Journal of AI-Powered Medical Innovations
ISSN: 3078-1930 DOI: 10.60087
2025, Vol. 04, No. 1
Home page https://japmi.org/

Explanations That Work: A Risk-Based Model for Meaningful AI Transparency in Regulated Decisions

Nicki James Shepherd

UK Research, London, United Kingdom
ORCID: 0009-0009-3480-8706

Abstract

Law increasingly gives people a right to an explanation when artificial intelligence (AI) contributes to decisions about them. The EU AI Act grants a right to a clear and meaningful explanation for decisions informed by high-risk systems, the Court of Justice of the European Union has held that handing over an algorithm or formula does not satisfy the right to meaningful information about automated decisions, and the UK has reformed its rules on automated decision-making while keeping rights to information and to contest. Yet the empirical literature on explainable AI shows that explanations frequently fail the people they are meant to serve. They can increase acceptance of AI advice whether or not it is correct, leave people less able to spot model errors, and be produced to satisfy compliance rather than to inform. This paper asks what makes an explanation work, and how the effort put into explanation should scale with the risk of the decision. It synthesises legal, regulatory and human-computer interaction research to propose the Meaningful Explanation Model (MEM). The model specifies explanations along three dimensions: the risk tier of the decision, the recipient and the purpose the explanation must serve, and a set of effectiveness tests that judge an explanation by what it enables rather than by whether it exists. The model is applied to five regulated decisions and its implications for regulators, deployers and designers are set out.

Keywords: explainable AI; right to explanation; automated decision-making; AI regulation; human oversight; contestability

1. Introduction

When a person is refused credit, flagged for a benefits investigation or ranked out of a shortlist with the help of an AI system, the law in Europe and the UK increasingly says that person is owed an explanation. Article 86 of the EU Artificial Intelligence Act gives anyone affected by

* Corresponding author: **Nicki James Shepherd Email:** nick.shepherd@uk-research.org

Received: 01-10-2025; **Accepted:** 25-11-2025; **Published:** 16-12-2025

Copyright: © The Author(s), 2025. Published by JAPMI. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

a decision based on the output of a high-risk AI system the right to obtain a clear and meaningful explanation of the role the system played and the main elements of the decision (Regulation (EU) 2024/1689). The Court of Justice of the European Union (CJEU) has held that a credit score on which a lender relies heavily can itself be an automated decision under the General Data Protection Regulation (GDPR) (*OQ v Land Hessen*, 2023), and that the right to "meaningful information about the logic involved" is not met by handing over complex formulas or technical documentation (*CK v Magistrat der Stadt Wien*, 2025). In the UK, the Data (Use and Access) Act 2025 has loosened restrictions on automated decision-making for most personal data, while keeping safeguards that require people to be told about significant automated decisions and to be able to contest them.

Much of the debate about explanation has concerned what the law requires (Wachter et al., 2017; Selbst & Powles, 2017; Kaminski, 2019). This paper asks a different question: when an explanation is given, does it work? A right to explanation is only as valuable as the explanations it produces. If explanations mislead, overwhelm or reassure without informing, the right becomes a ritual.

There is good reason to worry that this is happening. A substantial body of experimental research shows that explanations of AI recommendations can increase people's acceptance of those recommendations regardless of whether they are correct (Bansal et al., 2021), that more transparent models can leave people worse at spotting large model errors (Poursabzi-Sangdeh et al., 2021), and that people tend to form general rules about when to trust an AI system rather than engage with each explanation (Buçinca et al., 2021). Practitioners, meanwhile, mostly use explainability tools to debug their own models rather than to inform the people affected by them (Bhatt et al., 2020), and even data scientists misread the outputs of the interpretability tools they use (Kaur et al., 2020).

