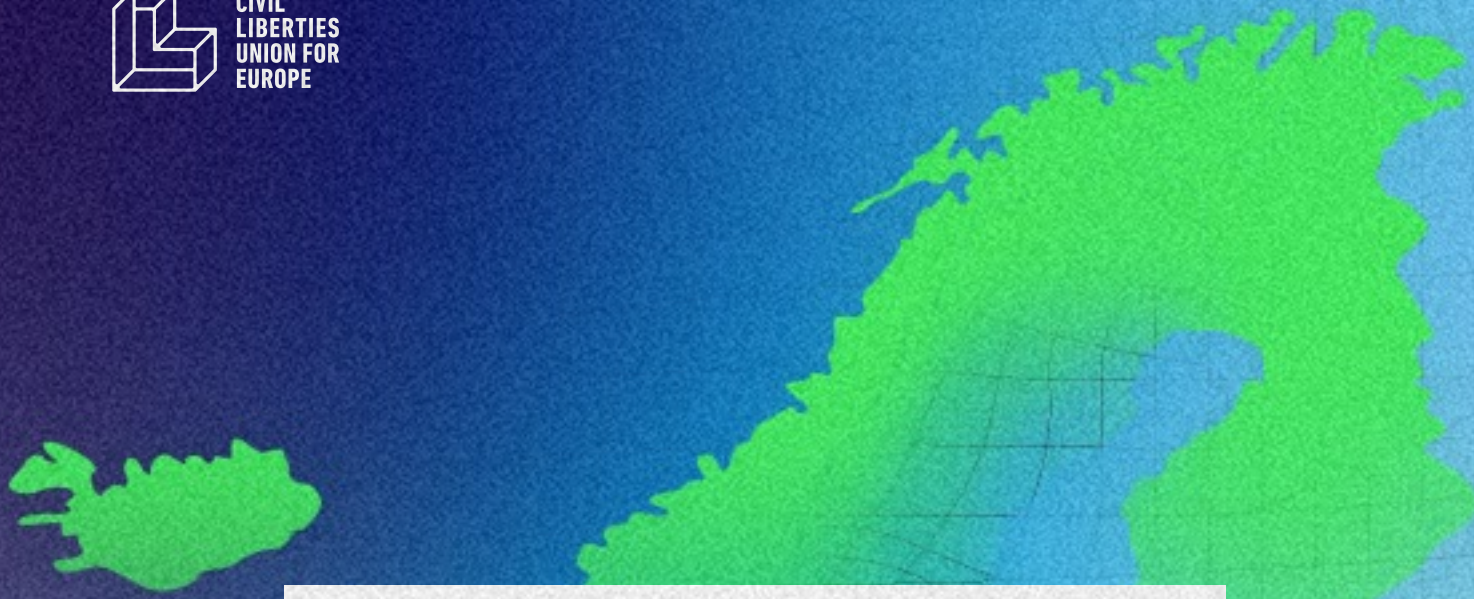
CIVIL LIBERTIES UNION FOR EUROPE



Closing the High-Risk Regulatory Gaps: Strengthening Classification and Fundamental Rights Protections under the EU AI Act

Input for the European Commission's Guidelines on implementing the AI Act's rules on high-risk AI systems.

September 2026

TABLE OF CONTENTS

Executive Summary

The European Union has played a landmark role in regulating the future of AI use and protecting fundamental rights. Despite the EU's intentions to ensure safeguards and the usability of AI alongside the highest standards of rights protection, through efficient, largely self-regulated structures manifested in its "high-risk classifications," significant gaps remain. Liberties' policy paper argues that, as currently drafted, the European Commission's Draft Guidelines on the classification of high-risk AI systems under Article 6 of the AI Act (published 19 May 2026; the "draft Guidelines") does not deliver on that promise: the "intended purpose" test and the Article 6(3) filter mechanism both depend, in the first instance, on a provider's own self-assessment. Although the draft Guidelines contain genuine textual resources that could support a more functional interpretation, this self-assessment architecture produces significant ex post regulatory enforcement problems and unresolved fundamental-rights concerns. Drawing on the text of the draft Guidelines themselves, EU legislative history, and comparative material from France, Germany, and Italy, the paper traces this problem across four areas. It proposes concrete textual and structural amendments for each.

The paper identifies four areas in which this structural weakness is particularly significant. First, the treatment of "intended use" gives providers considerable influence over whether their systems fall within Annex III at all. The draft Guidelines already contain anti-circumvention principles for complex and agentic systems but apply comparable functional reasoning unevenly to other horizontal concepts, including the "on behalf of" test and the distinction between natural and legal persons. The final Guidelines should generalise this functional approach so that classification turns on a system's actual capabilities, reasonably foreseeable uses, and real-world deployment rather than primarily on the provider's formal description.

Second, the Article 6(3) filter allows systems otherwise falling within Annex III to avoid high-risk classification where one of four conditions is met. Although these conditions are intended to be interpreted narrowly, their application still depends heavily on how providers characterise a system's function. The profiling override in Article 6(3), under which Annex III systems that perform profiling of natural persons remain high-risk regardless of the filter, is therefore an important but under-emphasised safeguard. The final Guidelines should reinforce this protection and ensure that systems that rank, prioritise, evaluate, or otherwise materially influence decisions cannot be recharacterised as merely procedural, supportive, or preparatory.

Third, paragraph 437's requirement that an AI system be "specifically intended" to influence electoral choice or turnout may risk leaving election-facing general-purpose AI systems outside the high-risk regime. Even if general-purpose chatbots and AI search interfaces are not marketed as electoral tools, many can

still provide party comparisons, voting-match scores, or other guidance that can influence voters. Liberties' research during the 2026 Hungarian parliamentary election campaign illustrates the risks posed by such systems. At the same time, conversational AI does not fit neatly within the Digital Services Act's Very Large Online Search Engine framework, creating a potential oversight gap between the two regimes. The final Guidelines should therefore address election-facing GPAI systems according to their functional and foreseeable influence on electoral behaviour rather than relying principally on how providers define their intended purpose.

Fourth, significant gaps remain in the administration of justice and law enforcement. AI-driven crime-analytics and predictive-policing tools were removed from Annex III during the legislative process. These restrictions were not reinstated, leaving important uses outside the high-risk framework and subject only to more limited safeguards. Civil litigants also lack protections comparable to those available in criminal proceedings when AI contributes to judicial or arbitral decision-making, as illustrated by France's 2019 restrictions on judicial analytics. At the same time, the undefined concept of "judicial authority" creates classification difficulties in Member States such as France and Italy, where prosecutorial and judicial functions overlap institutionally. Germany's decentralised policing structure further demonstrates how AI adoption can vary significantly within a single Member State, producing oversight gaps that national litigation alone cannot resolve.

This Liberties policy paper ultimately argues that these apparently distinct gaps share a common structural weakness: the AI Act's high-risk framework too often lets providers' formal characterisation of their systems determine whether regulatory safeguards apply, while meaningful correction occurs only after deployment. The final Guidelines should therefore adopt a more functional approach to classification centred on systems' actual capabilities, reasonably foreseeable uses, and real-world effects; reinforce the profiling override and other anti-circumvention safeguards; close regulatory gaps affecting election-facing GPAI and AI used in judicial and law enforcement settings; and strengthen transparency, documentation, and independent oversight where providers classify Annex III systems as non-high-risk.

Introduction

The European Union's AI Act plays a landmark role in regulating AI and protecting fundamental rights, and it leaves many open questions for further refinement. As AI's role across society, government, and industry expands, the EU has sought to manage risk through its high-risk system classifications and the accompanying draft Guidelines. The EU intends these classifications to safeguard both the usability of AI and the highest standards of rights protection through efficient, largely self-regulated structures. In several places, however, imprecision and over-reliance on provider-supplied definitions in the draft Guidelines create gaps and loopholes that run counter to the EU's stated goals.

The EU bases its determination of what counts as "high-risk" AI on how a provider defines a system's "intended use" and "intended purpose." That definition determines whether a system is high-risk at all, and therefore whether any of the Act's documentation and regulatory requirements apply. Even where a system's intended purpose is high-risk on its face, Article 6(3) offers a "filter mechanism": four conditions, any one of which lets the system avoid high-risk classification altogether. Even accounting for the legitimate goal of not stifling AI development, points (a)–(d), as currently drafted and interpreted, create considerable ex post regulatory enforcement problems and leave fundamental-rights concerns unresolved, because the initial classification decision sits with the party with the least incentive to classify conservatively, and correction, if it happens, comes only after the fact.

