

EXPLAIN YOURSELF:

THE LEGAL REQUIREMENTS GOVERNING EXPLAINABILITY

Authors

Marcus Evans, Rosie Nance, Lisa Fitzgerald, and Lily Hands

Agentic AI brings the promise of AI making a range of decisions autonomously. It has been proposed as the way forward for some of the most impactful decisions in our lives: interacting with customers and actioning requests, triaging requests for medical appointments, and hiring candidates — to name a few.

But many of the models we are looking to build agentic applications on (or to assist decisions in other ways) are black boxes: users can provide the system with data and receive corresponding output but cannot see the logic that leads to the system's output. As a result, organizations may be in the position of saying "computer says no" without being able to pinpoint why. This lack of information can create organizational and ethical challenges as well as legal challenges.

This article examines the legal obligations to explain decisions to affected persons from both data protection and AI-specific legal perspectives across three legal regimes in the UK and EU. It considers the EU's General Data Protection Regulation (EU GDPR), the UK's post-Brexit assimilated version (UK GDPR) (together the GDPR),[1] and the more recent EU AI Act,[2] as well as guidance from regulators in the EU and UK and relevant case law. The article provides practical tips for compliance and incorporating explainability into your wider governance program.

We highlight legal issues relevant to AI governance in one key area: explainability. However, a broader set of legal and governance considerations will apply — particularly for agentic applications. Depending on the context, relevant legal considerations around explainability may include the right to nondiscrimination, the UK Equality Act 2010 and public sector equality duty, and sectoral rules such as those governing financial services.

THE RIGHT TO AN EXPLANATION UNDER THE GDPR & AI ACT

The GDPR protects the fundamental rights and freedoms of individuals by putting in place rules around how their personal data can be used. It generally applies when anyone is doing anything with personal data with an EU or UK connection.[3]

Most obligations fall on the data controller — the party deciding how personal data will be used. Even when an organization uses third-party tech tools or a third-party platform to create agentic AI tools, it will be a data controller. Under the GDPR, it is the data controller who must comply with the obligation to provide an explanation. Not all privacy regimes divide up responsibilities in this way, and the position may be different in other jurisdictions. For example, the Australian Privacy Act 1988 (Cth) does not distinguish between controllers and processors.

> **ORGANIZATIONS MAY BE IN THE POSITION OF SAYING "COMPUTER SAYS NO" WITHOUT BEING ABLE TO PINPOINT WHY**

© 2025 ARTHUR D. LITTLE

The AI Act, in contrast, looks specifically at protecting health, safety, and fundamental rights where AI systems and AI models are used. Many of its provisions take a product-safety approach. This means that most of the obligations fall on the provider — the party developing the AI or having it developed. However, the AI Act right to an explanation is an exception, placing the obligation on the deployer — the party using the technology. This is because only the deployer has the necessary context to understand the role the AI system played in the decision

Under both the GDPR and AI Act, the party using AI or other technologies to make a decision is responsible for providing an explanation, even if they did not develop the technology themselves. To meet this obligation, they will need to engage with the vendor to understand how the technology works.



THE RIGHT TO AN EXPLANATION UNDER THE GDPR

Article 22 of the GDPR establishes a default prohibition on certain types of solely automated decision-making. This applies to decisions made "solely on automated processing," including profiling, that produce legal or similarly significant effects for the individual. Recruitment and employment-related decisions will likely be caught, as will decisions affecting access to finance, healthcare, or education.[4] There are some exceptions to this default prohibition.[5]

Decisions caught by Article 22 GDPR trigger a specific right to information for the individual.[6] The controller must provide "meaningful information about the logic involved" in the decision, as well as the significance and envisaged consequences of the processing for the individual. The data controller must provide information proactively, alongside other information about how the individual's personal data is processed (Articles 13(2)(f), 14(2)(g)), as well as reactively, when an individual exercises their right of access and requests the information (Article 15(1)(h)). Article 22 also builds in a right to human intervention and to contest the decision, requiring the controller to provide sufficient information for the individual to exercise this right.[7] We refer to these proactive and reactive obligations as the "right to an explanation" under the GDPR.

Despite the emphasis on solely automated processing, these obligations cannot be avoided by including a human in the loop. Regulators have emphasized that for processing not to be considered solely automated, human oversight must be meaningful. Processing must be carried out by someone who has the authority and competence to change the decision and who considers all the relevant data.[8]

With this high bar, demonstrating meaningful human involvement can be challenging, in part because the efficiency benefits from decision-assisting technology often rely on reducing the time spent or level of skill and experience needed from any humans involved.