The gap, then, is between a legal duty framed in terms of meaningfulness and a technical practice that rarely tests whether explanations mean anything to their recipients. This paper proposes a model to close it. Its main research question is: how should the content, form and testing of AI explanations in regulated decisions be specified so that explanations achieve their purpose, in proportion to the risk of the decision? Three subsidiary questions guide the analysis. What do law and regulatory guidance now require an explanation to do? What does empirical research show about when explanations succeed and fail? And how can those findings be turned into a specification that regulators can enforce and deployers can meet?

The contribution is the Meaningful Explanation Model (MEM), a three-dimensional specification that links decision risk, recipient and purpose, and effectiveness testing. The paper is structured as follows. Section 2 reviews the legal and regulatory position and the empirical research on explanations. Section 3 describes the methodology. Section 4 presents the model and applies it to five regulated decisions. Section 5 discusses implications and objections, and Section 6 concludes.

2. Literature Review

2.1 The legal duty to explain

The debate over a legal right to explanation began with the GDPR. Wachter, Mittelstadt and Floridi (2017) argued that, despite widespread claims, the GDPR did not create a general right to an explanation of individual automated decisions, only a right to limited information about the existence and general functionality of such systems. Selbst and Powles (2017) responded that the requirement to provide "meaningful information about the logic involved" in Articles 13 to 15 should be read functionally: information is meaningful if it enables the person to exercise their rights, including the right to contest. Edwards and Veale (2017) took a more sceptical position, arguing that even if such a right existed it was unlikely to be the remedy people needed, because individual explanations do little to address systemic problems and place the burden of challenge on those least able to bear it. Kaminski (2019) later argued that the right should be understood within the GDPR's wider system of accountability, in which individual transparency sits alongside oversight and impact assessment.

Case law has since moved toward the functional reading. In *OQ v Land Hessen* (2023), the CJEU held that the automated calculation of a credit score by a credit reference agency counts as an automated decision under Article 22 of the GDPR where a lender relies heavily on that score, rejecting the argument that the score was merely preparatory. In *CK v Magistrat der Stadt Wien* (2025), the Court addressed what "meaningful information about the logic involved" requires. It held that the person must be given relevant information about the procedures and principles actually applied to their data to reach the result, in a clear, understandable and easily accessible form, so that they can check its lawfulness and exercise their rights. Simply communicating a complex algorithm or a detailed technical description of every step would not meet that standard. Trade secrets cannot be used as a blanket reason to refuse access; where they are at stake, a court or supervisory authority must balance the competing rights, if necessary after reviewing the protected material itself (Stibbe, 2025). Commentators have noted the Court's suggestion that, for credit scoring, explaining how changes in a person's data would have changed the result might be sufficient, an approach close to the counterfactual explanations proposed by Wachter, Mittelstadt and Russell (2018).

The EU AI Act adds a separate right. Article 86 applies where a deployer takes a decision based on the output of a high-risk system listed in Annex III, the decision has legal or similarly significant effects, and the affected person considers it adversely affects their health, safety or fundamental rights. That person may request a clear and meaningful explanation of the role of the AI system in the decision and of the main elements of the decision taken (Regulation (EU) 2024/1689). The right applies only where other EU law does not already provide it, which means it will often operate alongside the GDPR rather than replacing it.

The UK has taken a different path. The Data (Use and Access) Act 2025, which received Royal Assent on 19 June 2025, replaces the general restriction on solely automated decisions with a more permissive, risk-based approach for most personal data, while keeping tighter rules for special category data. For significant decisions based solely on automated

processing, individuals must be given information about the decision and be able to make representations, obtain human intervention and contest it (Burges Salmon, 2025). The Information Commissioner's Office and the Alan Turing Institute had earlier published guidance identifying six types of explanation that organisations might provide: rationale, responsibility, data, fairness, safety and performance, and impact (Information Commissioner's Office [ICO] & The Alan Turing Institute, 2020).

Taken together, these sources point in a consistent direction. The law is less concerned with whether a system is technically interpretable than with whether the affected person receives information they can understand and act on. That is a standard of effectiveness, not of disclosure. But the law does not say how effectiveness should be judged.

