Guest editor

Eystein Thanisch

Contributing authors

Joseph Byrum
Marcus Evans
Joseph Farrington
Lisa Fitzgerald
Daniel Flatt
Lily Hands
Rosie Nance
Kitty Yeung

CUTTER
AN ARTHUR D. LITTLE COMMUNITY

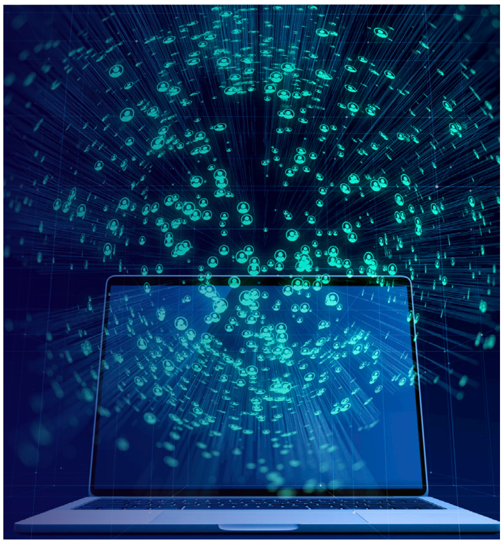
AMPLIFY

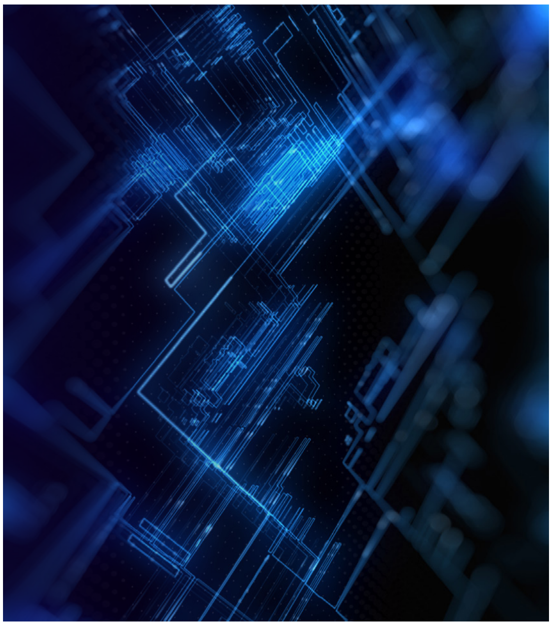
Vol. 38, No. 5, 2025

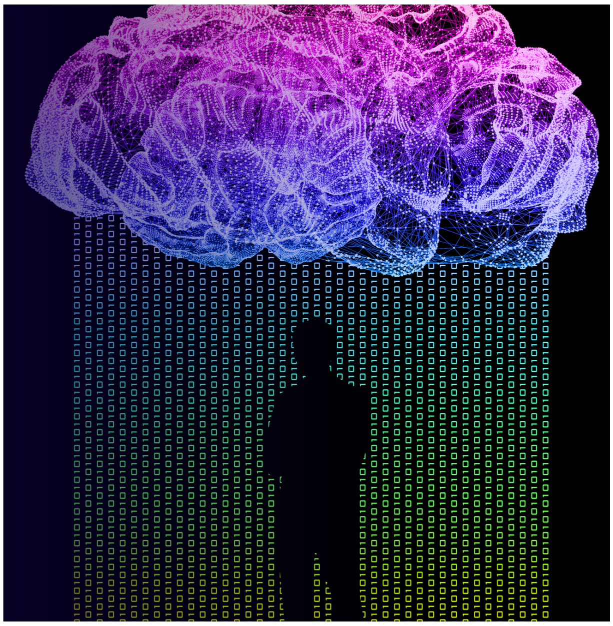
Anticipate, Innovate, Transform

Disciplining AI, Part I: Evaluation Through Industry Lenses

CONTENT







DISCIPLINING AI, PART I: EVALUATION THROUGH INDUSTRY LENSES

BY EYSTEIN THANISCH, GUEST EDITOR

For better or for worse, AI is set to have a major impact on business and society. Making AI technologies accountable through the disciplined and systematic evaluation of their effects is thus becoming both a matter of public safety and organizations' ROI.[1]

Many would argue that the AI industry is not accountable enough, particularly around intellectual property, privacy, bias, and social ramifications. AI-model benchmarking, however, is a prominent and influential aspect of the industry. Model developers and researchers have devised numerous sets of standardized tests that measure performance in areas like coding, math, reasoning, factual accuracy, and visual problem-solving, as well as aspects like safety or jailbreak vulnerability. The results can be compiled into leaderboards that purport to identify the best models for a given use case.

Nevertheless, AI benchmarking is increasingly seen as falling short of satisfactory evaluation of AI systems.[2] Benchmarking's inadequacies include underscrutinized test data, model developers teaching to the test, and even the possibility that models themselves "know" when they're being tested and feign the required responses.

Equally concerning, results from testing models on academic tasks taken out of context often have limited relevance to real-world applications, no matter how advanced the tasks may be. As frontier large language models show signs of plateauing in performance, attention is increasingly shifting toward how these models are applied in practice.

The discipline of AI evaluation aims to quantify the quality of the responses of entire AI systems, concrete and in context, as well as their individual components. Modeling even just the relevant aspects of the context and tracking the nondeterministic outputs from the AI is, of course, hard, but it is increasingly recognized by tech leaders and investors as critical.[3] Another outstanding question is whether and by what means one can evaluate *how* AI systems arrive at their output, with a view to verifying and explaining that output.

Testing, however elaborate, has proved imperfect at reliably evaluating AI's usefulness when integrated into human workflows across organizations and societies. In a recent study, access to market-leading agentic AI coding tools *increased* software developers' completion times of actual tasks on mature projects by 19%.[4] Even the technology's most ardent supporters acknowledge this to be a major frontier.[5]

Here in Part I of this two-part *Amplify* series on AI evaluation, we explore the impetus toward AI accountability that arises from tackling real problems in real-world settings. Understanding how AI can contribute, at what cost, and with what nth order effects in a given context requires rigorous socio-technical systems thinking.

IN THIS ISSUE

In pursuit of such thinking, drawing on experience across industries and disciplines, this issue of *Amplify* offers insights into the criteria that determine AI success.

First up, from Marcus Evans, Rosie Nance, Lisa Fitzgerald, and Lily Hands, AI explainability is a legal requirement as well as a scientific challenge. Despite the EU and UK's differing approaches in other aspects of AI regulation, both the EU and UK GDPR continue to uphold individuals' right to an explanation of automated or semiautomated decisions that impact them significantly. The EU AI Act also provides a right to an explanation for individuals or organizations.

> AI BENCHMARKING IS INCREASINGLY SEEN AS FALLING SHORT OF SATISFACTORY EVALUATION OF AI SYSTEMS

Organizations using AI-powered tools to make an impactful decision where there is an EU or UK connection must be able to explain how the decision is made, and the explanation must be intelligible to the lay citizen. In the face of these legal and social obligations, the explainability and evaluability of a system should be among the key factors considered when architecting or procuring. The authors conclude with practical advice on how to promote explainability via AI governance.

Next, Daniel Flatt contends that an evaluation framework that promotes accuracy and objectivity is a commercial necessity. He points to B2B publishing as a sector of particular note: an industry built on credibility and accountability that AI could undermine, even as it offers opportunities to accelerate time to output. In response, new tools for detecting inaccuracy and bias in AI-generated copy are emerging, alongside collaborations across the publishing workflow and the wider industry.

The relationship between AI and publishing is increasingly bidirectional. AI model developers seek partnerships with reputable publishers to access both content and brand credibility, while publishers must weigh such collaborations carefully, ensuring that models deliver quality output. With its long tradition of fact-checking, journalism brings valuable expertise to this challenge. Flatt calls for an approach that safeguards and advances journalism's core ideals while setting a broader standard for the responsible use of AI.

