### Equality of arms as regards access to information

“COI should be available to decision-making bodies and legal advisors of applicants in procedures for persons seeking international protection. Applicants must have access to the information a decision is based on, so that they may comment on it.” (ACCORD, January 2024, p. 43)

---

<sup>6</sup> Please note that the implications of this for issues such as plagiarism and copyright are beyond the scope of this paper.

The principle of equality of arms requires that all parties in international protection procedures have access to the same information and are able to scrutinise and challenge it. The introduction of AI into COI processes raises particular concerns in this respect, especially when such tools are used in decision making contexts. As noted earlier, many AI systems function as “black boxes”, generating outputs that are neither transparent nor reliably reproducible. When AI-supported COI findings cannot be meaningfully verified, or replicated, applicants and their legal representatives are hindered in their ability to understand and contest the information on which decisions rely. These asymmetries become even more pronounced when authorities deploy AI tools that are not equally accessible to other parties, creating an informational imbalance.

### Using public information

“To support fair procedures, publicly available information should be used. Public information is open to review and scrutiny by the applicant, experts, and the public at large.” (ACCORD, January 2024, p. 44)

In an AI-supported context, the principle of relying on publicly available information is somewhat compromised when generative models produce outputs that cannot be clearly traced back to verifiable public sources. The main difficulty here is not the opaque training data underlying the output generation, but the fact that AI-generated content can serve as COI, resulting in users being unable to determine whether it is based on publicly available information.

### Data protection

“The personal data of an applicant, as well as any information that could make the applicant identifiable, must be protected. This information should never be shared – either directly or indirectly – with the alleged persecutor.” (ACCORD, January 2024, p. 44)

Data protection constitutes a strict boundary for the use of AI in COI practice. Particularly generative AI systems operated through external platforms or cloud-based services carry risks related to data processing and storage. Any input of personal or otherwise identifiable information into such systems may expose applicants to significant harm, including the risk of unauthorised access. In light of these risks, AI tools should only be used with extreme caution when processing personal data or case-specific details of applicants.

## 5 Conclusion

This paper has examined whether and under which conditions generative artificial intelligence can be used responsibly in the field of COI. Given the relevance of COI for decisions that may profoundly affect applicants' lives, the use of AI in this context raises not only technological, but fundamentally normative questions. While generative AI has demonstrated its potential in various respects, the analysis reaffirms that COI constitutes a particularly sensitive application context. COI research is conducted using a multi-perspective approach that emphasises contextualisation and the transparent disclosure of uncertainty where necessary. Against this background, this paper has tried to move beyond questions of technical feasibility to focus on the ethical and methodological implications of AI-supported COI work. Accordingly, it examined whether AI use can be meaningfully aligned with established COI quality standards and research principles.

The analysis in chapter 4 demonstrates that several limitations commonly associated with generative AI become particularly salient when assessed against established COI quality standards and research principles. These limitations stem from structural characteristics of generative systems, including probabilistic output generation, limited contextual understanding, opaque training data, and the persuasive effect of linguistically fluent outputs. At the same time, it must be recognised that COI research has never been free from risks such as errors, biases, or incomplete knowledge, even when conducted by human experts, affecting how research questions are handled, how information is collected, which sources are prioritised, and how results are presented. It is precisely because they aim to make such risks visible and manageable that COI quality standards and research principles serve as such important normative safeguards.

Comparison with AI-supported practices, however, seems to reveal possible gaps in the applicability of these standards. When COI products are created with assistance from systems that operate fundamentally differently from human researchers, it may not be straightforward to apply standards tailored to human work seamlessly. Accordingly, these standards remain a central normative reference point, but the discussion shifts from asking whether AI use can be aligned with existing COI quality standards and research principles to asking how these standards need to be supplemented in order to address the new challenges posed by AI-supported COI work.

Chapter 2 presented and analysed various AI ethics guidelines, thereby revealing that certain core principles are considered relevant in almost all guidelines, regardless of their specific context. However, it was also emphasised that frequently cited principles are not inherently superior to any individual principles document. The relevance of ethical principles largely depends on the field of application. Taking this into account, we propose using the established COI quality standards as a starting point, and expanding them to specifically include those ethical guideline principles that address the specific challenges of using generative AI in the context of COI.

On this basis, the ethical principle of **accountability** emerges as particularly essential when dealing with the limitations associated with the use of generative AI in the context of COI. In traditional COI practice, it is generally clear who is responsible for the content of COI products and who can be held accountable in cases of inaccuracy or methodological shortcomings. In AI-

supported settings, this attribution may be less straightforward. Although responsibility formally remains with human actors, the involvement of generative AI could make lines of accountability less transparent, particularly when outputs are influenced by systems whose internal workings and training data are largely opaque. Existing COI quality standards and research principles were not developed with such scenarios in mind, which – as of now – leaves questions of accountability largely implicit.