2.2 What explanations do to people

Research in human-computer interaction has tested what explanations actually do, and the results are sobering. Poursabzi-Sangdeh et al. (2021) ran a series of large experiments in which participants predicted apartment prices with the help of a model presented either as a transparent model with few features or as a black box. Participants shown the transparent model could better simulate what the model would predict, but they did not follow its accurate predictions more closely, and they were less able to detect and correct its large errors, apparently because the extra information overloaded them. The authors concluded that intuitions about interpretability are often wrong and that interpretability should be tested with people rather than assumed.

Bansal et al. (2021) studied human-AI teams on tasks where the AI was about as accurate as the human. Teams outperformed either humans or the AI alone, but adding explanations did not improve that performance. Instead, explanations increased the chance that people would accept the AI's recommendation, whether the recommendation was right or wrong. Wang and Yin (2021) compared several explanation types across tasks and found that their usefulness depended on the task and the decision-maker's expertise, with some explanations failing to help people judge when to rely on the model. Buçinca et al. (2021) found that simple explanations did not reduce over-reliance on wrong AI advice, whereas "cognitive forcing" designs, which ask people to form their own judgement before seeing the AI's recommendation, did reduce it, at the cost of being less liked by participants. These findings fit with a longer literature on automation bias, in which people defer to automated advice even when it conflicts with other information (Goddard et al., 2012).

The pattern across these studies is that explanations change people's trust more readily than they change people's understanding. Jacovi et al. (2021) distinguish warranted trust, which tracks the system's actual trustworthiness, from unwarranted trust, which does not. An explanation that increases trust in a system regardless of whether the system is right in the case at hand produces unwarranted trust. In regulated decisions, where an operator is supposed to exercise meaningful oversight of AI outputs, that is precisely the failure the law is trying to prevent.

2.3 Who explanations are for

A second strand of research asks whom explanations serve. Miller (2019), drawing on philosophy and psychology, argued that human explanations are contrastive (they answer "why this rather than that?"), selective (they give a few reasons, not all causes) and social (they are part of a conversation shaped by what the questioner already knows). Most technical explanation methods meet none of these conditions. Lipton (2018) showed that "interpretability" is used to mean several different things, including simulatability, decomposability and post-hoc justification, and warned that the word often hides more than it reveals.

Studies of practice confirm the mismatch. Bhatt et al. (2020) interviewed organisations deploying explainability methods and found that they were used mainly by machine learning engineers for debugging, rarely to inform end users or affected people. Kaur et al. (2020) found that data scientists over-trusted and misused interpretability tools, often taking their outputs at face value. Liao et al. (2020) interviewed designers of AI products and organised user needs around questions people actually ask, such as "why", "why not", "what if" and "how to be that", showing that different questions call for different forms of explanation. Ehsan et al. (2021) argued for "social transparency", in which explanations include information about the people and organisational processes around an AI system, such as who else has relied on it and with what result, because recipients judge AI outputs in a social context rather than a technical one.

Rudin (2019) offered a more radical response: for high-stakes decisions, organisations should stop explaining black-box models after the event and use models that are interpretable by design. Her argument is that post-hoc explanations are approximations that can be wrong in exactly the cases that matter, and that in many high-stakes domains interpretable models perform as well as opaque ones.

2.4 Summary of the gap

The literature supports three conclusions. Law in the EU and UK now frames explanation as a functional duty: an explanation must enable the affected person to understand and challenge a decision, and technical disclosure is not enough. Empirical research shows that explanations often fail that functional test, increasing reliance on AI without improving understanding, and sometimes making errors harder to detect. And the recipients and purposes of explanation are varied, yet most explanation practice is designed for engineers. What is missing is a model that tells deployers what kind of explanation a given decision requires, for whom and for what purpose, and how to check that it works. The legal analysis supplies one input to such a model; the empirical literature supplies another.