Likewise, Kitty Yeung urges us to elevate our thinking and consider what we are trying to achieve — via AI or otherwise. She argues that the fashion industry has long failed to appreciate the imaginative journeys consumers are taking, journeys that weave together self, situations, social circles, and eclectic wearables. Destructive practices like fast fashion represent flawed attempts to address human complexity with incomplete information, cumbersome supply chains, and a narrow anthropology that undervalues consumers' creative agency.

In contrast, AI provides digital try-on tools that allow users to experiment with items in any combination or context — or even design new ones themselves. AI-enabled analysis can then surface the trends emerging from these creative interactions, helping to shape a smarter, leaner supply chain. For Yeung, realizing this potential requires AI evaluation to move beyond mere compliance with the status quo and instead align with the higher ideals of freedom, truth, and sustainability.

Joseph Farrington also emphasizes the importance of evaluating AI systems against their end goals. In healthcare, where developing and deploying AI models is especially challenging, he argues for first modeling the business context and processes the AI will interact with — before moving ahead with development or deployment. This approach can be used to assess, in advance, whether a plausible AI model will provide the intended benefit. It can also be used to run alternative scenarios to identify what else might need to change for the AI to really work or what else might work better if the AI were in place.

As a secondary benefit, context modeling brings engineers, stakeholders, and domain experts together much earlier than would typically occur in such projects. In this way, AI evaluation is not merely a corrective exercise after deployment; it should also anticipate, contextualize, and define — in concrete terms — what the AI is meant to achieve.

Closing the issue, Joseph Byrum introduces a framework to help organizations plan rationally and prudently for AI adoption. One element is defining performance thresholds beyond which emerging technologies become economically viable. Another is assessing how AI and humans should interact across different business functions: some tasks can be commoditized and handled by AI, while others remain critical differentiators under human responsibility, with hybrid possibilities in between.

This analysis is not straightforward. Byrum points to cases like UPS and its ORION route navigation system, where organizations had to undergo radical iterations to find the right balance between AI and human input. Complicating matters further are the rapid pace of technological development and the shifting nature of market differentiators, which make any framework less a static blueprint and more a matter of dynamic "adaptive sensing."

KEY THEMES

The contributors to this issue of *Amplify* all agree that true AI evaluation must go beyond assessing models or outputs alone. Instead, it should be grounded in an organization's strategies, end goals, and sources of differentiation.[6]

AI is on a trajectory to become too fundamental and impactful to be treated as just another tool evaluated only on its outputs. Its influence extends beyond organizational success into society at large, and society has expectations around human dignity and empowerment. Any evaluation framework must therefore consider the extent to which AI advances — or at the very least does not undermine — these core values.

It is also a mistake to imagine that business processes and structures will remain constant and provide an immutable viewpoint from which AI can be comfortably evaluated. As several contributors note, the rollout of the technology is already driving new ways of working, requiring more interdisciplinary/intermural collaboration and formalization of long-tacit knowledge. Looking ahead, it has the potential to transform supply chains and upend established economies of expertise. The status quo is not a valid yardstick.

Several contributors propose forward-looking approaches to AI evaluation — identifying where adoption will be most effective, estimating its likely benefits, and addressing additional requirements like explainability once deployment occurs. Synthesizing deep context and the interplay of action and reaction is central to systems thinking,[7] which provides a valuable framework for interpreting the contributions in this issue.

This issue also highlights profound challenges: the culture-specific nature of human imagination and self-perception, the difficulty of explaining systems that remain subjects of frontier research, and the organizational self-understanding required to model all the factors — including human expertise — that shape a process.

Indeed, the very notion of truth is in play — both in the journalistic sense and in the realm of authentic human expression. While no one suggests there are easy solutions, the contributions in this issue offer grounded and thought-provoking approaches.

In Part II of this *Amplify* series, we'll take a closer look at some of the engineering and conceptual challenges of AI evaluation.

ACKNOWLEDGMENT

Alongside the contributors and the ADL Cutter team, I would like to thank Alaa Alfakara, Michael Bateman, Maureen Kerr, Brian Lever, Olivier Pilot, and Greg Smith for their support.

REFERENCES

[1] Jones, Elliot, Mahi Hardalupas, and William Agnew. "Under the Radar? Examining the Evaluation of Foundation Models." Ada Lovelace Institute, 25 July 2024.

[2] Eriksson, Maria, et al. "Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation." arXiv preprint, 25 May 2025.

[3] Gupta, Aakash. "Why AI Evals Are the New Unit Tests: The Quality Assurance Revolution in GenAI." Medium, 11 June 2025.

[4] Becker, Joel, et al. "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." arXiv preprint, 25 July 2025.

[5] Mollick, Ethan. "The Bitter Lesson Versus the Garbage Can." *One Useful Thing*, 28 July 2025.

[6] For more on this line of thinking, see: Kolk, Michael, et al. "Innovation Productivity Reloaded: Achieving a 40% Boost Using a People-Centric AI Approach." Arthur D. Little, July 2025.

[7] Bansal, Tima, and Julian Birkinshaw. "Why You Need Systems Thinking Now." *Harvard Business Review*, September–October 2025.

About the guest editor

EYSTEIN THANISCH

Eystein Thanisch is a Senior Technologist with Arthur D. Little (ADL) Catalyst. He enjoys ambitious projects that involve connecting heterogeneous data sets to yield insights into complex, real-world problems and believes in uniting depth of knowledge with technical excellence to build things of real value. Dr. Thanisch is also interested in techniques from natural language processing and beyond for extracting structured data from texts. Prior to joining ADL, he worked on Faclair na Gàidhlig, the historical dictionary of Scottish Gaelic, on a team tasked with building a tagged corpus of transcriptions from pre-modern manuscripts. He also was involved in IrishGen, a project on the use of knowledge graphs to represent medieval genealogical texts. Dr. Thanisch also worked as a freelance editor and analyst for a number of IGOs and academics. He earned a master of science degree in computer science from Birkbeck, University of London, and a PhD in Celtic studies from the University of Edinburgh, Scotland. He can be reached at experts@cutter.com.

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**

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.

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]

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.

[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.

AI IN B2B PUBLISHING: THE PROMISE & THE PERIL

Author

Daniel Flatt

The B2B publishing sector has long provided critical insights and domain-specific intelligence to professional audiences. Today, it stands at a crossroads. With generative AI (GenAI) systems maturing rapidly, publishers that choose not to leverage these new technologies to sharpen their editorial and commercial edge may find themselves falling behind their AI-first competitors.

Drawing on two decades of editorial experience in financial media, this article examines areas where AI's promise must be balanced with due diligence, including deployment, legal accountability, commercial viability, and ethics. GenAI is less a silver bullet and more a powerful collaborative tool. Increasingly, the challenge for publishers is not deciding whether to adopt AI, but how to evaluate its effectiveness, reliability, and alignment with journalistic standards.

FROM PROMISE TO PRACTICALITY

The application of AI in B2B publishing is relatively nascent, but the pace of adoption is accelerating. Many publishers are experimenting with AI solutions, such as automatic earnings call summaries, reader personalization, automated interview transcription, and sophisticated data visualization. Nevertheless, confusion and skepticism remain about AI's reliability and impact.

Many publishers are approaching AI cautiously, typically starting with fundamental evaluation criteria. This can range from setting bars for the AI to hurdle to tailored metrics to measure. For example, does the tool significantly reduce production times (ideally by 30% or more) without increasing editorial revisions? Does it consistently reflect the publisher's distinct voice and editorial standards? Does it maintain accuracy, introducing fewer than two factual or numerical errors per 1,000 words?

One thing is certain: effective AI deployment requires hybrid workflows, including AI for efficiency, journalists for exclusivity, and editors for nuance and brand protection. Publishers must work closely with vendors to codesign tailored solutions, avoiding off-the-shelf deployments. Human oversight remains critical, particularly in regulated industries, underscoring the need for careful integration into existing workflows.

TRANSPARENCY & TRUST

Ethical and legal scrutiny around AI-generated content is intensifying. Business audiences demand complete transparency about data sources, making rigorous audits of training data's provenance, licensing, and transparency indispensable.