Alongside accountability, three additional principles frequently mentioned in AI ethics guidelines appear to be particularly relevant in our context, decisively complementing existing COI quality standards and research principles: **continuous human oversight, technical robustness and safety to prevent misuse, and transparency in AI usage**. Human oversight is closely linked to accountability, serving as a safeguard against the limitations of AI systems. While AI can process and generate information efficiently, only trained COI researchers and users have the expertise to critically evaluate sources, interpret context, and recognise when information is incomplete or biased. The so-called human in the loop-principle is therefore not only a procedural step, but also a fundamental safeguard to help maintain the (ethical) soundness of COI products.

In terms of technical robustness and safety to prevent misuse, ethical concerns extend from the design of AI systems to their practical use. In particular, there is a risk that generative AI may be inappropriately used directly for decision-making, even if it is intended solely as a supporting tool. Even future AI tools developed specifically for the COI field, which may already take COI-specific contexts into account and cite sources correctly, still pose ethical risks with regard to their application in practice particularly for this reason.

Transparency, already one of the COI research principles, may gain a new dimension in the context of the potential use of generative AI in COI research; namely transparency in regard to explaining how AI tools contribute to the creation of COI products, as well as disclosing any uncertainties or limitations associated with AI-generated content. This helps users to accurately assess the reliability of COI products and understand the role of AI in their creation.

Overall, this paper argues that creating context-specific ethical guidelines for using AI would benefit the field of COI. While existing COI quality standards and research principles should continue to serve as the main reference point, they may need further clarification and enhancement to address the unique challenges of AI-assisted practices. The four additional principles of accountability, continuous human oversight, technical robustness and safety to prevent misuse, and transparency in AI usage are not intended to replace established standards or make AI tools a quick fix for more profound problems. Rather, these principles could help address challenges that current COI standards do not fully capture and provide a basis for supporting and training COI practitioners and users in the responsible use of AI, ensuring that AI-assisted COI work remains both methodologically and ethically sound.

## 6 Sources

- ACCORD: Researching Country Information. Training Manual, January 2024  
[https://www.coi-training.net/site/assets/files/1036/accord\\_researching\\_country\\_of\\_origin\\_information\\_2024.pdf](https://www.coi-training.net/site/assets/files/1036/accord_researching_country_of_origin_information_2024.pdf)
- ACUTE: Deliverable 1 – Assessment of Needs and Requirements, forthcoming
- AI HLEG – High Level Expert group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI, 8 April 2019  
[https://ec.europa.eu/newsroom/dae/document.cfm?doc\\_id=60419](https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419)
- Artificial Intelligence Act (Regulation (EU) 2024/1689), Official Journal version of 13 June 2024, Recital 27  
<https://artificialintelligenceact.eu/recital/27/>
- EUAA – European Union Asylum Agency: EASO Country of Origin Information (COI) Report Methodology, February 2023  
<https://euaa.europa.eu/publications/coi-report-methodology>
- ENNHRI – European Network of National Human Rights Institutions: Technologies, migration and human rights: the role of European NHRIs – ENNHRI scoping paper, September 2024  
<https://ennhri.org/wp-content/uploads/2024/09/Technologies-migration-and-human-rights-the-role-of-European-NHRIs-an-ENNHRI-scoping-paper.pdf>
- euractive: Clear policies for AI in journalism, imperative for ethics, 29 October 2024  
<https://www.euractiv.com/section/economy-jobs/news/clear-policies-for-ai-in-journalism-imperative-for-ethics/>
- European Commission: Ethics guidelines for trustworthy AI, 8 April 2019  
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- EY: KI in der öffentlichen Verwaltung – der Hofrat ohne Gesicht? [AI in public administration – the faceless public councilor?], 27 March 2024  
[https://www.ey.com/de\\_at/insights/government-public-sector/ki-oeffentliche-verwaltung](https://www.ey.com/de_at/insights/government-public-sector/ki-oeffentliche-verwaltung)
- Fjeld, Jessica: Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI, in: Berkman Klein Center Research Publication, 2020  
[https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3518482](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482)
- FRA – European Union Agency for Fundamental Rights: Assessing High-Risk Artificial Intelligence, 2025  
[https://fra.europa.eu/sites/default/files/fra\\_uploads/fra-2025-assessing-high-risk-ai-fundamental-rights-risks\\_en.pdf](https://fra.europa.eu/sites/default/files/fra_uploads/fra-2025-assessing-high-risk-ai-fundamental-rights-risks_en.pdf)
- Janowicz, Krzysztof: GeoAI: Is intelligence spatial? in: Rudolphina Research Magazine, 3 July 2024  
<https://rudolphina.univie.ac.at/en/geoai-is-intelligence-spatial>
- Jarrahi, Mohammad Hossein et al.: Artificial intelligence and knowledge management: A partnership between human and AI, in: Business Horizons, Volume 66, Issue 1, 2023  
<https://www.sciencedirect.com/science/article/pii/S0007681322000222>