Similarly, identifying what qualifies as a "decision" may not be straightforward. For example, the Court of Justice of the EU (CJEU) found that a credit score may constitute a decision (and therefore be subject to Article 22) where the score was provided to a third party that drew strongly on it to establish, implement, or terminate a contractual relationship with that person.[9]

The right to an explanation is also not the only GDPR consideration relevant to interpretability and explainability. The GDPR also includes broad principles protecting individuals, including transparency, fairness, and accountability. These all impact the way organizations must use individuals' data when making decisions. Articles 13 and 14 of the GDPR impose obligations to inform individuals about how their data is used — obligations that apply in all cases, regardless of whether decisions are automated.

For any processing of personal data, organizations must find an appropriate "lawful basis" under Article 6 GDPR. "Legitimate interests" is often used for AI applications, but it requires weighing the individual's interests against the organization's. To rely on legitimate interests, organizations must be able to explain how their decision-making satisfies that balancing test. Alternatively, consent can prove a lawful basis. However, to be valid for GDPR purposes, individuals must be given sufficient information about the intended use and consequences of the processing to comprehend exactly what they would be consenting to.[10]

Changes to the UK GDPR are coming soon. The Data (Use and Access) Act amends the UK GDPR to lift the default prohibition for data that is not "special category" data.[11] This opens the possibility of greater use of automated decision-making. However, the right to an explanation remains in place. Individuals will have the right to make representations and a (reiterated) right to information, alongside the rights previously included in Article 22 to obtain human intervention and contest the decision.[12] The relevant provisions of the Act are not yet in force. The government has confirmed its plans to bring them into effect around the end of 2025.[13]

WHAT MUST BE PROVIDED WHEN THE RIGHT TO AN EXPLANATION APPLIES?

CJEU recently looked at what constitutes meaningful information in the context of an individual's exercise of their right to information under Article 15(1)(h) EU GDPR. It found that data controllers must explain the procedures and principles actually applied when using an individual's data to reach a decision. This must be done in a concise, transparent, and intelligible way.[14] Interestingly, the court suggested that providing too much information was neither necessary nor helpful, which means disclosing the algorithm does not constitute providing an intelligible explanation to the individual.[15]

Organizations may be concerned that even without a full explanation or disclosure of an algorithm, the obligation to explain decisions may expose their trade secrets. There may also be scenarios where an explanation could reveal third-party personal data. The CJEU has suggested that information could be disclosed to a court or competent authority to conduct a balancing exercise and decide what must be disclosed.[16] Many questions remain as to how this would play out.

CHANGES TO THE UK GDPR ARE COMING SOON

THE RIGHT TO AN EXPLANATION UNDER THE AI ACT

The AI Act's right to an explanation is very similar to the right under Article 15(1)(h) GDPR: deployers must provide affected persons with clear, meaningful explanations of the AI system's role in the decision-making process, as well as the main elements of the decision itself. This right applies to decisions that have legal effects, as well as those that significantly affect the relevant person in ways that they consider may negatively impact their health, safety, or fundamental rights. While "affected persons" will often be individuals, they can also be organizations.[17]

The right applies to decisions made by most types of AI systems classified as high-risk due to their potential impact on fundamental rights — including systems used in employment, recruitment, and consumer credit.[18] The decision only needs to be made "on the basis of" the AI system's output; it does not need to be solely automated.

© 2025 ARTHUR D. LITTLE

The AI Act includes a derogation that may exclude certain systems from the high-risk classification — meaning the right to an explanation would not apply — where they do not pose a significant risk of harm to health, safety, or fundamental rights, including by not materially influencing the outcome of decision-making.[19] However, this is not a general "human in the loop" exemption and is likely to be interpreted narrowly. The Act explicitly states that the derogation does not apply where individuals are subject to profiling.



Where an AI system is developed by a third-party provider, the deployer should be able to rely on information from the provider. Providers of high-risk AI systems are required to design and develop their systems so that deployers can interpret the output and to supply information relevant to explain its output.[20] However, what precisely this information must include remains unclear at this time.[21]

In addition to the AI Act's right to an explanation, deployers must assign competent human oversight, which requires ensuring that those providing this oversight have sufficient information and understanding to do their job.[22]

Significant risk management, governance, and documentation obligations fall on the provider.[23] Organizations may also become subject to these obligations — beyond those that apply to deployers — if they develop their own tools, including agentic applications, or significantly modify a third-party platform or model.[24]

In practice, deployers may encounter providers who claim that the above-mentioned derogation applies to their systems — and therefore that they are not subject to the AI Act's information requirements. If deployers disagree with this assessment, they will need to evaluate whether they have sufficient information to meet their own obligations, including the right to an explanation and the requirement to provide competent human oversight.