> **THE APPLICATION OF AI IN B2B PUBLISHING IS RELATIVELY NASCENT, BUT THE PACE OF ADOPTION IS ACCELERATING**

In the UK, for example, fierce debates arose when ministers wished to block a 2025 House of Lords amendment intended to mandate disclosure of copyrighted material used in AI training. Although ultimately withdrawn, the ministers' apparent lack of concern for transparency risked undermining journalistic rights.

For B2B publishers, the reputational stakes are particularly acute. Protecting credibility requires clear policies on data provenance and robust licensing agreements. It also requires meticulous auditing of all AI output, treating it as one would consider material from human journalists, incorporating source attribution, fact-checking, and named oversight.

MEASURING COMMERCIAL IMPACT

Of course, successful AI adoption isn't solely about editorial effectiveness — it's also about demonstrable commercial outcomes. Publishers increasingly rely on clear, quantifiable metrics to differentiate genuine ROI from technological novelty.

For instance, a midsize US publisher worked with Leverage Lab to implement AI-powered customer segmentation tools. The result was an 80% reduction in subscriber acquisition costs.[1] These improvements were meticulously tracked via real-time dashboards comparing AI-driven results against traditional methods.

Similarly, a UK trade publisher working with an AI data firm doubled its subscription conversion rate by gating AI-generated data insights. Within 12 weeks of going live, the publisher demonstrated clear commercial uplift directly attributable to AI.[2]

Evaluating commercial success requires ongoing monitoring of metrics like churn rates, reader engagement, and lead-conversion improvements. Analytics suites such as Mixpanel and Looker let publishers embed accountability and measure real-time commercial impacts.

EVALUATING SUCCESS: HOW PUBLISHERS KNOW AI IS WORKING

Our experience shows us that publishers generally evaluate AI systems across three dimensions: accuracy, editorial standards, and fairness.

Accuracy remains paramount. Tools such as DeepEval and Ragas measure coherence and faithfulness. Regular nightly batch tests of sampled prompts help publishers ensure outputs consistently surpass predefined accuracy thresholds.

The biggest benefit of these tools is objectivity at scale: automated nightly tests score hundreds of prompts for factual consistency, relevance, coherence, and faithfulness, flagging drifts long before human editors might notice.

This quantitative feedback accelerates model iteration, reduces editorial rework, and builds a defensible audit trail, which is critical when regulators or clients ask, "How do you know it's accurate?" Additionally, because metrics are standardized, publishers can benchmark one model version against another (or compare vendor solutions) using like-for-like scores rather than anecdotal impressions.

However, systematic evaluation carries risks. First, tools can create false confidence if the test set is unrepresentative. For example, models may "game" predictable prompts while still hallucinating on real news. Overreliance on numeric thresholds can nudge editors to publish borderline content because it "passed the score," weakening critical judgment.

Second, there is a resource burden: configuring, maintaining, and interpreting evaluation pipelines demands data science expertise that many mid-tier publishers lack. Finally, proprietary tools introduce vendor lock-in: if a scoring method is opaque, publishers may be unable to contest results or migrate historical benchmarks elsewhere. Used judiciously and paired with human review, evaluation suites are invaluable, but they must never replace newsroom skepticism.

Leading publishers treat AI-generated content with the same scrutiny applied to junior journalists. BloombergGPT, for example, requires rigorous editorial checks for source accuracy, numeric correctness, and clarity.[3] Increasingly, AI-generated articles carry a "double byline," attributing accountability both to the AI system and supervising editors.[4]

Fairness is another essential dimension. Fair-auditing frameworks such as Giskard, Microsoft Fairlearn, and IBM AI Fairness 360 give publishers a structured way to surface demographic bias before flawed copy reaches readers. Their main benefit is granular visibility: by testing model outputs across protected attributes (gender, ethnicity, age, geography, socioeconomic status), they quantify disparities in sentiment, ranking, or error rates that would otherwise lurk unseen.

Dashboards translate statistical measures (e.g., equalized odds, demographic-parity gaps) into color-coded risk flags, letting editors halt publication in seconds when bias scores exceed preset thresholds. This proactive gatekeeping safeguards brand reputation, reduces legal exposure under anti-discrimination law, and supports ethical commitments to diverse readerships.

These tools are not a panacea, however. "Metric myopia" is one hazard: optimizing for a single fairness score can inadvertently worsen others (reducing false positives might inflate false negatives). Second, fairness metrics hinge on the quality and completeness of attribute labels; many datasets lack reliable demographic tags, leading to spurious conclusions.

There is also a context gap: statistical parity may be inappropriate for finance, law, or medicine, where unequal treatment can be ethically justified by risk profiles. Finally, automated shutdowns can disrupt workflows if thresholds are too tight, causing alert fatigue or publication backlogs. Fairness audits are indispensable for modern newsrooms, but they must be accompanied by nuanced editorial judgment and continuous tuning of thresholds.

DEEPENING THE FRAMEWORK

Beyond immediate evaluation metrics, publishers need permanent guardrails. Many are forming AI editorial boards — small cross-functional teams of editors-in-chief, data scientists, commercial leads, and legal advisors. Usually meeting monthly, the board is in charge of overseeing these areas: (1) a risk register listing every AI workflow, its data sources, and known failure modes; (2) a metric charter that defines accuracy, bias, latency, and revenue targets plus escalation paths; and (3) an incident playbook that spells out how to pause or roll back a faulty model and communicate with subscribers or regulators. By minuting each review and circulating findings newsroom-wide, boards turn AI evaluation from a siloed data science task into an organization-wide discipline.

Oversight should also extend beyond the walls of a single publisher. Structured peer benchmarking lets competing outlets compare results without revealing proprietary data. Participants can export de-identified evaluation logs (e.g., DeepEval scores, bias indices, click-through lifts) to a neutral analytics partner that aggregates and ranks performance.

For example, quarterly reports can reveal whether a "good" 0.92 faithfulness score is industry-leading or merely average and spotlight systemic drifts after major model upgrades. Because identities are masked under nondisclosure agreements, competitive sensitivities remain protected while the sector as a whole moves toward shared accuracy, fairness, and reliability standards.

In early 2024, the *Süddeutsche Zeitung*, Germany's largest broadsheet, set up an internal board to coordinate all editorial AI initiatives.[5] The board includes the editor-in-chief, product and visual desk leaders, data science engineers, HR, IT, and legal counsel. It reviews every new GenAI workflow and signs off on evaluation metrics — and it can halt deployment if standards slip.

EMBEDDING TRUST, EXPLAINABILITY & FAIRNESS

Evaluating AI extends to broader issues of trustworthiness, explainability, and fairness — principles outlined by the US National Institute of Standards and Technology (NIST). Publishers are tasked with translating these standards into practical metrics, such as explainability ratios, severity indices for errors, and bias measures across protected classes.[6]

A 2021 study in the *Journal of Biomedical Informatics* highlights that explainability significantly influences user trust, particularly in high-stakes environments, underscoring the necessity of transparency around AI decisions.[7]

LICENSING RECIPROCITY: TOWARD A SUSTAINABLE AI ECOSYSTEM

The relationship between GenAI developers and publishers is evolving, with encouraging trends toward collaborative licensing agreements (tracked by nonprofit Ithaka S+R). These agreements (typically linked to allow AI developers access to content for training) provide new revenue streams and foster collective benchmarking, helping publishers establish shared industry standards.

A collaborative license agreement is a legal arrangement in which a publisher and a GenAI developer agree to share access to content and technology under mutually beneficial terms. This often includes the publisher granting the AI developer permission to use its content for training or output generation, while both parties collaborate on attribution, revenue sharing, or codeveloped tools.

But licensing deals shouldn't be just cash-for-content transactions. The most forward-looking agreements embed shared evaluation clauses. A well-structured contract can require the AI provider to:

- **Report model-level metrics** to the publisher at regular intervals (e.g., monthly DeepEval faithfulness scores, numerical error rates on domain-specific data, or Giskard bias indices).
- **Benchmark those metrics** against an agreed-upon industry baseline (e.g., Ithaka S+R consortium reports). If scores drift below a threshold, the provider must retrain or switch models to protect the publisher's brand.
- **Return performance telemetry** to determine how often publisher content is surfaced, clicked, or reused so the newsroom can correlate editorial investment with downstream impact.
- **Allow joint audits** in which the publisher and vendor co-run stress tests on sensitive topics (e.g., market-moving financial data) and publish a summary of findings.