3. Methodology and Design

3.1 Approach

This is a conceptual and analytical paper. It combines doctrinal analysis of legal sources with a structured narrative synthesis of empirical and design research (Popay et al., 2006), and uses the procedure for conceptual framework development set out by Jabareen (2009) to build the model. The model is then applied to five regulated decision scenarios as an analytical test of its coherence and usefulness.

3.2 Sources

Three bodies of material were analysed. Legal and regulatory sources comprised the EU AI Act, the GDPR provisions on automated decisions, two CJEU judgments on automated decision-making, the UK Data (Use and Access) Act 2025 and the ICO and Alan Turing Institute guidance. Scholarly legal analysis was drawn from the leading contributions to the right-to-explanation debate. Empirical and design research was drawn from peer-reviewed human-computer interaction and machine learning venues, prioritising controlled experiments with human participants and studies of deployed practice. All sources were checked against publisher or official records.

3.3 Procedure

Legal sources were analysed for what they require an explanation to achieve, as opposed to what they require it to contain. Empirical studies were tabulated by the explanation type tested, the recipient, the task, the outcome measured and the direction of effect. Relationships across studies were then examined, with particular attention to the gap between effects on trust and effects on understanding or error detection. Following Jabareen (2009), concepts were grouped into clusters, which became the three dimensions of the model, and the relationships between dimensions were defined. The model was then applied to five scenarios chosen to cover finance, public administration, employment, education and health, and revised where a scenario exposed a gap.

3.4 Limitations of the design

The model has not been tested empirically with recipients of real decisions; Section 6 proposes how this could be done. The legal analysis is limited to EU and UK law, and the UK provisions on automated decision-making were still being brought into force at the time of writing. The empirical studies synthesised were mostly laboratory or online experiments, which may not capture the stakes and emotions involved when a real decision goes against someone.

4. Results: The Meaningful Explanation Model

4.1 From disclosure to effectiveness

The synthesis produced a central finding. Legal duties to explain are framed functionally, but compliance practice and much technical work treat explanation as disclosure: something is

meaningful if it has been provided. The empirical evidence shows that this assumption fails. An explanation is meaningful only if it achieves its purpose for its recipient, and whether it does so is an empirical question. Table 1 contrasts the two approaches.

Table 1. Disclosure and effectiveness approaches to explanation compared

Dimension	Disclosure approach	Effectiveness approach
Question asked	Has an explanation been provided?	Does the explanation achieve its purpose for its recipient?
Typical output	Feature importances, model documentation, generic notices	Answers to the recipient's actual question in a form they can use
Assumed recipient	A technically literate reader	A defined recipient: affected person, operator, auditor or regulator
Evidence of compliance	The explanation exists	The explanation has been tested and meets defined criteria
Main risk	Explanations that reassure without informing	Cost of testing; risk of gaming the tests
Scales with risk?	Rarely	Yes, by design

4.2 The three dimensions

The Meaningful Explanation Model (MEM) specifies an explanation along three dimensions. A specification is complete when all three have been set for a given decision.

Dimension 1: Decision risk tier

The effort an organisation puts into explanation should be proportionate to the consequences of the decision. MEM uses three tiers, aligned where possible with existing legal categories.

Tier A (high consequence) covers decisions with legal or similarly significant effects on a person, including those informed by high-risk systems under Annex III of the EU AI Act and significant decisions under UK data protection law. Examples include decisions about credit, social security, employment, access to education and medical treatment.

Tier B (moderate consequence) covers decisions that affect people but are routinely reviewable and reversible, such as triaging a customer complaint or prioritising an internal case queue.

Tier C (low consequence) covers decisions with minor, easily reversed effects, such as content recommendations or routing a routine query.

The tier sets the minimum standard on the other two dimensions. At Tier A, every recipient with a role in the decision must receive an explanation that has been tested against all relevant criteria; at Tier C, a general notice of AI use may suffice.