- Kanoulas, Evangelos: Unlocking Artificial Intelligence’s Potential in COI Research, February 2025  
<https://www.refworld.org/reference/lpprs/unhcr/2025/en/149514>
- Krüger, Anne: Generative AI in knowledge work. In: Elephant in the Lab (Blog), 25 September 2024  
<https://elephantinthelab.org/generative-ai-in-knowledge-work/>
- Lauritsen et al.: Explainable artificial intelligence model to predict acute critical illness from electronic health records, in: Nature Communications, Volume 11, 31 July 2020  
<https://www.nature.com/articles/s41467-020-17431-x>
- Lu, Chan: Rethinking artificial intelligence from the perspective of interdisciplinary knowledge production, in: AI & Society, Volume 39, 2024  
<https://link.springer.com/content/pdf/10.1007/s00146-023-01839-2.pdf>
- Madan, Rohit, & Ashok, Mona: AI adoption and diffusion in public administration: A systematic literature review and future research agenda, in: Government Information Quarterly, Volume 40, Issue 1, 2023  
<https://www.sciencedirect.com/science/article/pii/S0740624X22001101>
- Möller, Hanna: Can AI be fair?, in: Rudolphina Research Magazine, 6 June 2024  
<https://rudolphina.univie.ac.at/en/how-to-make-ai-more-equitable>
- NYT – New York Times: A.I. Is Getting More Powerful, but Its Hallucinations Are Getting Worse, 6 May 2025  
<https://www.nytimes.com/2025/05/05/technology/ai-hallucinations-chatgpt-google.html>
- OeAW – Österreichische Akademie der Wissenschaft: KI-Content: Ein Fluch im Netz? [AI content: A curse on the internet?], 14 February 2025  
<https://www.oeaw.ac.at/news/ki-content-ein-fluch-fuer-das-netz>
- OHCHR – Office of the High Commissioner for Human Rights: AI and Human Rights 101, 23 October 2025  
<https://www.ohchr.org/sites/default/files/2025-10/ai-series-1-en.pdf>
- Papagiannidis, Emmanouil, et al.: Responsible artificial intelligence governance: A review and research framework, in: The Journal of Strategic Information Systems, Volume 34, Issue 2, 2025  
<https://www.sciencedirect.com/science/article/pii/S0963868724000672>
- Press Councils.eu: Media councils are developing guidelines on AI and ethics, no date  
<https://www.presscouncils.eu/media-councils-are-developing-guidelines-on-ai-and-ethics/>
- Radanliev, Petar et al.: Ethics and responsible AI deployment, in: Frontiers in Artificial Intelligence, Volume 7, 2024  
<https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1377011/full>
- RSF – Reporters sans frontières: Paris Charter on AI and Journalism, 10 November 2023  
<https://rsf.org/sites/default/files/medias/file/2023/11/Paris%20charter%20on%20AI%20in%20Journalism.pdf>

- Samios, Greg: First Do No Harm: Four Guiding Principles For Adopting Generative AI In Healthcare, in: Forbes, 12 January 2024  
<https://www.forbes.com/councils/forbesbusinesscouncil/2024/01/12/first-do-no-harm-four-guiding-principles-for-adopting-generative-ai-in-healthcare/>
- Secretary-General of the European Commission: The European Annual Asylum and Migration Report (2025), 12 November 2025  
<https://data.consilium.europa.eu/doc/document/ST-15196-2025-INIT/en/pdf>
- UNESCO – United Nations Educational, Scientific and Cultural Organization: Recommendation on the Ethics of Artificial Intelligence, November 2021  
<https://unesdoc.unesco.org/ark:/48223/pf0000381137.locale=en>
- UNESCO – United Nations Educational, Scientific and Cultural Organization: Ethics of Artificial Intelligence. The Recommendations, no date  
<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>
- UNHCR – United Nations High Commissioner for Refugees: Country of Origin Information: Towards Enhanced International Cooperation, February 2004  
<https://www.refworld.org/docid/403b2522a.html>
- University of Copenhagen: New project to map the impact of artificial intelligence on science, 3 October 2023  
<https://socialsciences.ku.dk/news/2023/new-project-to-map-the-impact-of-artificial-intelligence-on-science>
- Van Noordt, Colin, & Tangi, Luca: The dynamics of AI capability and its influence on public value creation of AI within public administration, in: Government Information Quarterly, Volume 40, Issue 4, 2023  
<https://www.sciencedirect.com/science/article/pii/S0740624X23000606>
- Xu & Shuttleworth: Medical artificial intelligence and the black box problem: a view based on the ethical principle of “do no harm”, in: Intelligent Medicine, Volume 4, Issue 1, February 2024, pp. 52-57  
<https://www.sciencedirect.com/science/article/pii/S2667102623000578#bib0022>