A CALL FOR RESPONSIBLE COLLABORATION

B2B publishers, though niche, are uniquely positioned to demonstrate responsible AI adoption. Indeed, they face an imperative and an opportunity: to deploy AI that enhances, rather than dilutes, journalistic quality.

The path forward requires transparency, robust explainability, and fair licensing practices. Publishers must embed clear evaluation standards (speed, accuracy, fairness, and commercial impact) into every AI initiative. Treating AI as a black box will compromise trust and viability.

Ultimately, responsible collaboration between publishers, technology providers, and industry bodies grounded in shared evaluation standards and collective benchmarks will ensure that GenAI results in smarter, faster, and more ethical journalism.

B2B publishing may be a niche industry, but it is one where accuracy is currency. In an age of automation, that currency must not be devalued. The way forward lies in collaboration, open eyes, fair contracts, and high standards.

REFERENCES

[1] "Built for Publishers, Trusted by B2B: Data That Drives Revenue." Leverage Lab, accessed 2025.

[2] "Clients." Flare, accessed 2025.

[3] "Introducing BloombergGPT, Bloomberg's 50-Billion Parameter Large Language Model, Purpose-Built from Scratch for Finance." Bloomberg, 30 March 2023.

[4] For example, Regulation Asia is a B2B platform covering compliance in Asia. It has standard human bylines, but when a summary is produced for a human byline story, it makes clear that it was created by AI.

[5] Jordaan, Lucinda. "How to Integrate AI into Your Newsroom: 'Not Just as a Tool, But as a Transformative Force.'" World Association of News Publishers, 30 May 2025.

[6] "AI Research — Explainability." US National Institute of Standards and Technology (NIST), 6 April 2020.

[7] Markus, Aniek F., Jan A. Kors, and Peter R. Rijnbeek. "The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies." *Journal of Biomedical Informatics*, Vol. 113, January 2021.

About the author

Daniel Flatt is cofounder and Editor-in-Chief of Flare Data, an AI powered insight platform. In this role, he applies machine learning techniques to identify trends, patterns, and anomalies across diverse datasets. Previously, Mr. Flatt launched and led *Capital Monitor* at the New Statesman Media Group, an award-winning journal focused on sustainable finance driven by data analysis. From 2016 to 2020, he served as Editorial Director of Haymarket Media's Financial Media division in Asia, where he oversaw *FinanceAsia*, *AsianInvestor*, and *CorporateTreasurer*, the latter of which he founded. With over two decades of experience in data mining, sustainability reporting, and cross-border finance, Mr. Flatt now channels his expertise to build AI-enabled editorial tools that drive smarter decision-making. He can be reached at daniel@flaredata.ai.

MEASURING AI'S IMPACT ACROSS THE FASHION VALUE CHAIN

Author

Kitty Yeung

Generative AI (GenAI) is reshaping creative industries with unprecedented speed. Fashion, often considered slow to adopt change, is becoming one of its most compelling and dynamic arenas. No longer just the domain of magazines, commerce, and runway shows, fashion is deeply intertwined with data, social behavior, and digital storytelling. From trend analysis and virtual try-ons to content creation, design, and made-to-order, the fashion industry is actively experimenting with AI applications.

Fashion is a trillion-dollar global industry, with more than half of spending coming from womenswear and about 90% of its content shaped by user-generated media. It thus offers a compelling case study in applied AI accountability. However, in the face of rapid transformation, traditional markers of success such as image resolution, model latency, or benchmark accuracy are no longer sufficient. In fact, they are fast becoming irrelevant. As GenAI models grow more accessible, performant, and cost-effective, high-quality output is becoming a baseline expectation rather than a competitive advantage.

Much like the evolution of IT infrastructure in the early 2000s, where capabilities once considered strategic became commoditized utilities, AI is reaching a point where what it can generate is less important than what it enables.

In consumer-driven markets like fashion, where emotional resonance, aesthetic judgment, and cultural context shape adoption, accountability must be measured by AI's ability to empower personalization, creative freedom, economic inclusion, and sustainable practices. The differentiator is no longer the tool itself but the ecosystems it unlocks and the value it creates at the interaction layer.

This article examines AI's real-world accountability across the fashion value chain. Drawing on firsthand experience and data from Wear It AI (a platform developed by the author that allows users to visualize themselves in any style, monetize their content, and customize products), we explore how AI is redefining fashion success metrics. We also extract broader insights that can inform AI accountability in other consumer-facing industries.

> **AI IS REACHING A POINT WHERE WHAT IT CAN GENERATE IS LESS IMPORTANT THAN WHAT IT ENABLES**

AI TOUCHPOINTS

Fashion is not a linear pipeline but a rich, circular journey from ideation to engagement, expression, and iteration. In fashion, AI is not isolated to a single function; it touches every point of the product and consumer journey, each with unique goals and success measures.

PLANNING & FORECASTING

AI helps brands identify what styles will resonate based on real-time social signals, e-commerce behavior, and image trends, essentially predicting what consumers will want next. This reduces guesswork and overproduction, dramatically decreasing planning errors and improving inventory efficiency. What once took months of trend tracking now happens in days.

DESIGNING & PROTOTYPING

General-purpose AI tools can already generate high-fidelity design visuals from text prompts, sketches, and product specs. AI can also match brand tone and preserve product fidelity, removing the need for manually created visuals. Thousands of mood boards, technical drawings, fabric swatches, and so forth, can be generated within minutes, shortening the decision-making cycle across teams. But more than speed, AI empowers a new generation of creators by lowering the barrier to professional-quality ideation, particularly for independent designers and stylists.

SELLING & MERCHANDISING

Decision-making for fashion consumers is largely based on visuals in e-commerce and marketing. Product photography used to cost US $50,000-$100,000 per collection for fashion brands. Now, editorial-quality images can be generated in minutes for a fraction of a cent using AI tools. Instead of photographing physical samples, AI-generated product imagery lets brands create customizable lookbooks featuring diverse models and settings. Products can be marketed before manufacturing, enabling extensive market testing, reducing waste, and empowering creators to build income-generating portfolios from pure imagination.

FITTING & VISUALIZATION

Flat-lay images can be mapped onto customizable AI-generated humans of any body shape and size. GenAI is enabling a new wave of virtual try-on. Users can see themselves styled and accessorized, turning browsing into self-actualization.

MANUFACTURING & SUPPLY CHAIN

This is the least automated but potentially most transformative segment. AI is beginning to support pattern extraction, digital twin development, and sample-less prototyping. Translating digital visions into physical garments remains a complex task due to gaps in material simulation and production readiness, but AI's evolving capabilities hold huge potential for this area.

In the coming year, we expect to see companies selling B2B software-as-a-service (SaaS) products in the above categories be disrupted by the democratized AI approach, which is driven by general-purpose models and open source tools. Essentially, these software packages can now be replicated with just a bit of experimentation, and companies aiming to capitalize on customers' "ignorance" will become obsolete. Many fashion-related SaaS products, including those based on AI, claim exclusivity but are actually repackaged open tools with gated user interfaces. Native AI companies are built in the open and empower every designer, creator, and brand to own their workflows.

CLOSING THE BEHAVIORAL GAP

The lines between designers, creators, brands, and consumers are blurring. Increasingly, consumer goods industries like fashion will be driven by AI-enabled personalization.

Through Wear It AI, we conducted research involving more than 20,000 consumers (89% were women between the ages of 16 and 35). We found that fashion consumers are no longer just shoppers; they play shared roles in creation and buying decisions. Among Wear It AI's users are (overlapping) 57% user-generated content creators, 49% social media influencers, 34% fashion stylists, 31% fashion designers, 20% artists, and 11% professional photographers.