The EU offers use-case guidance to help providers navigate this system. Even so, the line between a "high-risk" and "non-high-risk" use case can be so fine that a provider's chosen language, rather than a system's actual capabilities or foreseeable impact, does the decisive work. A worker-monitoring system labelled "workflow optimisation" and one labelled "employment evaluation" can perform materially the same function while receiving entirely different regulatory treatment, letting a provider decide, in effect, whether the law applies at all. Liberties' policy paper examines, first, the draft Guidelines' over-reliance on provider-specific definitions of "intended use" and the Article 6(3) filter, and the risks that follow from overly narrow interpretation and provider self-selection; and second, the loopholes that, left unaddressed, would undermine the very goals the draft Guidelines and the wider regulatory framework are meant to serve.

Liberties' policy paper focuses on the power asymmetry the draft Guidelines create between individuals on one side and the developers, providers, and deployers who control broad-reaching AI systems on the other. That asymmetry carries weight in settings with immediate, material consequences for fundamental rights: elections and democratic processes, access to healthcare and other essential services, and the administration of justice. In each setting, provider self-selection combined with limited transparency deepens the existing imbalance between individuals and the entities that control these systems, leaving individuals with little practical recourse. Without meaningful transparency, accountability, and safeguards, oversight risks exist only on paper.

Liberties' policy paper argues that the draft Guidelines should first shift the evaluation of "intended use" away from provider control toward a multi-factor test based on technical capabilities, high-probability foreseeable uses, and real-world deployment, rather than marketing labels or terms of service. The draft Guidelines should then narrow and tighten the Article 6(3) filter conditions (a)–(d), preventing an Annex III system from defaulting to low-risk treatment by prohibiting provider self-selection that mischaracterizes evaluative or bias-laden systems as merely "narrow," "preparatory," or "supportive." The draft Guidelines should also classify GPAI chatbots and search interfaces that offer election advice or voting-match scores as high-risk under Annex III point 8(b), bringing them within the Digital Services Act's risk-assessment obligations; and, finally, the draft Guidelines should mandate an EU-wide AI disclosure certificate for criminal evidence, extend transparency and explanation rights to private legal-tech software, and define "judicial authority" to close evidentiary oversight gaps across both criminal and civil proceedings.

Each section develops this argument through the specific textual and structural features of the AI Act and the draft Guidelines. Section 1 traces the "intended use" loophole through the draft Guidelines' own horizontal provisions, the treatment of complex and agentic systems, the "on behalf of" concept, and the natural/legal person distinction, showing that the draft Guidelines already contain unevenly applied anti-circumvention tools. Section 2 works through each of the four Article 6(3) filter conditions in turn, using the draft Guidelines' own worked examples, and highlights the profiling override as an under-emphasized structural safeguard that the final Guidelines should reinforce rather than erode. Section 3 focuses on paragraph 437's "specifically intended" test for election-facing GPAI systems, drawing on Liberties' own research into chatbot behaviour during the 2026 Hungarian election campaign, and on the definitional mismatch between the AI Act's point 8(b) and the Digital Services Act's Very Large Online Search Engine regime that currently leaves GPAI election chatbots without effective oversight from either instrument. Section 4 turns to the administration of justice: how AI-driven crime-analytics tools were removed from Annex III's high-risk list at the Council's general approach stage in 2022 and never reinstated; how the criminal/civil divide in Annex III points 6 and 8(a) leaves civil litigants without the disclosure rights available to criminal defendants, as illustrated by France's 2019 restrictions on judicial analytics; how the undefined term "judicial authority" creates a genuine classification problem in Member States such as France and Italy, where prosecutorial and judicial functions sit within a single professional body; and how Germany's decentralised policing structure illustrates how unevenly AI adoption can vary within a single Member State.

Taken together, these changes would help the Guidelines fulfill their stated role: a high-risk classification regime that is both rights-affirming and workable as the AI Act's real-world deployment continues to evolve. Sections 1 through 4 take each of these four areas in turn; the Conclusion elucidates structural commonalities.

1. Closing potential loopholes in intended use

Before the Article 6(3) filter mechanism is reached, an AI system must first fall within one of the use cases listed in Annex III of the AI Act[1] (the list of specific AI applications the Act treats as presumptively high-risk, spanning areas such as biometrics, employment, education, essential services, law enforcement, and the administration of justice). This creates two distinct opportunities to escape high-risk classification. First, if a provider defines the system's "intended purpose" so that it falls outside an Annex III use case, the system never reaches the filter mechanism. Second, even where a system falls within Annex III, Article 6(3) provides a further route out of high-risk classification.[2]

The draft Guidelines' general principles for classifying high-risk AI systems make clear that how a system is "intended to be used," its "intended purpose," is the key criterion for determining high-risk status. Article 3(12) of the AI Act defines "intended purpose" as: "the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation."[3]

The draft Guidelines confirm that intended use is identical to the provider's own definition of intended purpose. If that definition does not encompass an Annex III use case, the system escapes high-risk classification under Article 6(2) of the AI Act, regardless of what it can do.

This structural feature most concerns Liberties in this section. "Intended purpose" and "intended use" should be interpreted broadly to prevent systems deployed in Annex III high-risk domains from avoiding high-risk status despite their impact on fundamental rights. Providers can strategically draft instructions for use, terms of service, and promotional materials in narrow, technical language that explicitly denies any high-risk use. For example, a worker-monitoring system can be labelled "workflow optimisation" rather than "employment evaluation." This framing offers an easy route around conformity assessments and fundamental-rights impact assessments, and it may also let providers claim legal immunity by arguing that a deployer used the system outside its stated purpose.

1 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (AI Act), Annex III.

2 Regulation (EU) 2024/1689, Article 6(2)– (3).

3 Regulation (EU) 2024/1689, Article 3(12): "'intended purpose' means the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation."

Liberties are not alone in raising this concern. In September 2023, BEUC, European Digital Rights (EDRi), Access Now, and 115 other civil society organisations jointly urged EU lawmakers to address a loophole that lets AI developers determine for themselves whether their own system is "high-risk," effectively allowing providers to decide whether the law applies to them.[4]

The draft Guidelines' own tools for narrowing this loophole, and their limits

The draft Guidelines do provide genuine textual resources for a functional reading of "intended use," and Liberties' recommendations build on these rather than inventing them from nothing. Three horizontal provisions, meaning general interpretive rules the draft Guidelines apply across multiple Annex III categories rather than to any single use case, are especially relevant.

First, the draft Guidelines clarify that where several AI systems form part of a more complex system, such that their combined intended purpose or joint outputs materially influence an individual decision, the combined configuration is treated as a single AI system for classification purposes.[5] This principle is designed to prevent circumvention by system design (such as splitting one high-risk function across several formally separate, lower-risk components). It extends to agentic AI systems that coordinate through linked actions serving a joint high-risk purpose. This is a welcome anti-circumvention rule, but its practical impact depends entirely on market surveillance authorities, the national regulators responsible for policing AI Act compliance, investigating split-architecture claims rather than accepting a provider's own account. Liberties' broader policy work calls for adequately resourced, proactive market surveillance practice on exactly this point, rather than treating the rule as self-enforcing. The initial characterisation, again, remains the providers to make.

Second, the draft Guidelines' discussion of the "on behalf of" concept in public-sector use cases is instructive by analogy.[6] Under

4 BEUC, European Digital Rights (EDRi), Access Now and 115 other civil society organisations, joint statement, "EU legislators must close dangerous loophole in AI Act" (7 September 2023), https://www.beuc.eu/letters/eu-legislators-must-close-dangerous-loophole-ai-act-joint-statement; EDRi, "AI Act: EU must protect human rights" (7 September 2023), https://edri.org/our-work/civil-society-statement-eu-close-loophole-article-6-ai-act-tech-lobby/.

