Ethics of AI

for Science Journalists

A Practical Framework for Responsible Reporting

Daniela Ovadia (CESJ – Center for Ethics in Science and Journalism)

Disclosure: this presentation has been refined and graphically designed with the support of Claude Sonnet 4.6 and Notebook LM

Why Does This Matter — Right Now?

93%

of newsrooms are experimenting with AI tools
(Reuters Institute, 2025)

×4

increase in AI-generated content online
since ChatGPT launch (2022–2024)

1st

major binding AI law in the world:
EU AI Act, fully in force 2026

AI is no longer a future topic — it is the present context for all science reporting.

The Regulatory Landscape

What journalists must know about AI governance

EU AI Act (2024–2026)

World's first comprehensive AI law. Risk tiers: unacceptable / high / limited / minimal. Transparency obligations for generative AI. Journalism tools may fall under its scope.

GDPR & Data Rights

AI systems processing personal data must comply. Journalistic exemptions are narrow. Critical for AI-assisted investigations and any form of automated profiling of sources.

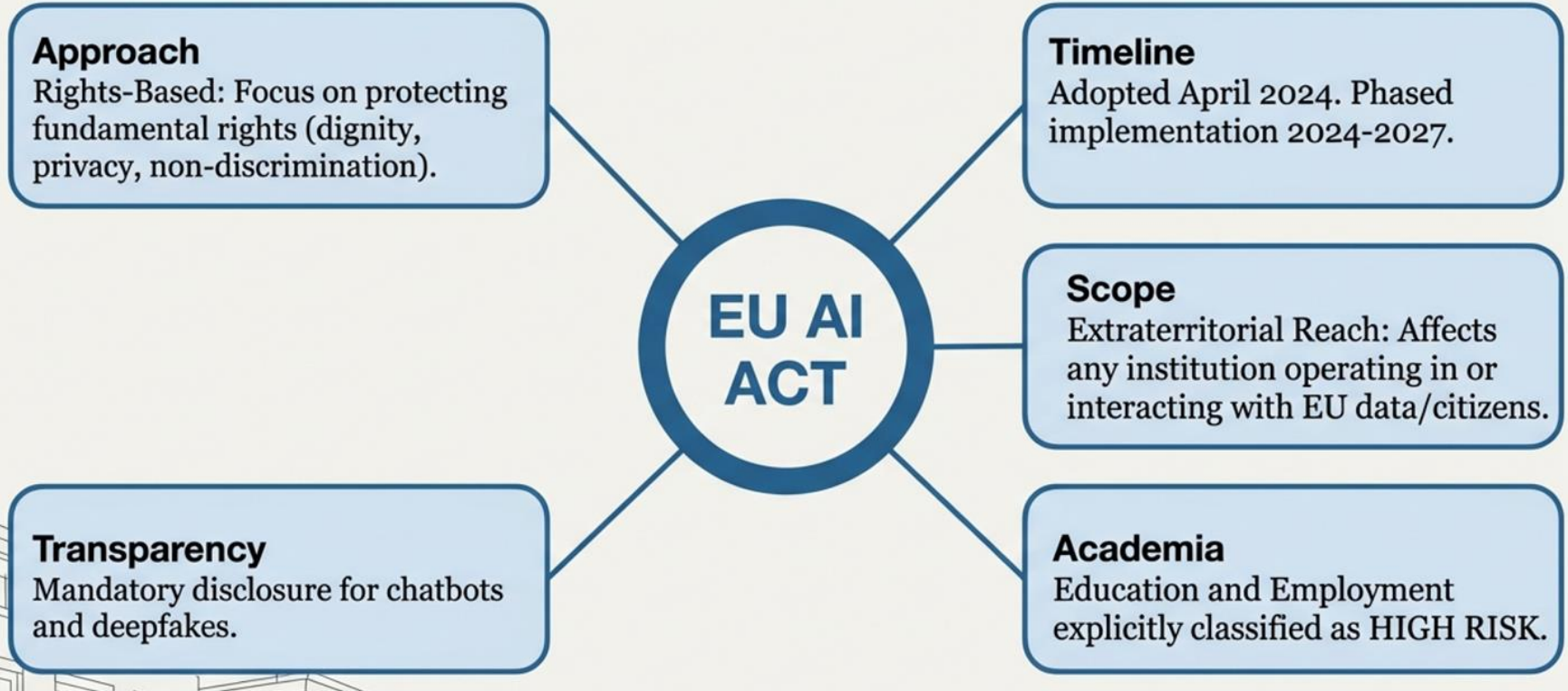
Digital Services Act (DSA, 2024)

Platforms hosting AI-generated content face new liability. Algorithmic transparency requirements affect how AI content is distributed and labelled across Europe.