Our research confirms that shoppers spend far more time consuming fashion content than they do shopping for garments. Traditional e-commerce platforms ignore this behavioral gap. Much of e-commerce and digital content creation still relies on outdated metrics: impressions, likes, and follower counts. As consumers shift from passive consumption to active participation (styling, curating, and generating content), social media metrics fail to capture true engagement and commercial value. A content creator might generate hundreds of likes but no conversions. Conversely, one with modest reach but highly tailored content might drive meaningful purchases and repeat engagement.

Given the interweaving of generating content, consuming content, and purchasing the items in the content, fashion companies should consider try-on more of a self-expression mechanism than a purchasing tool. Many virtual try-on tools claim to reduce returns or increase purchase confidence, but evidence remains elusive — will enough retailers adopt the solution compared to free shipping, physical try-on, and free returns?[1,2]

In fact, uncertainty ("Will this look good on me?") is often what drives sales. Despite efforts by Google, Amazon, Snap, and start-ups alike, why hasn't virtual try-on delivered on its promise? Existing efforts primarily focus on embedding virtual try-on solutions into e-commerce shopping sites, aimed at helping shoppers make decisions on the look and sizing of the products. With the advent of GenAI, a wave of standalone apps emerged to test whether virtual try-on can be an engaging consumer shopping experience that leads to more sales, higher conversion rates, and fewer returns. For now, the jury is still out.

Fashion, unlike logistics or search, is emotional and aspirational. Users don't just want to "try on" clothes. They want to become someone or clearly express who they already are. That is why static try-ons and disembodied product simulations feel underwhelming. The best-performing features are those that support storytelling, identity exploration, and social sharing.

This reveals an important lesson: technical fidelity matters, but psychological utility matters more. Through experimentation, we found the following indicators of AI success:

- **Try-on content-creation frequency** — how often users visualize themselves in fashion brands' catalog styles
- **Peer influence** — how likely a user would be to see another user's style and try it on themselves
- **Content-led transactions** — tracking sales driven by user-generated content
- **Retention and return rate** — behavioral metrics indicating satisfaction and utility

These metrics reflect a deeper insight: using GenAI for try-on isn't about automating output. It's about surfacing intent. The focus has shifted from fit-accuracy benchmarks to experiential ones. How many styles do users explore per session? Do they come back to restyle? Do they integrate these looks into their social content?

By allowing users to visualize themselves in any look — styled, customized, and shared from anywhere by anyone — we found that people are less focused on trying individual items and more driven by achieving a cohesive style. This is what fashion brands are missing as they try to sell items to consumers without knowing what they are looking for.

NEXT STEPS

The growth of GenAI coincides with cracks in the influencer economy. A new creator economy is forming, one that emphasizes value over viewership. So far, most of the creator economy remains under-monetized.

We surveyed more than 3,000 Wear It AI's users who identify as content creators and over 30 small and medium-sized fashion brands. They consistently expressed frustrations with existing advertising and sales channels. On platforms like TikTok and Instagram, content creation is saturated and undercompensated. Brands struggle to quantify ROI, and creators struggle against inconsistent income and burnout. Even talented stylists and designers find it hard to convert effort into income.

AI offers a new path — not by replacing their creativity, but by giving them platforms to scale it. By allowing fashion lovers to generate themselves realistically in any style, they can create and model entire collections, monetize directly from looks, and guide product engagement based on their taste. In contrast to traditional influencer marketing, this lets every consumer be an influencer as AI empowers them with high-quality visuals to use as consumable content.

Importantly, consumers can create before owning physical items, and brands don't have to own inventory. They can build style catalogs from digital imagery and AI-generated collections. This shifts power from social media toward individual consumers, defining a new layer of commercial engagement that traditional e-commerce and social media weren't built to support.

AI is poised to disrupt the industry's dominant operating model: fast fashion. Despite its well-documented environmental harm, fast fashion thrives economically by offering a quicker, cheaper solution to recurring human desires.

AI offers the chance to reinvent this equation by moving the data collection of consumer demand to earlier in the trend-prediction phase using imagery alone. By detecting trends on social media and using their established supply chain, fast-fashion brands can run experimental batches to collect validating sales data before scaling up new-product production.

As online retail becomes the dominant commerce channel and consumers rely on digital content to make decisions, AI-generated product imagery may soon precede physical samples entirely. If supply chains can keep pace — and factory automation advances — brands will gain an unprecedented ability to test assumptions about consumer preferences, shopping behavior, and self-expression through style. This shift toward on-demand manufacturing offers a powerful path to eliminating overproduction — ultimately, a critical step toward saving the planet.

These insights from the fashion industry expose the deeper implications of AI adoption. It's not merely a matter of efficiency or cost— it's about access, empowerment, and authenticity. As AI evolves, its true success should be measured by how effectively it enables self-expression. The future lies in open experimentation, where traditional gatekeeping becomes obsolete.

REFERENCES

[1] "What Statistics Indicate the Impact of Virtual Try-On on Returns?" Sustainability Directory, 15 April 2025.

[2] "Virtual Try-On in E-Commerce: A Research Summary." Focal, 12 February 2025.

About the author

Kitty Yeung is a physicist, engineer, and artist as well as cofounder and CEO of Wear It AI. Passionate about bridging science and art, she also founded Art by Physicist, a sustainable STEAM fashion brand, and established the fashion-tech incubation program at Microsoft, where she served as its founding CEO. Dr. Yeung also served as Senior Director of AI at Browzwear. At Microsoft, she led the development of quantum learning materials, including Microsoft Learn Quantum modules and the Quantum Learning website. Dr. Yeung is also creator of the comic series *Quantum Computing & Some Physics* and has delivered lectures on quantum computing through HackadayU and Microsoft Reactor. Her career centers on the integration of technology, science, design, and art. Dr. Yeung is a frequent speaker at conferences and events, delivering technical and career-focused talks that reflect her expertise and passion for quantum computing, fashion-tech, digital transformation, and start-ups. She earned a PhD in applied physics from Harvard University, USA. Dr. Yeung can be reached at artbyphysicist@kittyyeung.com.

A SIMULATION-FIRST APPROACH TO AI DEVELOPMENT

Author

Joseph Farrington

As AI and machine learning (ML) capabilities continue to advance, they will increasingly be embedded into complex operational workflows. In these settings, it is essential to evaluate their impact within an organization using KPIs rather than standard model prediction metrics or benchmark scores. Good models are unlikely to deliver value if they cannot be used within the constraints of a workflow.

In risk-sensitive and regulated sectors, simulation is used to estimate utility only after a model has been developed. As this article explores, however, adopting a simulation-first paradigm is more beneficial: building a workflow simulation to explore how predictions of varying quality will affect KPIs before starting model development. This gives teams an early answer to a critical question: "Would a model even help here, and if so, how good would it need to be?"

This piece first examines why models that perform well on narrow tasks may fail to deliver value in real-world settings. It then reviews how simulation is currently used to evaluate AI systems and introduces a simulation-first approach that places operational context at the center of model development. To illustrate the benefits of this perspective, the article describes a project from the author's research: one aimed at reducing blood-product waste at a major London hospital. (The examples presented are mainly drawn from healthcare; however, the article also considers how a simulation-first approach can be applied in other industries.)

WHY GOOD MODELS FAIL

It is common practice to evaluate traditional ML models using predictive metrics, such as precision, recall, or mean squared error. General-purpose AI models, such as large language models, are typically benchmarked across a wide range of tasks, while enterprise deployments often rely on custom evaluation sets to assess whether a model performs acceptably for a defined use case. These metrics are essential for judging performance on narrow tasks. In relatively simple feedback loops, such as recommending a product or serving an advert, they may be sufficient. In such settings, it is often straightforward to measure business value through A/B testing and iterate quickly using live deployments. But when AI and ML models are incorporated into more complicated workflows, these metrics are no longer a reliable proxy for real-world value. A wide range of factors can limit whether a well-performing model leads to meaningful impact.