5 Draft Guidelines (2026), para. 90 (referencing para. 72), confirming that an AI system otherwise satisfying an Article 6(3) filter condition remains high-risk if it forms part of a complex or agentic system whose combined intended purpose or joint outputs materially influence an individual decision within a high-risk use case.

6 Draft Guidelines (2026), Section 2.5 ("'On behalf of'"), paras. 79–81, including the anti-money-laundering worked example (an accounting firm using AI to detect money laundering to comply with its own EU AML obligations is not acting 'on behalf of' law enforcement).

this concept, a private entity is not treated as acting "on behalf of" a public authority, and therefore does not inherit that authority's high-risk classification, where the private entity acts on its own behalf, whether voluntarily or to comply with an independent legal obligation. The draft Guidelines' own example is an accounting firm using an AI system to detect money laundering to comply with its own EU anti-money-laundering obligations: this is not treated as acting on behalf of law enforcement, even though the output could plausibly feed into a later law-enforcement process. The example illustrates precisely the self-characterisation dynamic that concerns Liberties in the intended-use context more broadly: a provider's account of why a system exists, rather than an assessment of its downstream function, does the classificatory work.

Third, the draft Guidelines' natural-person/legal-person distinction (Section 2.2) means that a system evaluating only "business or company data" falls outside a use case that would otherwise apply to natural persons.[7] The draft Guidelines' own example confirms this even where the owner of a small business is assessed for creditworthiness to back a company loan, so long as the company rather than the individual is treated as the primary beneficiary. This is a defensible line in principle, but it is again a line drawn by reference to the provider's own account of who the system is "really" evaluating. That creates an incentive to frame borderline cases, such as sole traders, small partnerships, or gig-economy platforms scoring individual workers through a nominally corporate intermediary, on the more favourable side of the line.

Generally, Annex III systems are considered high-risk, and any deviation is exceptional and requires supporting evidence and documentation. This does not, however, resolve the structural problem that the regulated entity must make the initial classification decision. The implementation architecture does include ex post corrective measures: under Article 80, market surveillance authorities may reassess systems that a provider has not classified as high-risk and, if necessary, treat them as high-risk;[8] under Article 99, penalties may follow where misclassification is found to have been used to circumvent the AI Act's requirements;[9] and market surveillance and fundamental rights authorities may also request documentation and, in justified cases, technical testing. All this matters, but it functions essentially as an ex post protective mechanism. By the time reclassification takes effect, the system has often already affected people's rights or access.

7 Draft Guidelines (2026), Section 2.2 ("Limitations in some use cases to natural persons"), paras. 72–73, including the worked example that a system establishing the credit score of a small business owner using only business or company data is not intended to evaluate a 'natural person.'

8 Regulation (EU) 2024/1689, Article 80 (procedure for dealing with AI systems classified by the provider as non-high risk in application of Annex III).

9 Regulation (EU) 2024/1689, Article 99 (penalties).

The sector-specific sections already included in the draft Guidelines offer a workable template for the functional approach Liberties recommends.[10] In the education and employment sections, the draft Guidelines identify AI systems through the specific decision-making roles they play systems intended to evaluate learning outcomes, assess the appropriate level of education, recruit or select natural persons, or manage work-related relationships. Here, the provider's chosen market label does not do the classificatory work; the functional capability does, through the intended outcomes and the product's capacity to manage, evaluate, recruit, and assess. There is no principled reason this same functional logic could not be generalised as the default interpretive method across all of Annex III, rather than applied unevenly, section by section.

Toward a functional standard

The most direct improvement would be to shift the concept of "intended use" away from a provider-controlled monopoly and toward a rebuttable, multi-factor legal test that incorporates reasonably foreseeable misuse alongside the documented purpose. This approach would produce a more functional classification, at the cost of some increase in compliance workload, a cost that seems justified given the fundamental-rights stakes described above.

That shift would need to be paired with strengthened documentation requirements specifically for systems that fall within an Annex III use case but that a provider has classified as non-high risk, whether by defining the intended purpose narrowly enough to avoid Annex III altogether, or by invoking the Article 6(3) filter discussed in Section 2. Providers should be required to produce, for any such system, a detailed classification memo, a verifiable decision tree, a use-case inventory, a list of prohibited and permitted configurations, and a description of known real-world abuses, together with mandatory independent preliminary testing, third-party testing, or randomised regulatory sampling audits, particularly in employment, essential services, the judiciary, and law enforcement.

None of this will function without correspondingly strengthened testing channels for fundamental rights authorities: a dedicated pool of technical experts, common investigation protocols, an expedited emergency reporting channel, and mandatory notification of the market surveillance authority whenever a fundamental rights authority disputes a system's actual decision-making power. The liability dimension also merits strengthening. Where a provider intentionally, or unintentionally, defines the intended use too narrowly and thereby excludes the system from the high-risk regime, this should carry specific legal consequences: a reversal of the burden of proof in certain legal disputes, more severe sanctions, and exclusion from the safe harbour, meaning the limited protection from liability that

10 Draft Guidelines (2026), Annex III point 3(c) ("assess the appropriate level of education") and point 4(b) ("manage work-related relationships").

a provider otherwise receives by following the Act's documented self-assessment procedure in good faith. The right to file complaints, whistleblower protection, and market surveillance reclassification are existing foundations, but, as argued above, they are not sufficient on their own. Finally, the anti-circumvention logic the draft Guidelines already apply to complex and agentic AI systems should extend to the "on behalf of" and natural/legal person distinctions discussed above, so that the same functional standard governs every horizontal concept that determines Annex III scope, rather than only the one currently addressed.

2. Filter Mechanism

Under Article 6(3) of the AI Act, an AI system listed in Annex III can escape high-risk classification even after falling within one of Annex III's use cases, if it meets one of four specified conditions and does not pose a significant risk of harm to health, safety, or fundamental rights. This is the second escape route referenced in Section 1: the first is defining the intended purpose narrowly enough to avoid Annex III altogether; this second route applies only once a system has already been acknowledged to fall within Annex III and asks whether it can nonetheless be filtered back out. The draft Guidelines interpreting Article 6(3) stress that conditions (a)–(d) must be interpreted "narrowly" to protect fundamental rights.

Even so, the mechanism's heavy reliance on providers' own self-selection, documentation, and system interpretation currently introduces too many gaps to secure those protections or regulatory efficiency. As noted in Section 1, defining a product's "intended purpose" by the provider's own drafting, rather than by the system's demonstrated capabilities, creates numerous problems. Because providers draft the intended-purpose statement and supply the primary evidence against which that statement is judged, interpretive authority is almost entirely unilateral. The draft Guidelines themselves acknowledge how consequential this self-assessment is: where the filter applies, neither the provider obligations nor the deployer obligations set out in Chapter III of the AI Act (the chapter containing the Act's core high-risk compliance regime) apply to that system at all. The stakes of getting the initial characterisation right, or of a provider getting away with getting it wrong, are correspondingly high.

It is worth clarifying the filter's capabilities. The draft Guidelines confirm that Article 6(3) is exhaustive but alternative: a system need only satisfy one of the four conditions, and there is no separate, independent assessment of significant risk of harm besides those four conditions. Critically, however, the filter is subject to a hard override that Liberties considers under-emphasised in public discussion of the mechanism. An Annex III system is always classified as high-risk, regardless of which of the four conditions it might otherwise satisfy, where it performs profiling of natural persons

within the meaning of Article 4(4) GDPR,[11] that is, automated processing of personal data to evaluate personal aspects such as work performance, economic situation, health, personal preferences, reliability, behaviour, location, or movements. This profiling override is a meaningful structural safeguard, and Liberties' recommendations below are intended to reinforce it, not replace it. If the four conditions are drawn narrowly on paper but ambiguously in application, providers may characterize plainly evaluative systems as falling outside profiling altogether.

This structural unilateralism runs through each of the four filter conditions.