Global Standards

UNESCO AI Ethics Recommendation (2021). OECD AI Principles. Council of Europe AI Convention (2024). These inform international reporting and accountability frameworks.

The Global Benchmark: The EU AI Act



Implication: The world's first comprehensive AI law sets a standard that global institutions must monitor.

Determining Liability: The EU Risk Pyramid

Unacceptable Risk: Prohibited AI Systems

- **FOR RESEARCHERS:**
- **Don't develop these systems**
- **Even for research purposes, deployment on real people is prohibited**
- **Can be studied theoretically or in simulations, but not deploy in EU**

1. SOCIAL SCORING by public authorities

- Assigning 'trustworthiness scores' based on behavior
- China's social credit system = canonical example

2. REAL-TIME BIOMETRIC IDENTIFICATION in public spaces

- Face recognition to identify people in real-time in public
- Narrow exceptions for law enforcement (missing children, preventing terrorism)
- Even then: requires judicial authorization, strict safeguards

3. EXPLOITING VULNERABILITIES OF SPECIFIC GROUPS

- AI targeting children, people with disabilities, elderly to manipulate behavior causing harm

4. SUBLIMINAL MANIPULATION

- AI using techniques below conscious awareness to manipulate

5. EMOTION RECOGNITION in workplace or educational institutions

- Exceptions: medical purposes, safety (e.g., driver drowsiness)
- Rationale: Intrusive, unreliable, inappropriate surveillance

High Risk: Compliance obligations if you deploy them

1. RISK ASSESSMENT (Article 27)
   - Identify potential harms to fundamental rights
   - Document mitigation measures
   - Update when system changes
2. DATA GOVERNANCE (Article 10)
   - Ensure training data is high-quality, unbiased, representative
   - Document data sources and curation
3. TECHNICAL DOCUMENTATION (Article 11, Annex IV)
   - How system works, performance metrics, limitations
   - Keep for 10 years
4. TRANSPARENCY TO USERS (Article 13)
   - Users must know AI is being used
   - Understand how it affects them
5. HUMAN OVERSIGHT (Article 14)
   - Human must be able to intervene, override AI
   - Human-in-the-loop or human-on-the-loop
6. ACCURACY, ROBUSTNESS, CYBERSECURITY (Article 15)
   - Meet appropriate performance levels
   - Protect against attacks and failures
7. CONFORMITY ASSESSMENT (Article 43)
   - Third-party evaluation for some systems
   - Self-assessment for others
8. CE MARKING (for products)
9. REGISTRATION in EU database (Article 71)

Limited Risk: Transparency

PRACTICAL IMPLEMENTATION:

- **Add disclaimer to chatbot interface**
- **Include statement in participant consent forms**
- **Watermark or label AI-generated images**

RATIONALE:

Users have right to know when they're interacting with AI

Prevents deception, enables informed decisions

MUST DISCLOSE TO USERS THAT THEY'RE INTERACTING WITH AI:

1. CHATBOTS

- Must inform users it's automated/AI
- Example: 'This is an AI chatbot. Responses are generated automatically and may contain errors. For human assistance, contact [email].'

2. EMOTION RECOGNITION (when allowed)

- Users must be informed AI is detecting emotions
- Example in research: Participant information sheet states AI emotion recognition is used

3. BIOMETRIC CATEGORIZATION

- Users must be informed if AI categorizes them by biometric features

4. AI-GENERATED CONTENT (deepfakes, synthetic media)

- Must be labeled as AI-generated
- Machine-readable labels for synthetic media
- Example: 'This image was generated using AI (DALL-E 3)'

Transparency Obligations (Article 52)

If your AI system:

Interacts with users (e.g., chat-style interface),

Generates content (e.g., generative search responses),

Uses synthetic content or emotion detection,

Then you **must inform** users that they're interacting with an AI and explain its capabilities/limitations in a clear, understandable way.