THE TIMING OF A PREDICTION CAN BE CRITICAL

The timing of a prediction can be critical. For example, the utility of an ML model developed to identify patients who might benefit from palliative care planning was limited in simulation studies by the fact that hospital staff were often too busy to act on predictions before the patients were discharged.[1]

Resource constraints present another challenge. Accurate sepsis-prediction models may fail to improve patient outcomes if there are only a small number of intensive care unit beds; there may simply be no capacity to respond to the early warnings.[2]

Data quality can also break the link between model performance and value. A diabetic retinopathy screening tool deployed by Google in Thailand struggled in real-world conditions because the retinal scans collected by nurses were often of insufficient quality for the model to analyze.[3]

Human behavior and workflow fit can also limit the benefits of a good model. In the pharmaceutical supply chain, a study found that staff frequently overrode highly accurate algorithmic forecasts in an effort to incorporate their knowledge, and this led to poor predictions.[4] Primary care providers have reported that a key barrier to doctors using decision support tools is the perception that they disrupt the consultation and slow down work.[5]

These examples highlight the need to consider not just the performance on the predictive task, but the broader organizational context in which the model will be used. Good performance on an isolated benchmark may not translate into real-world value if the model's output cannot be acted on effectively within the constraints of the surrounding workflow.

The path from prediction to impact is often more complex than it first appears due to timing, resource availability, data quality, and human factors. To ensure models contribute meaningfully to business or organizational goals, their development should be guided from the outset by a clear understanding of the processes they are intended to support.

CURRENT USE OF SIMULATION FOR EVALUATION

By replicating key elements of the workflow, simulation allows teams to explore the potential impact of a model's predictions with fewer regulatory and safety hurdles than would be required for a pilot study or live deployment. For example, researchers used simulation to assess how predictive admission and discharge policies affect patient flow in hospitals, and another research group modeled discharge decision-making to estimate the potential impact of a triage support tool.[6,7]

Simulation is also central to the design and use of digital twins, which have become increasingly popular in manufacturing, logistics, energy, and healthcare. Digital twins are high-fidelity virtual representations of real-world systems that are frequently updated to reflect the current state of their physical counterparts. This allows ML models to be evaluated in realistic operational environments, such as testing a predictive maintenance model within a factory twin or inserting a forecasting model into a simulated warehouse to assess effects on stock levels or delivery delays. These tools are widely used for robustness testing, scenario evaluation, and deployment planning, but typically only after a model has already been developed.

In many sectors, there are significant barriers to starting model development. In healthcare, working with patient-level data often requires extensive ethical approvals, formal data governance procedures, and secure computing environments. In such cases, before committing resources to these processes, it is valuable to understand whether a model is likely to deliver impact even if it performs well.

A SIMULATION-FIRST APPROACH

A simulation-first approach deploys the same tools used in post hoc evaluation but applies them at the start of the process, changing the focus from validating a built model to deciding whether to build one at all.

By injecting synthetic AI or ML model outputs (e.g., predictions, recommendations, generated content) of varying quality into a simulator that models the workflow, teams can explore how the model's performance translates into operational value. Would perfect foresight improve outcomes? If not, there may be little reason to invest further. But the case for development becomes stronger if a reasonably accurate model could generate impact.

This early insight can support project prioritization. There are always many competing projects for a limited data science team and multiple places in a workflow where AI could be useful. Simulation allows teams to compare these opportunities based on likely business value before committing to data collection or model development.

A simulation-first approach also encourages early, cross-functional collaboration. To simulate a system, data scientists must engage with domain experts to understand decisions, constraints, and KPIs before model deployment is considered. These conversations ensure that the model's performance is measured against the outcomes that really matter to the organization.

Crucially, a simulation-first approach enables experimentation not only with ML models themselves but also with the workflows in which they operate. For instance, a model may only deliver value if surrounding processes are adapted to act on its predictions — and if decision makers are willing to trust and use them. Simulation provides a safe environment to (1) test those adaptations without disrupting live operations and (2) understand whether taking full advantage of new technology will require bigger changes than simply replacing one part of an existing process.

This approach is intuitive and has been applied in a limited number of studies on supply chain forecasting and resource allocation.[8-12] In settings where development is costly due to regulatory approvals, privacy risks, or high labeling effort, a simulation-first strategy offers a low-risk way to focus resources where they will most likely deliver value.

CASE STUDY: REDUCING WASTE OF BLOOD PRODUCTS

Platelets (blood components essential for clotting) present a unique inventory challenge. With a shelf life of (at most) five days, hospitals must carefully balance stock levels to ensure they have enough on hand to meet unpredictable demand while avoiding waste due to expired units. At a large London teaching hospital, my research team observed that many platelet units were returned from wards unused after being requested by clinicians. The standard policy — issuing the oldest available unit — is optimal when all issued units are transfused. However, when units are returned, this practice often prevents them from being reissued before expiring.

This seemed like a good opportunity to use data to improve practice. If an ML model could predict which requests were likely to result in returns, it could support a new policy: issuing the oldest units when a transfusion is likely and the youngest when a return is expected. This approach would increase the chances that returned units remain usable.

However, building the model would require patient-level data, involving long approval processes, integration of data from multiple health systems, and a significant investment of analyst time.

We therefore began by building a simulator to model the workflow in the hospital blood bank, including placing a replenishment order in the morning, selecting a unit to meet each clinical request, and disposing of expired units at the end of each day. We then simulated predictions from models with various levels of performance and assessed how they would affect key outcomes like waste and service level.

Model performance was controlled by adjusting the assumed sensitivity and specificity of the predictions. These measure, respectively, how well the model identified the cases we wanted to flag and how well it avoided false alarms. Each ranges from 0% to 100%. By setting both values to 100%, we could test whether even perfect predictions would make a difference. By varying sensitivity and specificity, we explored how different levels of performance would translate into improvements.

The results showed that the model was worth building. A moderately accurate model would meaningfully reduce waste, assuming the prediction was acted on. We observed that there was a much larger improvement when the issuing policy was combined with optimized replenishment orders (how many units the blood bank should order from its supplier each day). This operational insight, showing how changes to multiple decision-making processes interact, would have been impossible to learn from predictive performance metrics alone.

The simulation results gave our team the confidence to proceed with model development, knowing that the time and effort required to secure and process clinical data would be worthwhile. Once the model was developed, the workflow simulator was used to help tune the hyperparameters of the ML model and estimate its real-world impact. We also explored how these benefits might vary across hospitals, finding that the model would be especially valuable in hospitals where a greater proportion of units are returned and where units tend to be older upon delivery. This demonstrates that a good simulator can support evaluation throughout the model development process.

Just as importantly, building the simulator prompted early engagement between stakeholders. Blood bank staff, clinicians, and data scientists worked together to define the decisions that mattered, the constraints that applied, and the metrics that should be used to determine success. This collaboration ensured that any model developed would be evaluated against practical criteria and shaped from the outset to fit the real-world context in which it would operate.

The simulation-first approach helped us identify whether ML would add value, how good the model would need to be to be "good enough," and whether it would be useful in a workflow not originally designed for ML. The resulting policy is being tested further — not just because a model was built but because the system around it was understood, challenged, and carefully modeled.[13]

BROADER APPLICATIONS

This article focuses on healthcare, but a simulation-first approach is equally applicable to situations in which success depends not only on model quality but on understanding how to integrate the model into a workflow or navigating barriers to model development.

In the energy sector, forecasting models are critical for balancing supply and demand, especially with increasing reliance on renewables. Deploying a new demand-forecast model involves more than improving predictive accuracy; it requires understanding how grid operators will act on those forecasts and how their decisions affect overall system stability, cost, and emissions. A simulation-first approach could help teams investigate these dynamics safely before committing to model development or infrastructure changes.

In manufacturing and logistics, a simulation-first approach could help teams assess whether predictive maintenance models or delay forecasts would meaningfully reduce downtime or improve service levels, especially when predictions must be embedded in tight production schedules or just-in-time inventory systems. Similarly, in public service delivery such as social care, emergency response, or transport planning, simulation-first could help teams assess whether better predictions about risk or demand would lead to improved outcomes or simply shift bottlenecks elsewhere.

CONCLUSION

In domains like digital advertising or product recommendation, A/B testing and rapid iteration make it straightforward to link model performance to business value. In contrast, that connection is much harder to establish in settings where interventions are high-stakes, data access is restricted, and broader system constraints limit how model outputs can be used. The simulation-first approach helps teams prioritize projects with the greatest potential for real-world impact by assessing a model's likely business value before it is built — grounding that evaluation in a realistic simulation of how decisions are actually made.

