

DETECTOR: BUILDING TRUSTWORTHY AI FOR DEEPFAKE DETECTION

DETECTOR examines how Europe can develop high-risk AI for deepfake detection while safeguarding personal data and fundamental rights

Europe's artificial intelligence (AI) debate is moving from rule-making to implementation, with no higher stakes than in criminal justice. Manipulated media may shape an investigation, while authentic material may be dismissed as fake. Law enforcement authorities (LEAs) and forensic institutes therefore need more than a binary 'real or fake' score: they need methods whose results can be understood, tested, and challenged.

From detection to forensic interpretation

[DETECTOR](#) (Deepfake Evidence and Technology for Forensic Content Oversight and Research) is a 15-partner Horizon Europe project developing multimodal tools and multilingual datasets to detect manipulated or synthetic images, videos, audio and texts across Europe's linguistic and cultural contexts. Co-designed with LEAs and forensic practitioners, our tools pair eXplainable AI (xAI) with likelihood-ratio (LR) methods that express evidential strength under competing propositions. Across glass-box, partially interpretable and closed-box models, trustworthiness must attach to the full evidentiary process: LRs can structure validated

outputs, but cannot substitute for transparency, validation or auditability.

When does audiovisual evidence become biometric data?

Faces and voices may be personal data without automatically becoming biometric data. Under the General Data Protection Regulation (GDPR) and the Law Enforcement Directive (LED), biometric data result from specific technical processing that allows or confirms unique identification. A facial image or voice recording is therefore not enough. Authenticity analysis that does not create or use identifying features or embeddings may fall outside this category, but the full pipeline, including intermediate representations, must be assessed. Where the identification threshold is met, Article 9 GDPR or Article 10 LED applies.

The applicable regime turns on actor and purpose. Research organisations and providers generally fall under Article 2(1) GDPR. Competent authorities, defined by Article 3(7) LED, processing for Article 1(1) law-enforcement purposes fall under Article 2(1) LED and outside the GDPR under Article 2(2)(d). The same file may cross regimes, but

each transfer or repurposing requires separate justification.

Europe now needs a common operational rulebook, not another layer of abstract principles, for moving data from GDPR-governed research into LED-governed use. The Digital Omnibus on AI allows an Article 27 AI Act Fundamental Rights Impact Assessment to cross-reference or incorporate relevant parts of a GDPR or LED Data Protection Impact Assessment. That reduces duplication, but the assessments remain distinct. European Union (EU) guidance should integrate them without gaps: regulation should follow the actual processing, its purpose and capacity to identify, not the file format or institutional label.

When is deepfake detection high-risk AI

The decisive issue under the AI Act is not deepfakes, but intended purpose. Annex III, point 6(c), covers AI used by LEAs to evaluate evidence in criminal investigations or prosecutions; the Commission's 2026 [draft guidelines](#) expressly cite systems verifying whether evidentiary images are deepfakes. The same authentication tool used for fact-checking may therefore be

classified differently in a criminal-evidence workflow.

An AI system does not cease to be high-risk because a human makes the final decision. Where its output can shape an investigation or a suspect's treatment, safeguards must reflect that influence. Europe should treat the postponement of high-risk obligations for Annex III AI systems until 2nd December 2027 as a compliance runway, not a regulatory holiday. Articles 9-15 are lifecycle processes, not forms to be completed before deployment; the extra time should embed accountability into development, not defer it until the deadline.

Can Europe reuse pre-trained models without importing invisible risk?

The EU should reject the false choice between pre-trained models and compliance. The AI Act does not require training every component from scratch. A task-specific detector may fall outside Article 3(63)'s GPAI definition; adapting a broadly capable model does not automatically remove that status. Limited upstream documentation does not necessarily bar re-use, but the provider must still demonstrate compliance of the resulting high-risk system.

What does meaningful human oversight require?

The real bottleneck is the evidence chain. Without upstream information on training-data provenance and lawfulness, evaluation, limitations and security, downstream providers cannot demonstrate compliance. Open-source availability is not a compliance certificate: Article 53(2) exemptions do not displace Article 25 value-chain responsibilities or Articles 9-15. Personal-data processing needs an Article 6 GDPR basis and, for special-

category data, an Article 9(2) condition – not consent by default; LED processing must satisfy Articles 8 and 10. Europe's competitive weakness is not regulation, but fragmented and repeatedly duplicated proof of compliance. Shared datasets, documentation, testing environments, sandboxes and transparent European models could turn regulation from a recurring cost into a European strategic asset.

Europe should stop treating "human-in-the-loop" as proof of meaningful oversight. Operators need competence, information, time and authority to challenge or override the system. Articles 14 and 26 of the AI Act should become enforceable forensic standards requiring documented limitations, measured error rates, warnings when case material differs from validation data, intelligible uncertainty, recorded interventions and tamper-evident logs. Otherwise, humans risk becoming a procedural rubber stamp for automation bias. Systems whose analytical pathway cannot be reconstructed should not be procured.

LRs can strengthen this framework, but only if tightly governed. Consistent with the [2015 ENFSI Guideline](#), they express evidential strength under competing propositions, not the probability of guilt. A detector score is not an LR. Converting it requires explicit propositions, representative data, a validated statistical model and uncertainty analysis; otherwise, numerical precision merely disguises model uncertainty as forensic authority. LR methods must be validated, auditable and reported with limitations. Meaningful oversight is an institutional capability, not an interface feature.

Compliance is part of capability

DETECTOR points to four decisions Europe can no longer postpone: a

forensic-AI rulebook clarifying the GDPR-LED and biometric-data boundaries; common validation protocols, independent testing and lawfully sourced multilingual benchmarks; procurement rules making documentation, chain-of-custody functionality, lifecycle logging and practitioner competence mandatory, not optional deliverables; and shared standards, trusted testing environments, sandboxes and transparent European models for responsible reuse. Without this infrastructure, compliance will remain duplicated, and scale will continue to outrun demonstrable trustworthiness.

The policy lesson is blunt: criminal justice should not purchase accuracy without accountability. An AI output deserves evidential weight only when legally justified, independently validated, scrutinised and challenged by a competent human. DETECTOR's task is larger than detecting synthetic media: it must move deepfake detection from technical promise to accountable forensic practice.

Authors

Antonia Koumpoti, Martina Franciosi, Dr Luis Unzueta, and Dr Dimitrios Zarpalas



**Funded by
the European Union**



Dr Dimitrios Zarpalas
Principal Researcher
Information Technologies Institute – ITI
Tel: +30 2310 464160

[WEBSITE](#)