Dimension 2: Recipient and purpose

Explanations fail when they are written for the wrong person or the wrong job. MEM identifies four recipients, each with a primary purpose that the explanation must serve. Table 2 sets

them out, drawing on the question types identified by Liao et al. (2020) and the explanation types in the ICO and Alan Turing Institute (2020) guidance.

Table 2. Recipients, purposes and suitable explanation forms

Recipient	Primary purpose	Typical questions	Suitable forms
Affected person	To understand and, if necessary, contest the decision	Why this outcome? What would have changed it? Who decided, and how do I challenge it?	Contrastive and counterfactual reasons in plain language; responsibility and route-to-challenge information; data used about them
Operator (human in the loop)	To exercise meaningful oversight of the AI output in the case at hand	Should I rely on this recommendation here? Where is it likely to be wrong?	Case-specific uncertainty; indications of when the model is outside its reliable range; designs that prompt independent judgement before the AI output is shown
Auditor or internal assurance	To verify that the system behaves as intended across cases	Is it accurate, fair and stable? How does it behave for different groups?	Global model descriptions; performance and fairness metrics by group; documentation of data and changes
Regulator or court	To determine whether the decision and system were lawful	Were the rules applied? Were rights respected?	The above, plus access to protected material under confidentiality where necessary

Two points follow from this dimension. First, one explanation rarely serves all four recipients. A feature-importance chart may help an auditor and mean nothing to a person refused a loan. Second, the operator's explanation has a special status. Because the law relies on human oversight as a safeguard, an explanation that increases the operator's reliance on the AI without improving their ability to spot errors actively weakens that safeguard (Bansal et al., 2021; Buçinca et al., 2021). For operators, the explanation's job is to support judgement, not to justify the output.

Dimension 3: Effectiveness tests

The third dimension sets out how an organisation should test whether an explanation works. MEM proposes five criteria, each tied to evidence about how explanations fail.

Comprehension: Can a representative recipient correctly state, in their own words, the main reasons for the decision? This tests the recipient's understanding, not the explanation's content, and responds to evidence that people can misread even technical explanations designed for experts (Kaur et al., 2020).

Actionability: Can the recipient identify what, if anything, they could do differently or how to challenge the decision? This reflects the functional reading of the right to explanation (Selbst & Powles, 2017) and the CJEU's emphasis on enabling people to exercise their rights (*CK v Magistrat der Stadt Wien*, 2025).

Calibrated reliance: For operators, does the explanation improve their ability to accept correct recommendations and reject incorrect ones, rather than simply increasing acceptance? This responds directly to the finding that explanations can raise acceptance regardless of correctness (Bansal et al., 2021).

Fidelity: Does the explanation accurately reflect how the system actually reached its output, rather than offering a plausible story? This responds to Rudin's (2019) warning that post-hoc explanations can diverge from the model they describe, and to the CJEU's requirement that information concern the procedures actually applied.

Accessibility: Can the explanation be understood by recipients with different levels of literacy, language and disability? An explanation that works only for fluent readers of dense text excludes many of the people most affected by automated decisions.

Table 3 shows how the tier sets the testing standard for each criterion.

Table 3. Minimum testing standards by risk tier

Criterion	Tier A (high)	Tier B (moderate)	Tier C (low)
Comprehension	Tested with representative recipients before deployment and after significant change	Reviewed against plain-language standards	General notice in plain language
Actionability	Every explanation states what would change the outcome and how to challenge it	Route to challenge stated	Not required
Calibrated reliance	Operator studies measuring correct and incorrect acceptance	Monitoring of override rates	Not required
Fidelity	Interpretable-by-design models preferred; post-hoc methods validated against model behaviour	Post-hoc methods documented with known limits	Not required
Accessibility	Tested with disabled recipients and in alternative formats	Alternative formats available on request	Accessible general notice

4.3 How the dimensions interact

The three dimensions are applied in sequence. The organisation first assigns the decision to a risk tier. It then identifies every recipient with a role in the decision and the purpose each explanation must serve. Finally, it applies the testing standard set by the tier to each recipient's explanation. The output is an explanation specification: a short document listing, for each recipient, the questions the explanation must answer, the form it will take and the evidence that it works.