Standard performance metrics used to evaluate ML and AI models offer only a partial view of a model's usefulness. A simulation-first approach focuses evaluation on the KPIs that truly matter while encouraging early collaboration and exposing hidden constraints, ensuring that new models are accurate and impactful.

REFERENCES

[1] Jung, Kenneth, et al. "A Framework for Making Predictive Models Useful in Practice." *Journal of the American Medical Informatics Association (JAMIA)*, Vol. 28, No. 6, December 2020.

[2] Singh, Karandeep, Nigam H. Shah, and Andrew J. Vickers. "Assessing the Net Benefit of Machine Learning Models in the Presence of Resource Constraints." *JAMIA*, Vol. 30, No. 4, February 2023.

[3] Beede, Emma, et al. “A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy.” *CHI’20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery (ACM), 2020.

[4] Fildes, Robert, and Paul Goodwin. “Stability in the Inefficient Use of Forecasting Systems: A Case Study in a Supply Chain Company.” *International Journal of Forecasting*, Vol. 37, No. 2, April–June 2021.

[5] Meunier, Pierre-Yves, et al. “Barriers and Facilitators to the Use of Clinical Decision Support Systems in Primary Care: A Mixed-Methods Systematic Review.” *Annals of Family Medicine*, Vol. 21, No. 1, January 2023.

[6] Mišić, Velibor V., Kumar Rajaram, and Eilon Gabel. “A Simulation-Based Evaluation of Machine Learning Models for Clinical Decision Support: Application and Analysis Using Hospital Readmission.” *npj Digital Medicine*, Vol. 4, No. 98, June 2021.

[7] Wornow, Michael, et al. “APLUS: A Python Library for Usefulness Simulations of Machine Learning Models in Healthcare.” *Journal of Biomedical Informatics*, Vol. 139, March 2023.

[8] Dumkreiger, Gina. “Data Driven Personalized Management of Hospital Inventory of Perishable and Substitutable Blood Products.” Doctoral dissertation, Arizona State University, 2020.

[9] Fildes, R., and B. Kingsman. “Incorporating Demand Uncertainty and Forecast Error in Supply Chain Planning Models.” *Journal of the Operational Research Society*, Vol. 62, No. 3, 2011.

[10] Altendorfer, Klaus, Thomas Felberbauer, and Herbert Jodlbauer. “Effects of Forecast Errors on Optimal Utilisation in Aggregate Production Planning with Stochastic Customer Demand.” *International Journal of Production Research*, Vol. 54, No. 12, March 2016.

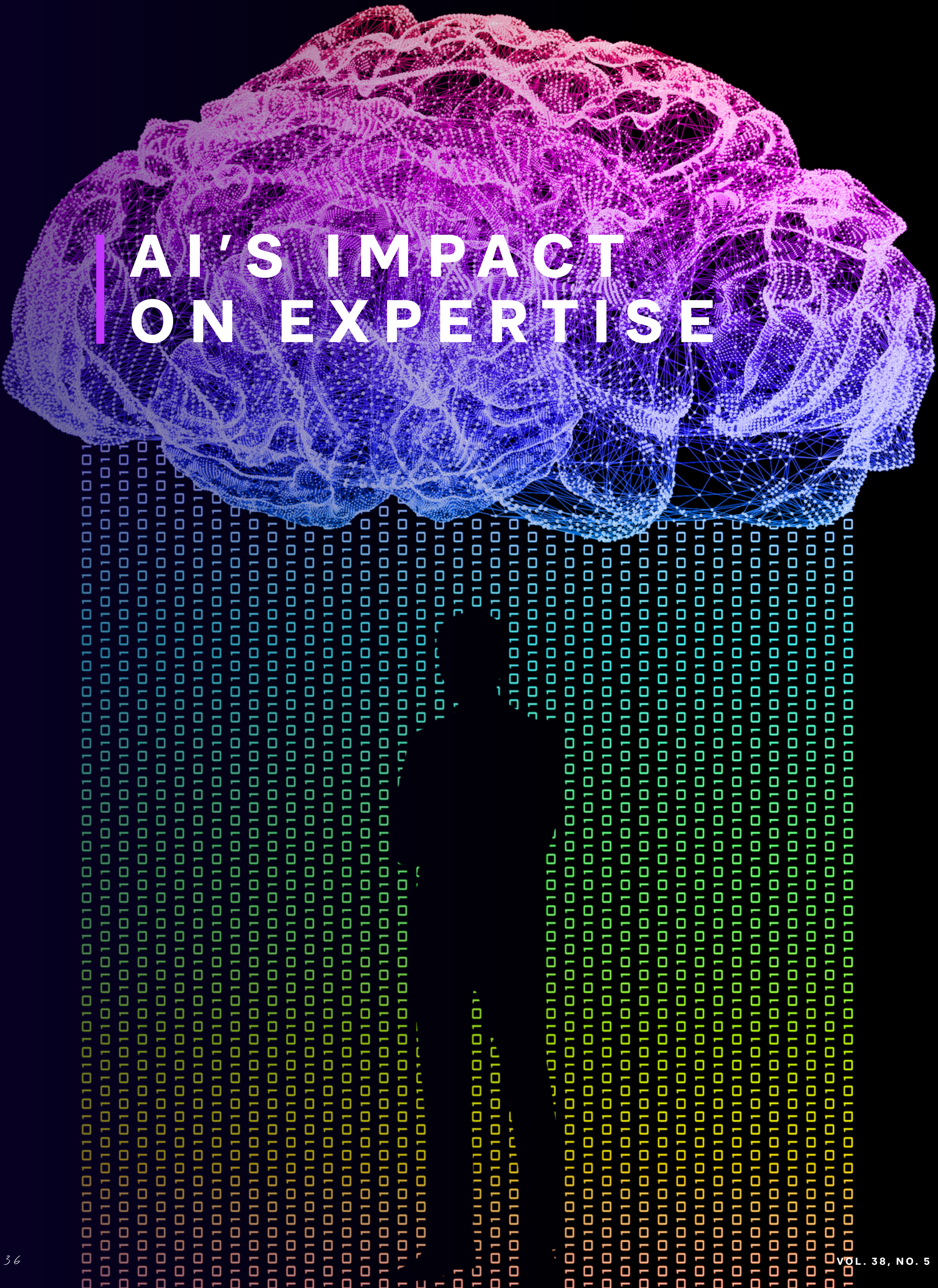
[11] Sanders, Nada R., and Gregory A. Graman. “Quantifying Costs of Forecast Errors: A Case Study of the Warehouse Environment.” Omega, Vol. 37, No. 1, February 2009.

[12] Doneda, Martina, et al. “Robust Personnel Rostering: How Accurate Should Absenteeism Predictions Be?” arXiv preprint, 26 June 2024.

[13] Farrington, Joseph, et al. “Many Happy Returns: Machine Learning to Support Platelet Issuing and Waste Reduction in Hospital Blood Banks.” arXiv preprint, 22 November 2024.

About the author

Joseph Farrington is a data scientist and chartered accountant. He recently earned a PhD in machine learning at University College London, UK, under the supervision of Ken Li, Wai Keong Wong, and Martin Utley. Dr. Farrington’s research was funded by UKRI training grant EP/S021612/1, the CDT in AI-enabled Healthcare Systems, and the NIHR University College London Hospital’s Biomedical Research Centre. He can be reached at joseph.farrington.18@alumni.ucl.ac.uk.



AI'S IMPACT ON EXPERTISE

Author

Joseph Byrum

Business leaders face a fundamental challenge in the AI era: identifying when emerging technologies will cross transformation thresholds that fundamentally reshape their markets. The genomics revolution provides a compelling preview. What once required decade-long agricultural innovation cycles now unfolds in 18 months, as AI systems analyze genomic patterns across vast combinatorial spaces. This compression of expertise development from careers to quarters creates what I call the "transformation threshold" challenge.