WHY THE RIGHT TO AN EXPLANATION MATTERS

Fines under the GDPR can be substantial: up to 4% of the total worldwide annual turnover of the preceding financial year or €20 million (£17.5 million under the UK GDPR),[25] with the current record fine coming in at €1.2 billion. Under the provisions of the AI Act discussed above, fines are up to €15 million or 3% of global turnover (whichever is greater).[26]

The GDPR has a one-stop-shop provision, which means that for cross-border processing, organizations would not generally be fined by multiple regulators for the same infringement. The AI Act does not include this mechanism, so fines could be issued in multiple member states. Organizations can also be fined for the same behavior under different legislation. While the AI Act requires regulators to take other other fines into account, it does not prevent parallel enforcement.[27]

Regulators' powers are not only financial; they can go to the heart of an organization's business model. Both the GDPR and the AI Act include powers allowing regulators to require organizations to stop a particular practice (in the case of the GDPR)[28] or withdraw an AI system (in the case of the AI Act).[29]

In terms of litigation risks, individuals are already active in pursuing compensation claims under the GDPR.[30] In addition, not-for-profit organizations can bring group claims — seeking compensation or injunctive relief — in the EU.[31]

The AI Act does not include a right to compensation for individuals (although qualified not-for-profits can still bring injunctions for alleged noncompliance).[32] Mass claims in the AI space will, however, be possible under the revised Product Liability Directive,[33] which extends the definition of "product" to include software and AI systems. Of course, AI (mis)use can prompt claims under other existing legislation, like the Equality Act 2010 and similar legislation in the EU, or claims in tort, such as for negligence.

A HIGH DEGREE OF INTERPRETABILITY IS ESSENTIAL FOR DECISIONS THAT SIGNIFICANTLY AFFECT INDIVIDUALS

DOES THE RIGHT TO AN EXPLANATION UNDER THE GDPR OR AI ACT PREVENT THE USE OF BLACK-BOX ALGORITHMS?

As with most questions in the data law world, this answer is context-dependent. Providing meaningful information does not necessarily require selecting a fully interpretable model. However, organizations will not be able to comply with their GDPR or AI Act obligations if they are unable to explain how a decision was reached or if they rely solely on vendor claims of high performance without interrogating how the technology functions.

A high degree of interpretability is essential for decisions that significantly affect individuals — particularly where denial of a service or opportunity could have serious consequences. Healthcare is a clear example: misalignment between human intent and the rules learned by an AI system can have fatal consequences. For example, while researching the potential application of AI systems for hospital triage, researchers found that an AI system was classifying asthma patients as low risk for pneumonia and recommending outpatient treatment. In reality, asthma patients had lower mortality rates because they were typically admitted directly to intensive care — a factor the model failed to recognize. Crucially, this issue was uncovered because the AI system was a fully interpretable, rules-based model.[34] If an opaque system that made similarly misguided inferences were rolled out, it would be impossible to comply with applicable GDPR or AI Act obligations.

The UK Information Commissioner's Office (ICO) has published joint guidance with the Alan Turing Institute to provide practical insights into explanations and explainability techniques in the AI governance process (the Guidance).[35] The Guidance emphasizes that these considerations are not only relevant for decision-making under Article 22 UK GDPR. As discussed above, even where Article 22 UK GDPR does not apply, the GDPR principles (e.g., transparency and accountability) continue to apply to decisions where there is meaningful human involvement.[36]

The Guidance suggests drawing on a mixture of process-based explanations (which describe the process and demonstrate good governance) and outcome-based explanations (which clarify the results of a specific decision). It guides controllers through the process of providing meaningful information based on the domain (sector or setting), use case, impact on the individual, data used to train and test the model, urgency (i.e., importance of receiving or acting on the outcome of a decision within a short time frame), and audience.[37]

© 2025 ARTHUR D. LITTLE

Controllers will likely need to draw on a range of explanations, including rationale explanations (which the Guidance considers to be the "why" of an AI decision) and responsibility explanations (the "who" involved in the development and management of the AI model).

In most cases, the Guidance indicates that the primary focus should be on providing rationale- and responsibility-based explanations — understanding what the system is doing and who is responsible for its outputs.[38] However, the Guidance acknowledges that the standard types of explanation may not suit every organization. Some may find that developing their own explanation framework is more effective. The Guidance confirms that this approach is "absolutely fine" provided the organization upholds the principles of transparency and accountability and carefully considers the specific context and potential impact.[39]



The Guidance cautions that organizations should only use black-box models if they have thoroughly considered their potential impacts and risks in advance.[40] Team members should also ensure that the use case — and the organization's capacity and resources — support the responsible design and deployment of the system. The Guidance further recommends using supplementary interpretability tools to deliver a domain-appropriate level of explainability. This level should be reasonably sufficient to mitigate potential risks and provide decision recipients with meaningful information about the rationale behind any given outcome.