Point (a): narrow procedural task

Point (a) requires that systems allowed to be "filtered" outperform only narrow procedural tasks. Recital 53 gives as examples systems that transform unstructured data into structured data, classify and categorise incoming documents, or detect duplicates.[12] The draft Guidelines' own boundary here can be arbitrary in application. Paragraph 92 specifies that AI systems which categorise, or change the format, structure, or presentation of data, can qualify as performing a narrow procedural task; the draft Guidelines' own worked example is a system that sorts incoming school or university applications by grade or educational level applied for, without evaluating suitability. Paragraph 93, by contrast, excludes systems that perform a "value judgement of data relevant for decision-making," for example labelling input as "useful" or "less useful" for a human assessor, or assigning a score or ranking.[13] The draft Guidelines' migration-context example is telling: a system that scans, converts, and files visa documents into predefined folders can benefit from the filter, but only provided it does not rank documents in a way that evaluates, filters out, or hides them, does not label material as "useful" or "less useful," and does not suggest next steps or credibility assessments. That is a fine distinction to draw and police in practice, and it depends heavily on how the provider describes the system's function in its own documentation. The final Guidelines should therefore clarify that the assessment under point (a) must hinge on the system's actual function in modulating which information is presented to a user and how, rather than on the provider's choice of labels. Systems that merely correct or standardise text, such as ordinary word processors or spell-check tools, can appropriately be treated as narrow procedural tasks where they do not alter the substance of what is available to the human reviewer. Systems that prioritise, suppress, or filter information in a way that

11 Regulation (EU) 2016/679 (GDPR), Article 4(4) (definition of profiling).

12 Regulation (EU) 2024/1689, Recital 53 applied ambiguously.

13 Draft Guidelines (2026), paras. 92–93, verbatim: para. 92, 'This condition may therefore cover AI systems that are intended to categorise, change the format, structure or presentation of data, or change its metadata,' including the school/university application-sorting worked example; para. 93 excludes systems that perform a 'value judgement of data relevant for decision-making.'

affects what the reviewer is likely to perceive or rely on should not qualify. The final Guidelines should also indicate what mechanisms will verify that systems framed as data-restructuring tools are not substantively performing evaluative tasks under another label.

Point (b): improving the result of a previously completed human activity

Point (b), covering AI systems intended to improve the result of a previously completed human activity, presents additional gaps. The draft Guidelines require three cumulative elements: a human activity must have been completed, that completion must have led to a result, and the AI system must improve that result without reverting to, or fully revising or replacing, the human activity. The draft Guidelines are explicit that the legislature deliberately chose "improve" rather than "review," precisely to exclude systems that provide a materially different result from the one a human reached. Their own worked examples draw the line at systems that flag contradictions or errors in already-finalised human work, or that map conclusions to evidentiary records to strengthen traceability, as against a system that, in the draft Guidelines' words, "checks a decision, plan, or construction made by a human and provides a substantially different solution." That distinction is sound in principle, but in practice it again depends heavily on the provider's own framing of what the system does, rather than on an independent assessment of the tool's actual output behaviour across cases. The final Guidelines should therefore specify that point (b) applies only where the system is limited to identifying clerical errors or mismatches in already-completed human work. Where the same model can effectively rewrite the underlying account, classification should focus on the system's functional capabilities rather than the provider's framing of its output. The Commission should clarify, in particular, that systems capable of materially altering the evidentiary or factual picture cannot be filtered out under point (b), and should require providers to document empirically, not merely assert, the rate at which the system's output diverges from the human-completed result it is meant to be improving.

Point (c): detecting decision-making patterns

Point (c), covering systems that detect decision-making patterns or deviations from prior patterns without replacing or influencing the prior human assessment, introduces the greatest risk to fundamental rights protections among the four conditions, because it is the only one of the four that can launder a biased baseline forward without ever being asked to justify that baseline. The draft Guidelines permit such a system to inform a human review that in turn influences or even replaces the prior assessment, if review is "proper," meaning meaningful, comprehensive, and considerate of the original decision's underlying elements. Nothing in this framework, however, asks whether the pattern being used as the comparator was itself sound. If a prior decision-making pattern was itself discriminatory or overly biased, whether instructively or unwittingly, a pattern-deviation system built on that baseline will treat any corrective departure from that

pattern as anomalous, flagging the correction rather than the original bias and effectively entrenching it. The draft Guidelines' own example, involving a system that detects deviations in public-benefits eligibility decisions for quality-assurance purposes, does not address what happens where the historical decisions being compared against were themselves systematically skewed against a particular group.

The final Guidelines should include three specific solutions. First, they should clarify that systems falling under point (c) cannot benefit from the filter where the baseline pattern against which deviations are measured may reflect historical discrimination or other unlawful bias. Second, providers relying on point (c) should be required to demonstrate that the historical decision pattern used as the comparator has been assessed for discriminatory effects, and that the system will not flag a corrective departure from a discriminatory pattern as anomalous. Third, the "proper human review" standard should be amended to explicitly require the reviewer to consider whether a flagged deviation might represent a correction of past unlawful bias rather than an error, with accessible human review and internal challenge procedures encouraged for any decision influenced by such a system.

Point (d): preparatory tasks

Point (d) treats the term "preparatory" as determinative. The draft Guidelines tie this to the task's proximity to the final human decision: preparatory tasks such as indexing, searching, processing, and linking add structure to inputs without themselves producing an assessment or outcome. The draft Guidelines are explicit that where a system's output is intended to inform a human operator's assessment, it counts as preparatory only where that output is "a general input or factor" rather than "a specific recommendation or evaluation of the case." In principle, this is a workable standard. In practice, most AI functionalities can plausibly be reframed as preparatory to a later human act, because "preparatory" is defined by the task's position in a workflow rather than by any inherent property of the task itself, so almost any function can be described as coming before a later human sign-off. The worked examples the draft Guidelines offer again depend substantially on the provider's own framing of the workflow and of what its system's output represents. The final Guidelines should therefore clarify that an AI system cannot benefit from the point (d) filter simply because a formal human decision occurs later in the workflow. Providers should instead be required to demonstrate that the system does not materially influence the follow-up human assessment. Where a system assigns scores, weights inputs, ranks or prioritises cases, recommends an outcome, or otherwise affects how the reviewer evaluates the case, it should not be regarded as performing a merely preparatory task, regardless of whether the provider characterises its output as "supplementary information" rather than a recommendation.

Self-assessment, registration, and the limits of ex post correction

The filter's operability depends on provider self-assessment across all four conditions. The draft Guidelines are explicit that whether a

system may benefit from Article 6(3) depends on the provider's self-assessment. Article 6(4) requires the provider to document that assessment, covering the system's intended purpose, why it would otherwise qualify as high-risk, which of the four conditions is considered to apply and why, and why the system does not perform profiling, before the system is placed on the market.[14] Providers relying on the exemption were, and remain, also subject to a registration duty under Article 49(2), so that deployers and, in principle, the public could verify use of the exception in the EU database. The Commission's original Digital Omnibus proposal would have deleted this duty for Article 6(3) self-assessed systems; the Council and Parliament rejected that deletion, and the Digital Omnibus on AI, which entered into force on 27 July 2026, instead simplified the information such providers must submit rather than removing the registration requirement.[15] The practical effect is a lighter but still public registration duty, closer to what civil society had pushed for than to the Commission's original proposal; the final Guidelines should still clarify how much the simplified Annex VIII information requirements narrow what the public can actually learn from a given registration entry.

This is a real safeguard, but it remains a documentation-and-disclosure model rather than a pre-market verification model. Nothing requires a regulator to review the assessment before the system reaches the market. The correction mechanism is again the ex-post one described in Section 1: Article 80 empowers market surveillance authorities to evaluate a provider's non-high-risk classification and request corrective action, and Article 99 allows penalties where a system was misclassified specifically to circumvent the AI Act's requirements. As with the intended-use loophole more broadly, the corrective mechanism operates only after a misclassified system has already reached the market and, in many Annex III domains, has already affected real decisions about real people.