United States Approach

- **No comprehensive federal AI law** – Sectoral approach through existing regulations
- **Executive Order on AI (Oct 2023):** Guidelines for federal agencies, safety standards, privacy
- **Agency-specific guidance:** FTC, EEOC, FDA regulate within their domains
- **Emphasis on innovation** over regulation
- **State-level initiatives:** California, Illinois, New York developing frameworks

US favors flexible, innovation-friendly approach vs. EU's comprehensive regulation

Claude's Constitution

Our vision for Claude's character

Claude's constitution is a detailed description of Anthropic's intentions for Claude's values and behavior. It plays a crucial role in our training process, and its content directly shapes Claude's behavior. It's also the final authority on our vision for Claude, and our aim is for all of our other guidance and training to be consistent with it.

Other Global Approaches

China

- Algorithmic regulation focusing on content control
- State-centric approach with government oversight
- Social credit systems integration
- Emphasis on national security

United Kingdom (post-Brexit)

- Pro-innovation framework
- Sector-specific regulators
- Five principles: safety, transparency, fairness, accountability, contestability
- Flexible, context-based regulation

Different regulatory philosophies: EU (rights), US (innovation), China (control), UK (flexibility)

GUIDELINES ON THE RESPONSIBLE IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE SYSTEMS IN JOURNALISM

Adopted by the Steering Committee
on Media and Information Society (CDMSI)
on 30 November 2023

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

The decision by media organisations and journalists to implement AI systems

The decision to implement journalistic AI systems **should not be purely technology or commercially-driven**, but also mission-driven in that it will help achieve the goals and align with the values of the news organisation in question.

The decision to implement journalistic AI systems **constitutes an editorial decision**

Conducting a **systematic risk assessment** is an important precondition for the responsible development and deployment of journalistic AI. News organisations should have procedures in place to recognise, and where feasible, assess and mitigate risks that result from the way journalistic AI systems are implemented

Identification and acquisition of AI systems by media organisations and professional users

Once automatable journalistic tasks have been identified, there are decisions to be made about the journalistic AI systems' acquisition.

When choosing a particular AI technology provider, it is important to consider **the extent to which the technology provider has made efforts to ensure the responsible use of data**.

Incorporating AI tools into professional and organisational practice

It is recommended that organisations build and maintain this infrastructure by **hiring new staff or upskilling existing staff**. News organisations should avoid simply replacing trained journalists with technical staff

In the case of automation and use of generative AI, **editorial oversight** is required to avoid incorrect or biased processes and outputs.

News organisations should disclose when and how they use AI systems to both subjects and the audience

News organisations should provide ongoing training on the use of journalistic AI systems for staff

The use of AI tools in relation to users and society

Both the **rights and responsibilities under Article 10 of the European Convention Human Rights extend to technology, involving an obligation to use digital technology responsibly and securely**, i.e., in accordance with the ethics of journalism, aligned with professional codes, and in a way that does not impinge upon the human rights of others.

Traditional journalistic values such as fairness, autonomy, accuracy, diversity, lack of bias, truthfulness, and objectivity remain relevant in the context of journalistic AI systems

The implementation and use of some journalistic AI systems could alter the relationship with the audience and should bring audience-centred values to the fore. Key audience-centred values are transparency and explainability, accuracy, privacy and data protection, accessibility, diversity, audience members' right to form opinions and take independent decisions.

Responsibilities of external technology providers and platforms

Technology providers should recognise that although automation can help with some tasks in the journalistic production chain, journalists will require some (if not most) tasks to be completed by humans.

Technology providers should also understand some of the unique or heightened risks faced by the news media in terms of how their output is interpreted.

Platforms (that disseminate news) need to develop appropriate internal governance responses to ensure that content is universally available, easy to find and recognised as a source of trusted information by the public

Obligation of States

- States have a positive obligation to **protect and create favourable conditions for the realisation of human rights and media pluralism**. There is a need for the diversification of funding schemes to support short- and long-term projects on the development of responsible journalistic AI systems…

Core Ethical Principles

Bridging AI ethics and journalism ethics

Accuracy

Verify AI outputs through independent expert sources. Never treat model output as evidence. Distinguish 'AI says' from 'evidence shows' in your writing.

Transparency

Disclose when AI was used in research, writing or editing. Report on the opacity of systems you cover. Name the model, version, and known limitations.

Accountability

Identify who is responsible when AI causes harm. Developer? Deployer? Clinician? Accountability gaps are not a legal footnote — they are the story.