The Guidance highlights that controllers need to think both locally (aiming to interpret individual predictions or classifications) and globally (capturing the logic of the model's behavior as a whole across predictions or classifications) when choosing supplementary explanation tools.

Because it was published in 2020, the Guidance did not envisage the way organizations are currently exploring agentic AI. When using third-party models to build agents, the Guidance may still serve as a useful framework for requesting information from vendors. However, in practice, major vendors may not provide additional detail — leaving organizations to make risk-based decisions based on the information available. It is also worth noting that EU regulators may not adopt the same approach as the ICO, and the ICO itself may issue updated guidance in the coming months.

BUILDING EXPLAINABILITY INTO AI GOVERNANCE PROGRAMS

To build explainability into your AI governance program, consider the following steps:

1. **Think about explanations early and often.** Consider explanations at the outset and throughout the system's lifecycle. Ensure you can provide relevant information as applicable about data-collection choices, data cleaning and labeling, the algorithm selected and used, validation and testing, and any decisions about how a system will or should be deployed.
2. **Gather the relevant stakeholders** (legal, compliance, procurement, and the proposed project team) ahead of rollout and ensure that appropriate consultation and communication takes place throughout the lifecycle to manage risks.
3. **If you do not have an AI governance program** that triages agentic use cases (or any use cases making decisions about individuals) for enhanced review by legal and compliance teams, **put one in place.**
4. When using an external platform or third-party provider, **factor in the right to an explanation and other applicable legal requirements** — such as those under the GDPR and AI Act — during procurement or tool selection. Request additional information from vendors as needed, and assess whether what they provide is sufficient to meet your legal obligations.

5. **Include sufficient data-sharing obligations and confidentiality provisions in your contracts** with vendors, developers, and/or deployers to ensure explainability can be achieved on relevant terms and on an ongoing basis.
6. **Maintain comprehensive documentation and logging** throughout the system's lifecycle to ensure you can provide meaningful information when required.
7. If you are using technology to assist rather than make decisions, **assess how meaningful any human involvement truly is** — and identify the specific legal obligations that apply.
8. **Think about the needs of the individuals** impacted by the AI system. Audiences have different needs, and different domains require different approaches. Translate the rationale of your system's results into usable, easily understandable reasons for decisions.
9. **Implement an AI literacy program** to ensure that everyone involved in developing, deploying, or governing AI has the necessary technical understanding — and is aware of the organizational, legal, and societal risks associated with inadequate explainability.

REFERENCES

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data [2016] OJ L 119/1.

[2] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) OJ 2024 L 1/144.

[3] GDPR, art 2 and 3.

[4] Article 29 Data Protection Working Party, "Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679" (WP 251 rev01, version 2 2018), p. 22; see also recital 71 GDPR.

[5] For "regular" personal data, necessity under a contract between data subject and a data controller, authorization under EEA/member state law (EU) or domestic law (UK), or explicit consent (art 22(2) GDPR)). For "special category" data (e.g., data about health), explicit consent or substantial public interest on the basis of EU/UK law is required, alongside suitable safeguards.

[6] The right to explanation is defined as applying to "at least" the automated decision-making/profiling referred to in Article 22 GDPR. The wording does not exclude the application of the right to an explanation to other forms of processing. However, at this stage, we are not aware of any applications of Articles 13(2)(f), 14(2)(g), or 15(1)(h), except in the context of Article 22 GDPR.

[7] GDPR, art 22(3).

[8] Article 29 Data Protection Working Party, "Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679" (WP 251 rev01, version 2 2018), p. 20.

[9] OQ v SCHUFA Holding AG (Court of Justice of the European Union, C-634/21 ECLI:EU:C:2023:957), 7 December 2023.

[10] Article 29 Data Protection Working Party, "Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679" (WP 251 rev01, version 2 2018), p. 15.

[11] Described in GDPR, art 9(1), including data about health.

[12] Data (Use and Access) Act, s.80.

[13] "Data Use and Access Act 2025. Plans for Commencement." Gov.UK, 25 July 2025.

[14] CK v Dun & Bradstreet Austria GmbH and Magistrat der Stadt Wien (Court of Justice of the European Union, C-203/22, ECLI:EU:C:2025:117, 27 February 2025), para. 66.