The final Guidelines should clarify that meaningful human involvement, referenced throughout the four conditions, requires that the reviewer have sufficient information, competence, time, and authority to assess the

14 Regulation (EU) 2024/1689, Article 6(4).

15 Regulation (EU) 2024/1689, Article 49(2) (registration of high-risk AI systems and of non-high-risk systems self-assessed under Article 6(3)). The Commission's original Digital Omnibus proposal (19 November 2025) would have deleted the Article 49(2) registration duty entirely for Article 6(3) self-assessed systems; the Council and European Parliament rejected full deletion and reinstated the registration obligation, while simplifying the Annex VIII information that must be submitted. See Lewis Silkin, "The Digital Omnibus on AI enters into force today" (27 July 2026); Freshfields, "EU AI Act unpacked #34" (2026); Cooley, "Digital AI Omnibus Delays Key Deadlines, Introduces New Rules" (2026). Civil society, including a February 2026 open letter from the European Disability Forum and others, had opposed the original deletion proposal: https://www.edf-feph.org/publications/a-call-to-eu-legislators-protect-rights-and-reject-the-call-to-delete-transparency-safeguard-in-ai-act/.

system's output independently, and a genuine possibility to depart from it without adverse consequences. This standard should be made explicit and uniform across all four points rather than left to be inferred separately in each.

3. Elections and referendums

The Commission should take further steps within the draft Guidelines to secure the integrity of elections, referendums, and other democratic processes. The draft Guidelines prioritise the objective a provider describes for a system over its functional and foreseeable impact on an election. This logic particularly favours search, recommendation, and chat systems that, while not marketed as "election AI," can nonetheless guide party or candidate selection during a campaign period.

Recital 62 explains the rationale for classifying election-facing AI systems as high-risk in the first place:[16] to address the risks of undue external interference with the right to vote enshrined in Article 39 of the Charter of Fundamental Rights,[17] and of adverse effects on democracy and the rule of law. The Recital carves out one exception, for AI systems whose output natural persons are not directly exposed to, such as tools used to organise, optimise, and structure political campaigns from an administrative and logistical point of view.[18] That exception is sensible on its face. A campaign's internal volunteer-scheduling tool is not the same kind of risk as a public-facing chatbot. But its boundary depends on the same self-characterisation dynamic discussed in Sections 1 and 2: a provider need only show that the natural persons whose votes are ultimately at stake are not "directly exposed" to a given output for the exception to apply, and a system's outputs can shape voter-facing content indirectly, through campaign targeting or messaging decisions, without any voter ever directly seeing the system's raw output.

Paragraph 437 and the "specifically intended" test

Liberties' own research indicates that GPAI chatbots are unreliable sources of information during election campaigns, at times providing false or misleading information capable of shaping voters' opinions and behaviour.[19]

16 Regulation (EU) 2024/1689, Recital 62.

17 Charter of Fundamental Rights of the European Union, Article 39 (right to vote and stand as a candidate).

18 Regulation (EU) 2024/1689, Recital 62.

19 The underlying research is documented across Day, J. and Simon, E. 'Should AI Tell You Who to Vote for? Testing Political Advice in General Purpose AI systems during the 2026 Hungarian Election Campaign, https://www.liberties.eu/f/uz6rz9, 2026. Summarized in Raul Arning, "2026 Hungarian Elections and AI: ChatGPT Will Not Tell You to Vote for Orban's Contender, but Will Give You Very Confident Advice," *Civil Liberties Union for Europe*, 10 Apr. 2026, https://www.liberties.eu/en/stories/hungary-elections-ai/45665; and Eva Simon, "What Happens When AI Gives You Voting Advice?," *Civil Liberties Union for Europe*, 13 July 2026, https://www.liberties.eu/en/stories/ai-voting-2026-jul/45754.

In our study of how ChatGPT and Gemini responded to political questions during the 2026 Hungarian parliamentary election campaign, neither tool disclosed a clear, consistent methodology for how it arrived at a party match or comparison with responses varying meaningfully in ways a voter would have no way to detect or audit. The final Guidelines should shift from a formal "intended purpose" test toward the functional, foreseeable influence a system has in a campaign context.

Paragraph 437 of the draft Guidelines introduces a "specifically intended" test for point 8(b) of Annex III[20] (the use case covering AI systems intended to influence the outcome of an election or referendum, or the voting behaviour of natural persons):[21] an AI system must be "specifically intended to have an effect on the electorate's choice or turnout" to fall within that use case. As currently drafted, and as commentary on the draft Guidelines has already noted, this test places a GPAI chatbot that declines to give voting advice outside the scope of point 8(b) by default, alongside general content-recommender systems that merely surface political content incidentally, campaign management and donor-analytics tools, and human-reviewed generative drafting of political content. What remains squarely in scope under the current draft is AI-targeted political advertising, AI-enabled political chatbots and spokespersons, dedicated voter-advice applications, and constituency-boundary analysis tools.

Liberties considers the treatment of general-purpose chatbots the central weakness in the current draft. A disclaimer of the kind described above is frequently inconsistent with the rest of a chatbot's output, and it is an insufficient safeguard where the same system goes on to offer substantive voting guidance, whether through voting-match scores or party and candidate comparisons, in the same or a later exchange. The mere fact that a general-purpose AI system replies, when asked directly, that it cannot give voting advice is not by itself a sufficient safeguard. Where that disclaimer is accompanied or followed by additional text that in substance offers voting advice, the system should fall within the scope of the high-risk obligations.

The Digital Services Act as a parallel, relevant lever

The AI Act is not the only instrument available to address this gap, and Liberties considers the draft Guidelines' current treatment of the interaction with the Digital Services Act (DSA) a missed opportunity. Under Article 33 DSA, providers of Very Large Online Platforms (VLOPs) and Very Large Online

20 Draft Guidelines (2026), para. 437, verbatim: 'Only AI systems intended to be used to influence the outcome of an election or referendum or the voting behaviour of natural persons, i.e. the AI system would have to be specifically intended to have an effect on the electorate's choice or turnout, fall within the scope of the use case listed in point 8(b) of Annex III.'

21 Regulation (EU) 2024/1689, Annex III, point 8(b).

Search Engines (VLOSEs), those with at least 45 million average monthly active recipients in the Union, are designated by the Commission and subject to heightened obligations, including the systemic risk-assessment duty in Article 34. Article 34(1)(c) requires designated VLOPs and VLOSEs to assess, at least annually, any actual or foreseeable negative effects of their service on civic discourse and electoral processes, and Article 35 requires placing reasonable, proportionate, and effective mitigation measures in response.[22]

As of 31 August 2026, the DSA framework's relevance to conversational AI has increased significantly with ChatGPT's new designation as a VLOSE and accompanying increased compliance requirements on "public security and electoral process." [23] With over 159 million monthly search users across the EU, ChatGPT is at over three times the 45 million threshold number for DSA and is exemplar of the pervasive role of chatbot usage.[24] As the first chatbot under this framework, and with its regulations set to take effect in January of 2027, ChatGPT is considered a "hybrid service," by the Commission as it responds to prompts and conducts web-searches.[25] Such a classification like the designation currently offered by the Commission does not separate out ChatGPT's prompt and web-search functionalities, thereby failing to clarify whether the mechanism where the voting-match scores and electoral comparisons are generated (the conversation layer) is included in ChatGPT's regulatory scope.[26] Moreover, since DSA's regulatory scheme is triggered by user numbers rather than function, this leaves room for many chatbots with similar electoral influence capabilities to slip through DSA's schema altogether.

The underlying challenge is terminological rather than substantive. AI chatbots and AI-native search interfaces do not fit neatly within the DSA's concept of an "online search engine," which was drafted with traditional link-based search in mind, and a conversational AI system generating a direct, synthesised, personalised answer sits uneasily within categories built around indexed results and content moderation. This blurring of lines between

22 Regulation (EU) 2022/2065 (Digital Services Act), Article 33 (designation of VLOPs/VLOSEs, 45 million average monthly active recipients' threshold), Article 34 (systemic risk assessment, including Article 34(1)(c) on civic discourse and electoral processes), and Article 35 (mitigation of systemic risks).

23 European Commission (2026), *DSA*

24 Vermeulen, M. and Lemoine, L. 'What ChatGPT's DSA Designation Means for OpenAI and the EU', *Tech Policy Press*, 4 September 2026, https://www.techpolicy.press/what-chatgpts-dsa-designation-means-for-openai-and-the-eu/.

25 Ibid.

26 Stan, A. M. 'The EU has regulated ChatGPT as a search engine. Calling it a platform would have handed OpenAI a liability shield', *The Next Web*, 4 September 2026, https://thenextweb.com/news/eu-chatgpt-vlose-designation-safe-harbour-gap.

search engines, platforms, and AI products creates a governance gap that neither instrument currently closes on its own. Where a provider's chatbot or AI search interface is not itself designated a VLOSE, none of the DSA's systemic risk-assessment machinery applies to it either, leaving only the AI Act's own point 8(b) classification, subject to the paragraph 437 gap described above, as a backstop.