This sequence has a practical advantage. It concentrates effort where the law and the risk demand it. Most decisions in most organisations fall into Tiers B and C, where the burden is

modest. Tier A decisions, where explanation failures have the greatest consequences, receive the full set of tests.

4.4 Application to five regulated decisions

Table 4 applies the model to five decisions across regulated sectors, showing how the specification differs from current common practice.

Table 4. MEM applied to five regulated decisions

Decision	Tier	Common practice	MEM specification (selected elements)
Consumer credit refusal informed by a credit score	A	Generic adverse-action reasons; score bands	Affected person: counterfactual reasons ("the outcome would have changed if...") tested for comprehension; route to challenge. Operator: score uncertainty and out-of-range flags. Regulator: access to scoring logic under confidentiality, consistent with *OQ v Land Hessen* (2023) and *CK v Magistrat der Stadt Wien* (2025)
Referral of a social security claim for fraud investigation	A	Internal risk flag with no explanation to claimant	Affected person: notice of AI involvement, main factors and right to contest. Operator: cognitive forcing design requiring an independent assessment before the flag is shown. Auditor: referral rates and outcomes by group
Ranking of job applicants for a shortlist	A	Ranked list given to recruiter	Operator: reasons for each ranking with indications of low confidence; periodic test of whether recruiters can identify mis-ranked candidates. Affected person: main factors and challenge route on request
Prediction that a university student is at risk of withdrawal, used to trigger support or progression review	A where it informs progression; B where it only triggers an offer of support	Dashboard of risk scores for staff	Student: plain-language reasons and how to correct data; accessible formats. Staff: guidance on uncertainty and on not treating a score as a judgement of ability. Auditor: outcomes by disability, ethnicity and other characteristics

Decision	Tier	Common practice	MEM specification (selected elements)
Clinical decision support suggesting a triage category	A	Suggested category with confidence score	Clinician: case-specific uncertainty and model limits, evaluated by calibrated-reliance testing. Patient: information that AI was used and who made the decision

Two observations emerge from the application. First, in every Tier A case the explanation that matters most for preventing harm is the operator's, because it is the operator's judgement that the law relies on as a safeguard. Yet current practice rarely tests operator explanations for their effect on reliance. Second, the affected person's explanation is usually the least developed, despite being the one the law most clearly requires. The fourth scenario also exposed a gap in the first version of the model: a single decision can move between tiers depending on how its output is used, so the specification must follow the use of the output, not the system that produced it. The model was revised accordingly.

4.5 A worked example: a credit refusal

To show the model in operation, consider a lender that uses a third-party credit score as a major input to its lending decisions. Under *OQ v Land Hessen* (2023), the score itself may be an automated decision under the GDPR. The decision is Tier A.

The lender identifies three recipients. The applicant needs to understand why they were refused and what they can do. The underwriter who reviews borderline cases needs to know when the score is unreliable. The regulator needs to be able to check that the scoring and the decision were lawful.

For the applicant, the specification requires an explanation that names the two or three factors that most affected the outcome, states in counterfactual terms what change would have altered it (for example, a lower level of existing borrowing), explains how to correct any inaccurate data and sets out the route to human review. Before deployment, the lender tests a sample explanation with a representative group of applicants, including people with low literacy and disabled people using assistive technologies, and checks that most can correctly state the reasons and the next steps. Where they cannot, the wording is revised and retested.

For the underwriter, the specification requires that cases near the decision threshold or with unusual data be flagged as uncertain, and that the underwriter record their own assessment before seeing the score in those cases. The lender monitors how often underwriters overturn the score and whether their overrides are borne out by later repayment, as a measure of calibrated reliance.