© 2025 ARTHUR D. LITTLE

[15] Dun & Bradstreet Austria GmbH, para. 59. This is consistent with the EDPB's view, which emphasizes that an explanation should enable individual to understand the rationale behind and reasons for a decision; see Article 29 Data Protection Working Party, "Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679" (WP 251 rev01, version 2 2018), p. 25.

[16] CK v Dun & Bradstreet Austria GmbH and Magistrat der Stadt Wien (Court of Justice of the European Union, C-203/22, ECLI:EU:C:2025:117, 27 February 2025).

[17] AI Act, art 86.

[18] Specifically, it apples to the AI systems listed in Annex III other than critical infrastructure, see Article 86(1).

[19] AI Act, art 6(3).

[20] AI Act, art 13.

[21] CEN/CENELEC's JTC21, the committee responsible for producing AI Act standards, is currently developing a transparency taxonomy that may shed more light; see FprEN ISO/IEC 12792, marked as "Under Approval" on the Cen/CENELEC work program as of 18 June 2025.

[22] AI Act, art 26.

[23] AI Act, Chapter III and art 72.

[24] AI Act, art 25.

[25] GDPR art 83(5).

[26] AI Act, art 99(4).

[27] AI Act, At 99(7)(b) and (c).

[28] GDPR, art 58(d) and (f).

[29] AI Act, Art 79(2).

[30] A right provided under GDPR, art 82.

[31] Under Directive (EU) 2020/1828 of the European Parliament and of the Council on representative actions for the protection of the collective interests of consumers.

[32] AI Act, art 110.

[33] Directive (EU) 2024/2853 of the European Parliament and of the Council on liability for defective products.

[34] Christian, Brian. *The Alignment Problem: How Can Machines Learn Human Values?* W.W. Norton, 2020, p. 106 onward.

[35] "Explaining Decisions Made with AI." ICO/Alan Turing Institute, accessed 2025.

[36] ICO Guidance (see 35), p. 13.

[37] ICO Guidance (see 35), pp. 35-41.

[38] ICO Guidance (see 35), pp. 20-35.

[39] ICO Guidance (see 35), p. 51.

[40] ICO Guidance (see 35), pp. 74-83.

About the authors

Marcus Evans is a Partner at Norton Rose Fulbright, London, and Head of Information Governance, Privacy, and Cybersecurity, EMEA. He advises global clients across sectors on a range of complex AI and data projects. Mr. Evans can be reached at marcus.evans@nortonrosefulbright.com.

Rosie Nance is a Senior Knowledge Lawyer for the AI and data team at Norton Rose Fulbright, London. She focuses on analyzing legal developments and crafting strategic approaches to ensure compliance with emerging regulatory requirements. Ms. Nance can be reached at rosie.nance@nortonrosefulbright.com.

Lisa Fitzgerald is a Partner in the Corporate and Technology practice at Norton Rose Fulbright, London. She specializes in privacy, data protection, cyber/AI, and advising private and public sector clients across a range of regulatory and transactional matters. Ms. Fitzgerald can be reached at lisa.fitzgerald@nortonrosefulbright.com.

Lily Hands is an Associate with the Technology, Commonwealth, and Government teams at Norton Rose Fulbright, Canberra, advising on a range of AI-related matters. She earned a PhD in the regulation of AI-driven automated decision-making from the University of Cambridge, UK. Dr. Hands can be reached at lily.hands@nortonrosefulbright.com.

© 2025 ARTHUR D. LITTLE

AMPLIFY

Anticipate, Innovate, Transform

Cutter is Arthur D. Little's Open Consulting community, bringing expert academics and business leaders together to advance thinking in key areas of business and technology.

Arthur D. Little has been pushing the boundaries of innovation since 1886, linking people, technology and strategy to help our clients overcome today's most pressing challenges, while seizing tomorrow's most promising opportunities.

Our people are present in the most important business centers around the world, combining strong practical industry experience with excellent knowledge of key trends, technologies and market dynamics. We are proud to work alongside most of the Fortune 1000 companies and other leading firms and public sector organizations, supporting them to accelerate performance, innovate through convergence and digital and make a positive impact on the world.

It's what we believe makes *The Difference*.

Founding Editor: Ed Yourdon
Managing Editor: Christine Generali
Publishing Manager: Linda Mallon Dias
Copyeditor: Tara K. Meads

© 2025 Arthur D. Little. All rights reserved. For further information, please visit www.adlittle.com.

CUTTER

AN ARTHUR D. LITTLE COMMUNITY

For more content, visit www.cutter.com