INTELLIGENCE · ARTIFICIAL AND COLLECTIVE INTELLIGENCE

AI Explainability Requirements in Intelligence Analytic Support

When AI supports intelligence analysis, analysts and decision-makers need to know why a system produced its output and how far to trust it. This article explains what explainability means, the main techniques, their limits, and the practical requirements an intelligence function should set for any AI tool it uses.

ModQu research article · 1,586 words · 17 references (APA 7th edition)

Intelligence analysis has always been judged not only by its conclusions but by its reasoning. Analytic standards require analysts to describe the quality and credibility of their sources, to explain their assumptions and to distinguish evidence from judgement (Office of the Director of National Intelligence [ODNI], 2015). When AI tools contribute to analysis, by scoring risks, ranking documents, flagging anomalies or drafting text, the same standards apply. A conclusion that cannot be explained cannot be properly checked, challenged or defended. **Explainability** is the property that allows people to understand why an AI system produced a given output.

What explainability means

The terms are used loosely. Doshi-Velez and Kim (2017) define **interpretability** as the ability to explain or present something in understandable terms to a human, and argue that what counts as understandable depends on the user and the task. Lipton (2018) warns that "interpretability" is often used as if it were a single property when it covers several distinct goals: trusting a model, understanding cause and effect, ensuring a model will transfer to new conditions, informing human decisions and ensuring fairness. Different goals call for different kinds of explanation.

Miller (2019) draws on research in philosophy and psychology to show how people actually explain things to each other. Human explanations are **contrastive**: they answer "why this rather than that?" They are **selective**, citing a few main causes rather than all of them. And they are **social**, shaped by what the person asking already knows. Many technical explanation methods ignore these features, producing lists of factors that are accurate but unhelpful.

Why it matters in intelligence

Explainability serves several purposes in an intelligence setting.

- **Checking.** Analysts need to know whether a system's output rests on sound evidence or on a quirk of the data.
- **Calibrating trust.** Analysts should rely on a system where it is strong and question it where it is weak. Explanations, along with known accuracy, help them tell the difference.
- **Communicating confidence.** Products must state how confident analysts are and why. An unexplained machine score cannot support that statement.

- **Accountability.** Oversight bodies and courts may need to know how a conclusion was reached, especially when it affects individuals.
- **Learning.** Explanations help analysts and developers find and correct errors.

The principles for AI in the American intelligence community published by the Office of the Director of National Intelligence (2020) include transparency and accountability, asking agencies to provide appropriate transparency about how systems work and to keep humans responsible for judgements. In the American Defense Advanced Research Projects Agency's explainable AI programme, Gunning and Aha (2019) set the aim as systems whose users could understand, appropriately trust and effectively manage them.

The four principles

Researchers at the American National Institute of Standards and Technology proposed four principles for explainable AI (Phillips et al., 2021). Figure 1 summarises them.