Fairness & Equity

Examine which populations were included or excluded from training data. Algorithmic discrimination in health, hiring and justice is a documented, reportable risk.

Human Oversight

Advocate for meaningful human control in high-stakes domains. AI assistance is not AI decision-making. This distinction is consequential and must be reported clearly.

Epistemic Humility

Uncertainty is scientific honesty. Avoid hype framing ('AI beats doctors'). Contextualise performance within real-world deployment conditions and actual patient outcomes.

Using AI in Your Own Journalism — Ethically

Permitted & Best Practice

- Initial literature searches (always verify findings)
- Transcription and translation assistance (with review)
- Structural editing suggestions (retaining your judgement)
- Background research on technical terminology
- Flagging potential conflicts of interest in datasets
- Accessibility: subtitles and lay summaries for wider audiences

Risks & Red Lines

- Publishing unverified AI-generated factual claims
- Using AI to fabricate or paraphrase sources
- Relying on LLM citations (they hallucinate references)
- Concealing AI use from editors and audiences
- Letting AI set the editorial angle or framing
- Using AI to generate clinical advice or patient-facing content

Reporting ON AI — Covering It Responsibly

The four most common failure modes in AI journalism

01	**Benchmark Laundering**	Reporting a single metric without context — dataset, comparator, real-world vs controlled setting. Always ask: 'Compared to what, in which population, under what conditions?'
02	**Anthropomorphism**	Framing LLMs as 'thinking', 'understanding' or 'reasoning'. These metaphors mislead audiences and inflate expectations. They are also scientifically inaccurate.
03	**Premature Efficacy Claims**	Pilot study ≠ clinical evidence. Press release ≠ peer review. Regulatory approval ≠ proven superiority. Distinguish each stage of the evidence ladder explicitly.
04	**Conflict of Interest Blindness**	Who funded the study? Who owns the AI? Disclosure standards for AI research are lagging behind other fields. Journalists must actively surface undisclosed financial relationships.

Verification, Misinformation & Source Integrity

Step 1: Confirm the Claim

Is this from peer-reviewed literature, pre-print, or a press release? Each has different epistemic weight. Never treat pre-prints as final evidence in your reporting.

Step 2: Trace the Model

Which AI system? Which version? Open or closed source? Is there a published model card? Who validated it, and on which data was it tested?

Step 3: Interrogate the Study

Prospective vs retrospective? Single-centre vs multi-site? Synthetic vs real-world data? What was the population size and how diverse was the sample?

Step 4: Seek Independent Expertise

At least one expert with no connection to the research or the company. Clinical expertise for clinical claims. Ethicists for rights-related implications.

Step 5: Label Your Own AI Use

If you used AI in your research or drafting, disclose it. Your audience has the right to know. Transparency about methods is a journalistic standard, not a weakness.

Emerging & Horizon Challenges

What is coming — and what you should be watching now

Synthetic Scientific Literature

LLM-generated papers already appear on pre-print servers and have passed peer review. The scientific evidence base itself is at risk of systematic pollution.

Agentic AI in Research

AI systems are beginning to design experiments autonomously. Who is the author? Who bears responsibility? Publication ethics frameworks are not ready.

Personalised Health Misinformation

AI can generate individually tailored health disinformation at scale. Traditional fact-checking models are wholly insufficient to respond to this emerging threat.

Epistemic Homogenisation

When all journalists use the same LLMs, coverage converges. Diversity of sources, angles and interpretations — core to good science journalism — may erode.

AI & Political use of Public Health

Deepfake scientists and fabricated clinical data are already used as political props. Science journalists are on the front line of combating these information operations.

Regulatory Gaps & Evolution

The EU AI Act does not cover all risks. Foundation models in research remain partially ungoverned. Following the legislative process is itself a reportable story.

Key Takeaways

1. AI ethics and journalism ethics are not separate domains — they are the same professional obligation in a new context.

2. Know the regulatory framework: the EU AI Act and GDPR create real obligations for the systems you report on and use.

3. Scepticism is your professional tool. Question benchmarks, demand transparency, trace funding, seek independent expertise.

4. Disclose your own AI use without apology. Transparency about your methods is foundational to public trust in science journalism.

5. The story is the system, not the headline claim. Report on governance, accountability, and real-world impact — not just metrics.