Revising the draft Guidelines' treatment of election-facing AI

The final Guidelines should address GPAI models directly and retain the existing example of "AI-enabled voter advice applications" as a use-case example. Paragraph 437 should be updated to specify that chatbots providing election advice, including voting-match scores and suggestions intended to inform voting behaviour, fall within point 8(b) of Annex III even where such outputs are accompanied by a disclaimer stating that the system cannot provide voting advice, unless the chatbot's only substantive reply is that it cannot give voting advice. Relatedly, the draft Guidelines' current safe-harbour language, meaning the provision that lets a system avoid high-risk treatment entirely by showing it carries "sufficient safeguards against a use influencing electoral processes" (such as a chatbot that clearly replies it cannot give voting advice, or that provides only neutral, factual, informational content about elections), should be clarified to exclude chatbot voting-advice use cases from that safe harbour, except where the chatbot's only reply is that it cannot give voting advice.

The Recital 62 exception for campaign-organisation tools "whose output natural persons are not directly exposed to"[27] should likewise be clarified so that indirect voter exposure, for example where a tool's targeting or messaging recommendations shape what a voter is later shown even though the voter never sees the tool's own output, does not qualify for the exception. More broadly, general-purpose AI systems used in relation to elections and referendums should be transparent, well-documented, evaluated, non-manipulative, secure, and continuously monitored for risks to fundamental rights, democracy, and the rule of law. The final Guidelines should cross-reference the Digital Services Act by system use case, so dual oversight under the DSA and the AI Act helps close the regulatory gap around GPAI voting advice.

Aligning the DSA response

Closing this gap also requires action on the DSA side. DSA enforcement bodies should interpret the concept of "online search engines" functionally rather than purely technically in relation to GPAI systems that retrieve, process, and present information from across the internet in response to user queries and perform a search-like function. The Commission should clarify that AI-generated answers may constitute search results "in any format" where they serve an access-to-information function,

27 Regulation (EU) 2024/1689, Recital 62.

so that providers cannot avoid DSA oversight merely because information is presented as a generated answer rather than as hyperlinks. ChatGPT's designation is an important step in the solution, but it is not the answer. Where AI search interfaces or conversational GPAI tools perform search-engine functions and meet the relevant user thresholds under Article 33 DSA, they should be designated as Very Large Online Search Engines accordingly. These GPAI systems should also be regulated according to the chat function, thereby ensuring that the conversational layers and actual systemic elements generating election advice are also included under the regulatory body.

AI-generated electoral guidance should, in turn, fall within the DSA's systemic risk-assessment framework for civic discourse and electoral processes under Article 34(1)(c) DSA, with risk assessments covering misleading or biased party matching, unequal visibility of political actors, and inaccurate political recommendations, conducted both before and during election periods. None of this will close the gap identified above, however, unless regulators combine the DSA's systemic risk oversight of VLOPs and VLOSEs with the AI Act's GPAI systemic-risk management obligations under Article 55, rather than treating the two regimes as parallel silos with no shared evidentiary base.[28]

4. Judicial Integrity

The AI Act's Annex III classifications present numerous law-enforcement use cases for AI systems. Though these use cases vary in their individual utility, several key gaps remain in the evidentiary processes themselves and in how these systems ultimately function in relation to judicial decision-making. Law enforcement's actions, particularly regarding criminal investigations, are eventually reviewed by judges, with final decision-making considered subject to human oversight.

Absent the post-remote biometric identification (RBI) reporting and documentation requirements of Article 26(10),[29] other high-risk law enforcement systems under Annex III points 6(a)–(e) carry only general documentation duties. The gap is even worse for AI-driven crime-analytics tools. The Commission's original 2021 proposal included broader law-enforcement analytics and predictive-policing use cases. The Council deleted these use cases from Annex III at the general approach stage in December 2022, alongside deep-fake detection by law enforcement and verification of the authenticity of travel documents. They were not reinstated during the subsequent trilogue negotiations that concluded in December 2023. The Digital Omnibus on AI, finalised in 2026, did not reinstate them either. While the Digital Omnibus on AI's amendments defer high-risk compliance deadlines and adjust

28 Regulation (EU) 2024/1689, Article 55 (obligations of providers of general-purpose AI models with systemic risk).

29 Regulation (EU) 2024/1689, Article 26(10).

several procedural obligations, they do not alter Annex III's list of law-enforcement use cases.[30] These tools, used to gather, organise, and visualise investigative data that may go on to inform prosecution files, now sit among the non-high-risk tier, governed only by the voluntary codes of conduct available under Article 95.[31]

The draft Guidelines confirm this reading directly and go further than the bare text of Annex III in excluding such tools from point 6. Section 3.6.7 of the draft Guidelines, addressing AI systems falling outside point 6's five use cases, states that those use cases primarily address situations where AI systems are applied directly to natural persons or have the capacity to materially influence decision-making, and gives as an example a system that sits close to the crime-analytics scenario itself: an AI-enabled crime-linkage tool that analyses police reports and case files to identify patterns across incidents and link them into possible "series" attributable to the same offender, which the draft Guidelines place outside point 6(e) specifically because its output is "not linked to a particular individual."[32] This is a meaningful omission given that such systems are acknowledged to be prone to repurposing under the Act's own "reasonably foreseeable misuse" concept: a tool built for pattern-based crime analytics can shift toward individual-based profiling without triggering any change in its regulatory classification. Without reporting requirements, tailored or general, judges tasked with evaluating evidence may lack any understanding of its origin or provenance. The trust underlying a free and fair trial is implicated where judges themselves are left unaware of the particularities of AI's role in processing evidence, particularly given that ECtHR doctrine ties the equality-of-arms and adversarial-hearing guarantees under Article 6 ECHR directly to a defendant's ability to contest the admissibility, reliability, and completeness of evidence,[33] the very function that point 6(c) systems are meant to perform.

The lack of classification of AI systems used internally by parties in court proceedings raises additional risks. There may be concerns that placing additional regulation on individual

30 Regulation (EU) 2026/1744 (Digital Omnibus on AI), published in the Official Journal 24 July 2026, entered into force 27 July 2026. The Omnibus defers Chapter III standalone high-risk obligations under Annex III to 2 December 2027 and Annex I embedded high-risk obligations to 2 August 2028; it does not amend Annex III's list of law-enforcement use cases and does not reinstate crime-analytics or predictive-policing tools. See Freshfields, "EU AI Act unpacked #34: The final Digital Omnibus on AI" (2026); Dastra, "Digital Omnibus on AI: Parliament Votes, Deadlines Redrawn" (2026).

31 Regulation (EU) 2024/1689, Article 95 (codes of conduct for voluntary application by non-high-risk systems).

32 Draft Guidelines (2026), §3.6.7 ("Other AI systems falling outside the use cases of point 6 of Annex III") and §3.6.6 (crime-linkage example, Point 6(e)).

33 Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights), Article 6.

parties would stifle judicial efficiency. Still, this omission implicates both the values underlying judges' role as fact evaluators and the right to a fair trial.

Civil and criminal asymmetry

This asymmetry sharpens once the criminal/civil distinction is drawn out explicitly. Annex III point 6 is definitionally limited to criminal matters: the Act's own definition of "law enforcement" ties the term to the prevention, investigation, detection, or prosecution of criminal offences,[34] so none of point 6's safeguards, thin as they are, extend to civil proceedings at all. Point 8(a), by contrast, is not so limited. It covers AI systems used by a judicial authority to research and interpret facts and law, and it explicitly extends to alternative dispute resolution bodies, reaching into commercial arbitration and civil mediation as much as criminal courts. Yet nearly every procedural safeguard discussed above, the right to information in criminal proceedings under Directive 2012/13/EU,[35] the presumption-of-innocence protections and the right to be present at the trial in criminal proceedings under Directive 2016/343/EU,[36] and the criminal limb of the Charter of Fundamental Rights Articles 47 and 48, along with Article 6 ECHR itself,[37] applies only to criminal defendants. Though Article 6 of the ECHR offers procedural fairness requirements for civil cases, a civil litigant left facing an AI-assisted judicial or arbitral decision has no equivalent statutory entitlement to disclosure of the tool's design, training data, or error rate.