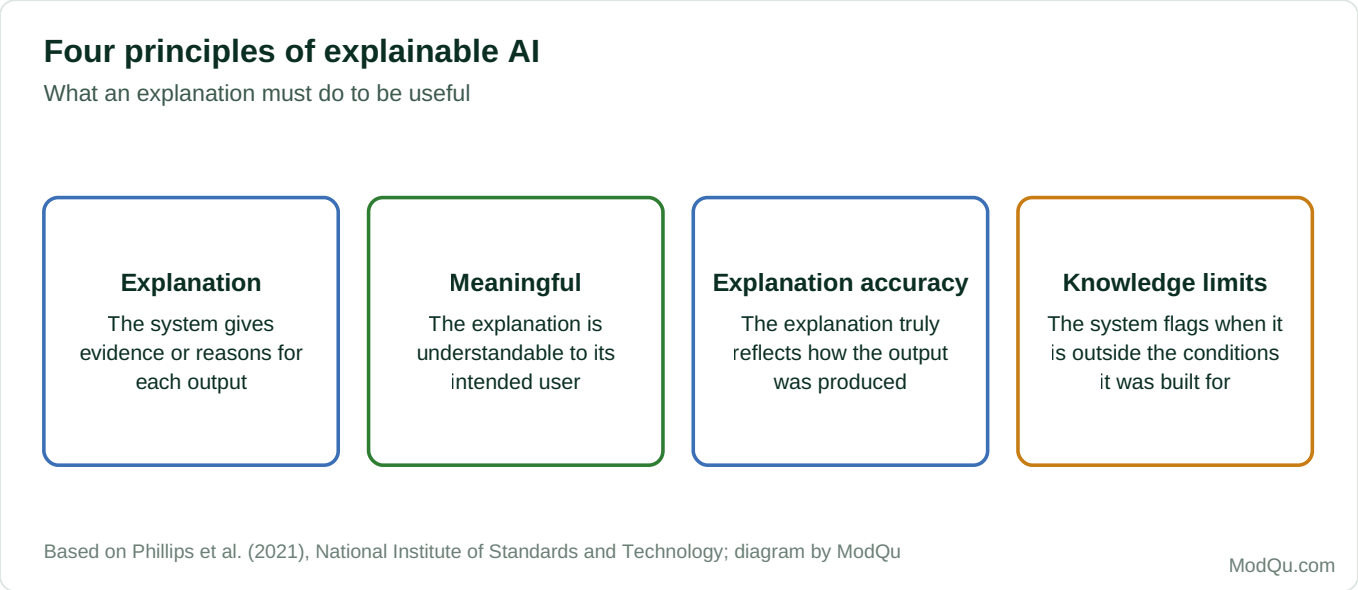


Figure 1. Four principles of explainable AI. An explanation is useful only if it is given, understood, faithful to how the system worked and honest about the system's limits. Based on Phillips et al. (2021); diagram by ModQu.

The third and fourth principles are often neglected. An explanation can be clear and persuasive yet fail to describe what the system actually did. And a system that cannot recognise when it is outside its competence will give confident outputs in conditions where it should not be trusted.

Techniques for explanation

Adadi and Berrada (2018) survey the main approaches. Two distinctions are useful.

Interpretable models versus post-hoc explanations. Some models are understandable by design: short decision rules, small decision trees, or scoring systems with a few weighted factors. Others, such as large neural networks, are too complex to inspect directly, so separate methods are used to explain their outputs after the fact.

Global versus local explanations. A global explanation describes how a model behaves in general, for example which factors matter most across all cases. A local explanation describes why it produced one particular output.

Two widely used post-hoc methods give local explanations. LIME, proposed by Ribeiro, Singh and Guestrin (2016), explains a single prediction by fitting a simple model to the complex model's behaviour in the neighbourhood of that case. SHAP, proposed by Lundberg and Lee (2017), assigns each input a contribution to the prediction using a method from cooperative game theory. Both can show, for example, that a document was flagged mainly because of three particular phrases.

Wachter, Mittelstadt and Russell (2017) proposed **counterfactual explanations**: statements of the smallest change to the inputs that would have changed the result, such as "the application would have been approved if annual income had been £5,000 higher". Counterfactuals match Miller's (2019) finding that people seek contrastive explanations, and they can be given without revealing the full workings of a model.

The limits of explanation

Explanation methods have important weaknesses. Rudin (2019) argues that post-hoc explanations of black-box models are approximations that may be unfaithful to the model, and that for high-stakes decisions organisations should use inherently interpretable models instead. Rudin points out that in many domains interpretable models perform as well as complex ones, so the supposed trade-off between accuracy and interpretability is often smaller than assumed.

Explanations can also mislead users. Bansal et al. (2021) found that providing AI explanations increased people's acceptance of recommendations whether those recommendations were right or wrong. Explanations made people more confident without making them better at spotting errors. Heuer (1999) showed that analysts tend to accept information that fits their expectations; a plausible explanation for a machine output can reinforce that tendency.

Regulatory requirements

Explainability is increasingly a legal requirement as well as good practice. The European Union's Artificial Intelligence Act requires high-risk AI systems to be designed so that those deploying them can interpret their outputs and use them appropriately, and gives people affected by certain decisions based on such systems a right to an explanation of the role the system played (European Union, 2024). The AI risk management framework of the National Institute of Standards and Technology (2023) treats explainability and interpretability as characteristics of trustworthy AI, alongside validity, reliability, safety, fairness and accountability.