For the regulator, the specification requires documentation of the scoring logic and its validation, and arrangements to provide protected material under confidentiality if required, consistent with the balancing approach in *CK v Magistrat der Stadt Wien* (2025).

The example shows how the model turns an abstract duty to provide "meaningful information" into concrete content, forms and tests, and how the testing burden falls where the risk lies.

5. Discussion

5.1 Meaningful as a measurable standard

The central claim of this paper is that "meaningful", the word that recurs in the GDPR, the AI Act and the CJEU's case law, should be treated as a measurable standard rather than a description. An explanation is meaningful if it enables its recipient to do what the explanation is for. That claim is consistent with the functional reading of the right to explanation advanced by Selbst and Powles (2017) and adopted in substance by the CJEU (*CK v Magistrat der Stadt Wien*, 2025). What the model adds is a way of checking it.

This matters because the alternative, treating meaningfulness as a property of content, is easy to satisfy and easy to game. An organisation can always produce more information. The empirical literature shows that more information is not the same as more understanding, and can reduce people's ability to spot errors (Poursabzi-Sangdeh et al., 2021). A standard based on what recipients can do is harder to satisfy with volume alone.

5.2 The operator problem

The most important practical implication concerns operators. Both the AI Act and UK data protection law rely on human involvement as a safeguard against harmful automated decisions. But the experimental evidence suggests that explanations given to operators often increase deference to the AI (Bansal et al., 2021; Buçinca et al., 2021), reproducing the automation bias long documented in other settings (Goddard et al., 2012). If that is so, a deployer could comply with an oversight requirement in form while weakening it in substance. MEM addresses this by making calibrated reliance a test criterion for operator explanations at Tier A, and by recommending designs, such as asking operators for their own judgement first, that the evidence suggests reduce over-reliance. The cost, as Buçinca et al. (2021) found, is that such designs are less popular with users. Regulators may need to require them, because organisations and operators are unlikely to choose them unprompted.

5.3 Interpretable by design

The model's fidelity criterion favours interpretable-by-design models for Tier A decisions, following Rudin (2019). This is a preference rather than a rule. Some high-stakes tasks may require complex models, and post-hoc methods can be useful when validated against the model's actual behaviour. But where an interpretable model performs comparably, using an opaque model and then approximating its reasoning after the event introduces a risk of infidelity that the deployer has chosen to take. Under a functional standard that requires explanations to reflect the procedures actually applied, that choice should have to be justified.

5.4 Objections

Three objections can be anticipated. The first is cost. Testing explanations with representative recipients is more expensive than producing them. The model limits this by confining full testing to Tier A decisions, which are a minority in most organisations, and by allowing tests to be run on templates rather than on every individual explanation.

The second objection is that testing can be gamed: organisations might select recipients who understand easily, or design tests that are easy to pass. This is a real risk, and it is the reason the model specifies representative recipients, including people with lower literacy and disabled people. Regulators could reduce the risk further by publishing standard test protocols, much as accessibility standards specify how conformance is checked.

The third objection is that some decisions cannot be explained in terms a lay person can understand. The CJEU's answer is that the complexity of a system does not excuse the controller from the duty to explain (*CK v Magistrat der Stadt Wien*, 2025). MEM's answer is similar: if no explanation of a Tier A decision can pass the comprehension test for its affected recipients, that is evidence that the decision process is not suitable for its use, not evidence that the test is wrong.

5.5 Relationship to existing guidance

The model complements rather than replaces existing guidance. The ICO and Alan Turing Institute (2020) guidance sets out what kinds of explanation organisations might provide; MEM adds who each is for, how much effort each deserves and how to tell whether it works. Because explanation duties differ between the EU and the UK, MEM also offers a common specification that can be applied across regimes, with the risk tier aligned to whichever legal category applies.