This is not a hypothetical concern. France banned the use of judicial analytics for litigation-prediction purposes in 2019 through Article 33 of Law No. 2019-222 (the "Justice Reform Act"), which prohibits reusing the identity data of judges and court clerks for the purpose or effect of evaluating, analysing, comparing, or predicting their professional practices, with violations carrying penalties of up to five years' imprisonment. That provision was a direct response to legal-tech firm building tools to forecast how individual judges tend to rule, largely for civil and commercial litigation strategy rather than criminal defence purposes.[38] The French Constitutional Council upheld the ban in Decision No. 2019-778

34 Regulation (EU) 2024/1689, Article 3 (definitions, including "law enforcement authority" via cross-reference to Directive (EU) 2016/680, Article 3(7)).

35 Directive 2012/13/EU of the European Parliament and of the Council of 22 May 2012 on the right to information in criminal proceedings.

36 Directive (EU) 2016/343 of the European Parliament and of the Council of 9 March 2016 on the strengthening of certain aspects of the presumption of innocence and of the right to be present at the trial in criminal proceedings.

37 Charter of Fundamental Rights of the European Union, Articles 47 (right to an effective remedy and to a fair trial) and 48 (presumption of innocence and right of defence).

38 Malcolm Langford and Mikael Rask Madsen, 'France Criminalises Research on Judges' (Verfassungsblog, 22 June 2019) https://verfassungsblog.de/france-criminalises-research-on-judges/ accessed 8 September 2026.

DC of 21 March 2019, reasoning that such profiling could enable pressure on judges or forum-shopping strategies liable to "alter the functioning of justice," and that the restriction did not infringe the right to a fair and equitable procedure or the balance of the rights of the parties.[39] The episode shows both that civil-context AI transparency concerns are serious enough to have already triggered national legislative and constitutional scrutiny, and that this scrutiny developed entirely outside the AI Act's framework, since the tools at issue, used by litigants and law firms rather than by or on behalf of a judicial authority, likely fall outside point 8(a) altogether and sit in the same unclassified space as the party-side AI systems discussed above.

On the recommendations side, Liberties considered, and for the European context ultimately decided against, proposing a parallel public inventory of AI tools used by law firms, maintained by national bar associations. The concern was primarily one of institutional fit: European bar associations are not positioned in the way U.S. state bar associations sometimes are to maintain this kind of technology registry, and Liberties' network of member organisations spans jurisdictions where the analogy would not transfer cleanly. The parallel public inventory concept may be worth developing separately for a U.S. audience, where bar-association-maintained technology inventories have a more established precedent. Still, it is omitted from the recommendations below.

Institutional Overlap

The responsibilities of judicial authorities and law enforcement are not fully delineated in the AI Act. Article 3 defines a "law enforcement authority" as any public authority competent for the prevention, investigation, detection, or prosecution of criminal offences, but offers no parallel definition of "judicial authorities," despite Annex III point 8(a)'s reliance on that term.[40] The gap is not confined to point 8(a): Article 26(10), governing the authorisation of post-RBI law-enforcement searches, requires deployers to seek authorisation from "a judicial authority or by an administrative authority whose decision is binding and subject to judicial review," using the same undefined term to describe who may approve a search that can determine whether a targeted biometric identification proceeds at all.[41] In Member States where prosecutorial and judicial functions overlap institutionally, this creates a genuine

39 Article 33, Loi n° 2019-222 du 23 mars 2019 de programmation 2018–2022 et de réforme pour la justice; Conseil constitutionnel, Décision n° 2019-778 DC du 21 mars 2019. Checked for currency as of August 2026: no repeal or amendment identified in subsequent French legislation or commentary; the prohibition remains codified in Article L. 10 of the Code de l'organisation judiciaire McCann FitzGerald, "France Bans Analytics of Judges' Decisions" (June 2019), https://www.mccannfitzgerald.com/knowledge/data-privacy-and-cyber-risk/france-bans-analytics-of-judges-decisions

40 See note 34 above.

41 See note 29 above.

classification problem that reaches beyond point 8(a) into this authorisation gateway as well.

In France, prosecutors are *magistrats* drawn from the same corps and the same training institution (the École nationale de la magistrature) as sitting judges. *Magistrats* can move between prosecutorial and judicial posts over a single career. In Italy, though judges and public prosecutors are both drawn from a single, unified magistracy (magistratura), governed by the same Consiglio Superiore della Magistratura, there are only limited possibilities to switch between the two functions under a 2022 legislative reform.[42] A constitutional amendment that would have formally separated the two career tracks, the so-called "Nordio Reform," was put to a confirmatory referendum on 22–23 March 2026 and was rejected by voters, meaning the Constitution continues to treat judges and prosecutors as belonging to a single magistracy.[43] An AI tool used by a French or Italian prosecutor's office to evaluate the reliability of evidence during an investigation can plausibly fall under Annex III point 6(c) as a law-enforcement system, while the same office, exercising its judicial function, falls within point 8(a)'s coverage of AI systems assisting a judicial authority in researching and interpreting facts and law. Such institutional overlaps mean the gaps highlighted above can persist even during criminal investigations, before reaching a judges' chambers.

This raises concerns about the practicability of human oversight, sharpened by the fact that classification determines which obligations attach to a system in the first place: only high-risk systems trigger the AI Act's individual right to explanation under Article 86, which does not extend to the non-high-risk crime-analytics tools discussed above, and which, in any event, is itself subject to a carve-out under Article 86(2) for AI systems whose disclosure is restricted by other Union or national law, a carve-out secondary commentary consistently flags as reaching precisely the law-enforcement and national-security contexts this section

42 École nationale de la magistrature (ENM), official description: "The National School for the Judiciary (ENM) is the only institution in France that trains future and serving French judges and prosecutors", https://www.enm.justice.fr/en/homepage. Consiglio Superiore della Magistratura (CSM), Italy: self-governing body for judges and prosecutors under Articles 104–105 of the Italian Constitution; see EJTN, "Italy – High Council of the Judiciary", https://ejtn.eu/office/italy-high-council-of-the-judiciary/; Verfassungsblog, "Reforming the Italian 'Magistracy': Context, Content, and Risks" (2026).

43 International Association of Judges, "Italian Constitutional Referendum on the Judiciary Rejects the Government's Anti-Judge Reform" (March 2026), https://www.iaj-uim.org/iuw/italian-constitutional-referendum-on-the-judiciary-rejects-the-governments-anti-judge-reform/; Verfassungsblog, "Neither What Italy Needed, Nor What it Deserved: The 2026 Constitutional Referendum and the Rejection of the Nordio Reform" (April 2026), https://verfassungsblog.de/nor-what-it-deserved/.

addresses.[44] Where classification between points 6 and 8 is left to provider self-assessment before deployment rather than to a court or regulator, a defendant has no clear statutory hook to demand disclosure of which track, if any, governed the tool's development.

Structural variability across Member States

Structural variability across EU Member States further complicates regulating AI use in law enforcement. This is because individual Member States with federal or highly decentralised internal structures devolve policing, and sometimes judicial administration, to sub-national units that may adopt AI tools independently of one another.

Each of Germany's sixteen Länder maintains its own police force and legal basis for data-driven investigative tools, and adoption has been uneven and, in at least one case, reversed, providing a strong use case for such complications. Bavaria deployed the predictive-policing tool PRECOBS, focused on burglary-pattern prediction, from 2015–2016 onward, and Baden-Württemberg ran its own PRECOBS pilot from 2015 before discontinuing the tool in 2019 following an external Max Planck Institute evaluation that cited data-quality concerns. Hesse, meanwhile, deployed hessenDATA, a Palantir Gotham-based data-mining and cross-referencing platform used since 2017.[45] The German Federal Constitutional Court's judgment of 16 February 2023 (1 BvR 1547/19, 1 BvR 2634/20) found that the specific provisions authorising this kind of automated data analysis in Hesse (§25a(1), first alternative, of the Hessian Security and Public Order Act) and in Hamburg (§49(1), first alternative, of the Hamburg Police Data Processing Act) were unconstitutional as applied to precautionary crime prevention.[46] The court held that the provisions' thresholds for interference were insufficiently defined and

44 Regulation (EU) 2024/1689, Article 86 (right to explanation of individual decision-making) and Article 86(2) (carve-out for exceptions or restrictions following from Union or national law).