Setting requirements for intelligence tools

Different audiences need different explanations. Figure 2 sets out four.

Who needs what explanation

Explainability requirements differ by audience

Audience	What they need to know	Suitable form
Analyst	Why this output, and can I trust it here?	Key factors, sources, confidence, similar cases
Manager or client	How reliable is the product overall?	Accuracy record, limits, plain summary
Oversight body	Is the system lawful, fair and controlled?	Documentation, tests by group, audit logs
Affected person	Why was I treated this way, and what would change it?	Main reasons and what would alter the result

Compiled from Miller (2019), Gunning and Aha (2019) and Wachter et al. (2017); table by ModQu

ModQu.com

Figure 2. Who needs what explanation. Requirements should be set for each audience the system serves. Compiled from Miller (2019), Gunning and Aha (2019) and Wachter et al. (2017); table by ModQu.

An intelligence function adopting an AI tool can set practical requirements:

1. **Prefer interpretable models** where they perform comparably, especially for decisions affecting individuals.
2. **Require source traceability.** Every output should link to the documents or data that produced it.
3. **Require local explanations** that show the main factors behind each output, in terms analysts understand.
4. **Require knowledge limits.** The system should flag cases unlike its training data or where its confidence is low.
5. **Require documentation.** Model cards, as proposed by Mitchell et al. (2019), should describe intended use, performance across conditions and known limitations.
6. **Test explanations with users.** Check that explanations actually help analysts identify wrong outputs, not just feel more confident.
7. **Record reliance.** Log when analysts accepted, adjusted or rejected outputs, so that the value of the tool can be assessed.

Explaining language model outputs

Language models raise particular difficulties. Their outputs are fluent text, and they can produce an explanation of their own reasoning on request. But such self-explanations are generated in the same way as any other text and are not guaranteed to describe how the answer was actually produced, which breaks the explanation accuracy principle. For intelligence use, the most reliable form of explanation for a language model is not its account of itself but **traceability**: requiring every statement in its output to be linked to a passage in a specified source, so that an analyst can check the evidence directly. Where a statement cannot be traced, it should be treated as unsupported.

Common mistakes

- **Treating a feature list as an explanation.** A ranked list of inputs may be accurate but does not answer the analyst's real question: why this case, rather than another?
- **Assuming explanations improve decisions.** They may only increase confidence. Test their effect.
- **Explaining a model that should have been simpler.** If an interpretable model performs almost as well, use it.
- **Ignoring the limits.** An explanation of an output produced outside the model's competence is still an explanation of an unreliable output.

A business illustration

A market intelligence team uses a model to score the likelihood that each of several hundred distributors will switch to a competitor. Sales managers are sceptical of the scores and ignore them. The team adds two features: for each distributor, a short list of the three factors that most raised or lowered the score, and a counterfactual statement of what would most reduce the risk. It also replaces the original complex model with a simpler scoring model that performs almost as well. Managers begin to use the scores, often adding information the model lacks, and the team records where they disagree. Within two quarters, the combined judgement of model and managers is more accurate than either.

Conclusion

Explainability is the bridge between AI output and human judgement. In intelligence work, where conclusions must be checked, qualified and defended, an output that cannot be explained cannot be relied on. Good explanations are meaningful to their audience, faithful to how the system actually works and honest about its limits. Intelligence functions should prefer interpretable models where possible, require traceability and knowledge limits from every tool, and test whether explanations improve judgement rather than merely increasing confidence.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

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* (Article 81). ACM.

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning* (arXiv: 1702.08608). arXiv.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689.

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence, Central Intelligence Agency.

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

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).

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

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM. https://doi.org/10.1145/3287560.3287596

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Office of the Director of National Intelligence. (2015). *Intelligence Community Directive 203: Analytic standards*.

Office of the Director of National Intelligence. (2020). *Principles of artificial intelligence ethics for the intelligence community*.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence* (NIST Interagency Report 8312). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

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

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.