6. Conclusion

Law in Europe and the UK now asks for explanations of AI-informed decisions that are meaningful to the people who receive them. Empirical research shows that many explanations fail that test: they raise trust without raising understanding, overload rather than inform, and are designed for engineers rather than for the people affected. This paper has argued that the meaningfulness of an explanation should be judged by what it enables its recipient to do, and has proposed the Meaningful Explanation Model to put that into practice. The model specifies explanations by decision risk tier, by recipient and purpose, and by five effectiveness tests: comprehension, actionability, calibrated reliance, fidelity and accessibility. Applied to five regulated decisions, it shows that current practice most often neglects two explanations: the one given to the affected person, which the law most clearly requires, and the one given to the operator, on which human oversight depends.

The model has limitations. It has not been tested with real recipients, the legal analysis covers only the EU and UK, and the UK reforms were still being implemented at the time of writing. Three lines of research would address these limitations. First, field studies in which deployers

of Tier A systems apply the comprehension and actionability tests to real explanations, measuring whether affected people can understand and challenge decisions. Second, operator studies in regulated settings measuring whether explanation designs improve calibrated reliance, extending laboratory findings to practice. Third, research with disabled recipients and those with lower literacy, to develop accessible explanation formats and test protocols that do not exclude the people most often affected by automated decisions.

A right to explanation was always meant to give people a way to understand and question decisions made about them. The test of any explanation is whether it does that. Organisations, regulators and designers now have the legal basis to apply that test. The model proposed here offers a way to do so.

Declarations

Funding: This research received no external funding.

Conflicts of interest: The author declares no conflicts of interest.

References

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. https://doi.org/10.1145/3411764.3445717

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 648–657). ACM. https://doi.org/10.1145/3351095.3375624

Burges Salmon. (2025). *Automated decision-making: The impact of the Data (Use and Access) Act 2025*. https://www.burges-salmon.com/articles/102kprb/automated-decision-making-the-impact-of-the-data-use-and-access-act-2025-and-a/

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

CK v Magistrat der Stadt Wien (Dun & Bradstreet Austria GmbH), Case C-203/22 (Court of Justice of the European Union, 27 February 2025).

Data (Use and Access) Act 2025, c. 18. https://www.legislation.gov.uk/ukpga/2025/18

Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a "right to an explanation" is probably not the remedy you are looking for. *Duke Law & Technology Review, 16*, 18. https://doi.org/10.31228/osf.io/97upg

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). ACM. https://doi.org/10.1145/3411764.3445188

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Information Commissioner's Office & The Alan Turing Institute. (2020). *Explaining decisions made with AI*. Information Commissioner's Office. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/

Jabareen, Y. (2009). Building a conceptual framework: Philosophy, definitions, and procedure. *International Journal of Qualitative Methods, 8*(4), 49–62. https://doi.org/10.1177/160940690900800406

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 624–635). ACM. https://doi.org/10.1145/3442188.3445923

Kaminski, M. E. (2019). The right to explanation, explained. *Berkeley Technology Law Journal, 34*, 189. https://doi.org/10.31228/osf.io/rgeus

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM. https://doi.org/10.1145/3313831.3376219

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). ACM. https://doi.org/10.1145/3313831.3376590

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43. https://doi.org/10.1145/3233231

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

OQ v Land Hessen (SCHUFA Holding AG), Case C-634/21 (Court of Justice of the European Union, 7 December 2023).

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC Methods Programme*. Lancaster University. https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–52). ACM. https://doi.org/10.1145/3411764.3445315

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233–242. https://doi.org/10.1093/idpl/ipx022

Stibbe. (2025). *CJEU confirms "right of explanation" in battle between trade secrets and algorithmic transparency*. https://www.stibbe.com/publications-and-insights/cjeu-confirms-right-of-explanation-in-battle-between-trade-secrets-and

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law, 7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *26th International Conference on Intelligent User Interfaces* (pp. 318–328). ACM. https://doi.org/10.1145/3397481.3450650