45 On PRECOBS in Bavaria and Baden-Württemberg and its 2019 discontinuation in Baden-Württemberg following an external Max Planck Institute evaluation citing data-quality and effectiveness concerns: Max Planck Institute for Foreign and International Criminal Law, evaluation report on the P4 predictive-policing pilot (2016–17, published via mpg.de); IPS-X case record, "PRECOBS – Predictive policing in German administrations", https://ips-soeu.github.io/ips-explorer/case/10433.html (citing Mayer, 2019, for the 2019 discontinuation). On HessenDATA (Palantir Gotham-based platform, in use since 2017): Lima Blog, "A quest for the legality of predictive data analytics in law enforcement", https://limablog.org/a-quest-for-the-legality-of-predictive-data-analytics-in-law-enforcement-landmark-judgment-on-automated-data-analysis-by-the-federal-constitutional-court-of-germany/.

46 Bundesverfassungsgericht, Judgment of 16 February 2023, 1 BvR 1547/19 and 1 BvR 2634/20, striking down §25a (1) of the Hessian Security and Public Order Act and §49(1) of the Hamburg Police Data Processing Act. Official press release: https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2023/bvg23-018.html.

disproportionate given the scope of profiles the tools could generate from disparate data sources. The Hesse provision was permitted to continue applying, subject to restrictions, only until new legislation was enacted, and no later than 30 September 2023; the amended Hesse law was itself the subject of a fresh constitutional complaint filed by the Gesellschaft für Freiheitsrechte in 2024.[47] Other Länder's equivalent tools were not before the court and left untouched. This disparate approach to AI investigative tools illustrates that a single national ruling does not necessarily harmonise practice across a federal state's own sub-units, let alone across the EU, even within a single tool's history, adoption and discontinuation varying by Land. Spain shows a related pattern, with autonomous communities such as Catalonia (Mossos d'Esquadra) and the Basque Country (Ertzaintza) running their own regional police forces alongside national ones, each capable of adopting investigative AI tools on its own timeline and under its own procurement rules.[48]

Such expansive and unevenly regulated use cases, spanning data mining, RBI, and increasingly the prediction of litigation outcomes, support expanding high-risk classification under Annex III, or at minimum expanding the specific use cases enumerated within it. This aligns with the European Union Agency for Fundamental Rights' repeated calls for a holistic, rights-based approach to AI regulation, one that guarantees a high level of protection against possible wrongdoing related to new technologies across the full spectrum of fundamental rights, rather than the narrower, use-case-by-use-case classification the AI Act currently provides.[49]

47 Gesellschaft für Freiheitsrechte, press release, "GFF erhebt Verfassungsbeschwerde gegen Gesetzesnovelle: Hessen setzt verfassungswidrige Big-Data-Analyse fort" (25 June 2024), confirming GFF filed a fresh constitutional complaint against the amended Hessian Police Act on 25 June 2024, arguing the novelised HessenDATA powers remained unconstitutional after the 2023 ruling: https://freiheitsrechte.org/en/themen/digitale-grundrechte/hessen-big-data-analyse. Independently corroborated by netzpolitik.org, "Hessendata: Erneute Verfassungsbeschwerde gegen polizeiliche Big-Data-Analysen" (25 June 2024).

48 On Catalonia's Mossos d'Esquadra and the Basque Country's Ertzaintza as autonomous regional police forces established under regional statutes: La Moncloa (Spanish Government), "Security Policy", https://www.lamoncloa.gob.es/lang/en/espana/stpv/spaintoday2015/security/paginas/index.aspx.

49 European Union Agency for Fundamental Rights, 'Assessing high-risk artificial intelligence' (2025/2026), https://fra.europa.eu/en/publication/2025/assessing-high-risk-ai, and 'Getting the future right – Artificial intelligence and fundamental rights in the EU'; see also FRA's own summary: "FRA calls on the EU and its Member States to define AI broadly to capture all possible risks, guide developers in assessing how AI affects people's rights and ensure independent oversight and control," https://fra.europa.eu/en/video/2026/fundamental-rights-assessment-high-risk-ai.

Toward a coherent judicial framework

Addressing the gaps traced above requires action on several fronts at once. The most immediate is an EU-wide AI disclosure requirement for criminal evidence produced or processed with AI assistance, covering intended purpose, known error rates, and processing methodology, to secure Article 6 ECHR guarantees on equality of arms and adversarial hearings. That disclosure requirement should be paired with an update to Annex III point 6 explicitly classifying generic crime-analytics and predictive-policing software as high-risk, reversing the categories deleted at the Council's general-approach stage and closing the gap left by the current five-use-case list, the fallback to voluntary Article 95 codes of conduct, and the Digital Omnibus's deferral of the deadlines that would otherwise apply to any reinstated category.

The civil side of the ledger needs equal attention. An equivalent, proportionate disclosure obligation should extend to AI systems used internally by parties to civil and commercial proceedings, including litigation-prediction and legal-tech analytics tools of the kind France's Article 33 was designed to restrain, rather than leaving this space entirely to fragmented, ad hoc national legislation. The institutional-overlap problem, in turn, calls for a clarified legal definition of "judicial authority" in Article 3, with criteria distinguishing an investigative or prosecutorial capacity (Annex III point 6) from a judicial fact-finding or interpretive capacity (Annex III point 8(a)), addressing the classification gap directly in the unified-magistracy Member States discussed above. High-risk status should extend beyond tools used strictly "by or on behalf of" courts, so that private legal-tech software undergoes mandatory transparency, bias auditing, and risk management before its outputs influence judicial strategy or judicial fact-finding. The final Guidelines should confirm that the Article 86 right to explanation applies directly to affected persons for law-enforcement outputs under Annex III point 6 and should clarify how that right interacts with the more limited disclosure available under national criminal procedure.

Finally, given the demonstrated unevenness of AI adoption across sub-national policing and judicial bodies in federal and regionally devolved Member States, Member States should be required to maintain and periodically report a public inventory of AI tools in use by regional and local law-enforcement and judicial bodies, the kind of visibility that no single national court ruling, however significant, can otherwise provide.

Conclusion

The four areas examined in Liberties' policy paper, general intended-use classification, the Article 6(3) filter, election-facing GPAI systems, and the administration of justice, are substantively distinct, but they share a common, ex-post regulatory, structural weakness. In each, the AI Act asks the regulated entity to characterise its own system in the first instance, while only providing ex post, resource-intensive mechanisms- Article 80 reassessment, Article 99 penalties, national

court challenges-to correct a self-serving characterisation after the fact.

The draft Guidelines already contain the raw material for a better approach: the profiling override in Article 6(3), the anti-circumvention rule for complex and agentic systems, and the paragraph 437 test for election-facing GPAI all point toward classifying systems by what they actually do rather than by what a provider calls them. What the draft Guidelines do not yet do is treat that functional approach as the default rule. Instead, each of these tools currently operates as a narrow exception bolted onto a classification process that a provider still controls in the first instance. The final Guidelines should close that gap: make the functional standard the baseline for every Annex III category, not a carve-out available only where the draft Guidelines happen to have already spelled it out.

The recommendations in Sections 1 through 4 do not ask the co-legislators to rebuild the AI Act's risk-based architecture from scratch. They argue, instead, that the architecture already agreed can, and should, be read in the final Guidelines in a way that treats a system's actual capabilities and foreseeable use as at least as important as the label a provider chooses to put on it.

Contact

The Civil Liberties Union for Europe

The Civil Liberties Union for Europe (Liberties) is a non-governmental organisation promoting the civil liberties of everyone in the European Union. We are headquartered in Berlin and have a presence in Brussels. Liberties is built on a network of 22 national civil liberties NGOs from across the EU.

c/o Publix
Hermannstraße 90
12051 Berlin
Germany
info@liberties.eu
www.liberties.eu

Author:
Emma Siegel

Editor:
Eva Simon
eva.simon@liberties.eu