The core question is not whether AI will transform expertise. Rather, it's: "How can organizations systematically identify and prepare for the capability thresholds that trigger market transformation?" This article introduces a framework for recognizing these transformation thresholds and navigating the transition from expertise scarcity to abundance.

We're witnessing the commoditization of expertise itself — the metamorphosis of knowledge from a scarce resource jealously guarded by organizations into an abundant capability that AI can access, replicate, and scale with unprecedented speed. This represents a fundamental restructuring of how organizations generate and capture value from human knowledge.

TRANSFORMATION THRESHOLDS: THE CORE FRAMEWORK

Understanding when AI capabilities cross performance thresholds that trigger market transformation requires a systematic approach. Drawing from my work on the intelligent enterprise concept,[1] I propose that transformation thresholds manifest across three critical dimensions:

1. **Performance parity thresholds** — when AI capabilities match human expertise in measurable outcomes
2. **Economic viability thresholds** — when AI implementation costs fall below human expertise costs
3. **Adoption acceleration thresholds** — when organizational resistance to AI implementation collapses

> **WE'RE WITNESSING THE COMMODITIZATION OF EXPERTISE ITSELF**

These three thresholds rarely align temporally, creating complex transitions that challenge traditional strategic planning. The intelligent enterprise framework, which uses the Adaptive Response Framework (observe, orient, decide, act [OODA]), provides a methodology for continuously monitoring these threshold approaches.

THE GREAT ACCELERATION

To comprehend the velocity of this shift, consider the trajectory of agricultural knowledge. For 10,000 years, farming expertise passed from generation to generation through oral tradition and apprenticeship. The mechanization of agriculture unfolded across 150 years, during which workforce participation in farming declined from approximately 70% in 1840 to less than 2% in developed nations today.[2] What previous generations achieved through centuries of gradual progress, today's AI systems accomplish in months.

This compression of innovation timelines tells a compelling story. Operations research (the discipline of optimizing complex decisions through mathematical analysis) offers a revealing precedent. Born from wartime necessity, it evolved over seven decades from the exclusive province of PhD mathematicians into software any competent manager can deploy. The Franz Edelman Award winners alone generated US $250 billion in savings by encoding expert judgment into replicable algorithms.[3] That figure represents a quarter-trillion dollars' worth of value created by transforming scarce human expertise into abundant computational capability.

AI achieves comparable transformations in much less time. The Human Genome Project consumed $2.7 billion and 13 years to sequence the first human genome (1990 to 2003). Next-generation sequencing accomplishes the same task for less than $1,000 in under 24 hours.[4]

Between 2018 and 2023, language models progressed from simple text completion to demonstrating complex reasoning capabilities, with each iteration exhibiting emergent properties that surprised even their creators.[5] Academic research from multiple institutions confirms these rapid capability improvements, though the true reasoning capabilities of current AI systems remain subject to debate.

This acceleration fundamentally alters organizational transformation dynamics. Economist Paul David documented how electric motors, despite installation in the 1890s, did not yield meaningful productivity gains until the 1920s (after factories reimagined their entire operational architecture around distributed rather than centralized power).[6] Of course, unlike industrialists who enjoyed a generation to adapt, today's executives face expertise disruption cycles measured in months, not decades.

THE SYMBIOSIS IMPERATIVE

Conventional analyses of AI transformation stumble because they frame the question as human versus machine, replacement rather than recombination. Nature offers more sophisticated models. Consider the peculiar partnership between zebras and ostriches on the African savanna. The ostrich possesses exceptional eyesight but poor hearing and smell. The zebra's sensory profile is precisely opposite: acute hearing and smell but mediocre vision. Together, they form a defensive system superior to what either could achieve alone.

This biological principle (mutualism) provides the blueprint for human-AI collaboration in the intelligent enterprise. As I have argued in previous work on this concept, the intelligent enterprise integrates AI throughout organizations to augment human capabilities rather than replace them. Machines excel at processing vast datasets with unwavering precision. Medical AI systems can process every journal article ever published, flagging obscure symptoms mentioned in foreign-language footnotes that might unlock a diagnosis, a capability that transforms how we think about medical expertise distribution. AI has achieved parity-level accuracy in medical imaging, matching board-certified radiologists. This doesn't eliminate the need for human doctors; it transforms their role from image analysis to complex decision-making.[7]

However, the success of human-AI collaboration is far more complex than optimistic projections suggest. Research demonstrates that human-in-the-loop systems can actually reduce AI performance compared to full automation, depending on the specific task, human operators involved, and implementation context. This complexity requires structured decision-making frameworks rather than assumptions about synergy.

Machines remain remarkably inept at capabilities humans consider trivial. They cannot read the subtle contextual cues that experienced professionals detect, such as the way a patient describes pain that suggests psychological rather than physical origins, the almost imperceptible tension in a negotiation that signals a deal is about to collapse, or the behavioral patterns of team members that indicate brewing conflict.

This tacit knowledge, which polymath Michael Polanyi estimated comprises 70%-80% of organizational knowledge, resists codification because it emerges from lived experience rather than explicit rules.[8,9]

THE OODA LOOP AS A TRANSFORMATION ENGINE

Military strategists have long understood how to operate in environments of extreme uncertainty. Their OODA loop framework offers surprising insights into how expertise commoditization unfolds in practice. Originally developed for fighter pilots making split-second decisions, the framework now illuminates how organizations can navigate the turbulent waters of AI transformation.

In the **observe** phase, AI systems capture and process volumes of data that would overwhelm human analysts. But raw observation without interpretation is merely noise. The **orient** phase (in which patterns are recognized and theories formulated) represents the first level of expertise commoditization. AI systems can now generate multiple strategic scenarios, each backed by statistical analysis of probable outcomes. What once required teams of strategists working for weeks can be produced in minutes.

The **decide** phase remains fundamentally human. Choosing between AI-generated options requires understanding contextual factors that often exist in unstructured forms: organizational culture, stakeholder relationships, and long-term vision. Modern AI increasingly processes unstructured text, but the challenge lies in capturing experiential knowledge that rarely gets documented, rather than quantification. Critically, the **act** phase creates new realities that no algorithm could fully anticipate. Each decision changes the environment in ways that require human judgment to interpret and manage.

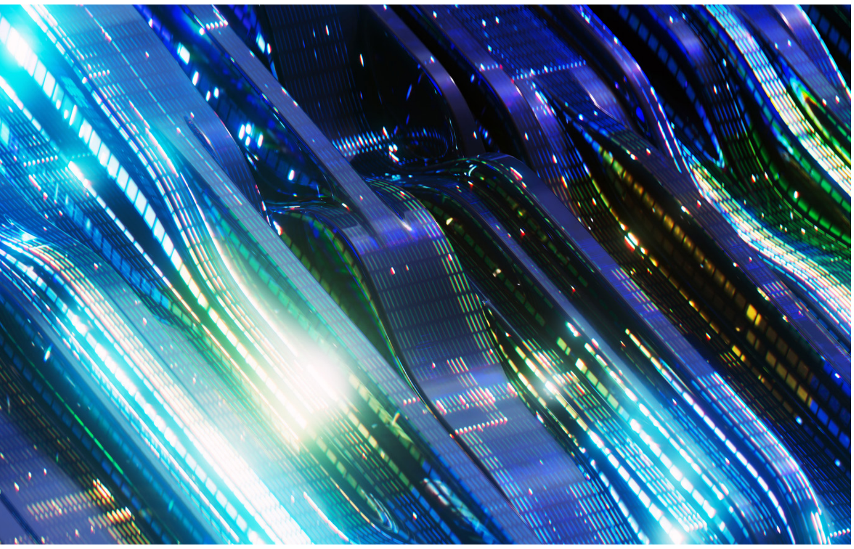
This framework sheds light on why simple automation fails while human-AI collaboration succeeds. JPMorgan's Contract Intelligence (COIN) system initially promised to eliminate legal work by reviewing commercial loan agreements in seconds rather than the 360,000 hours annually consumed by human lawyers. The system achieved 99% accuracy in routine term extraction.[10] But the real transformation came when lawyers, freed from document review, redirected their expertise toward complex deal structuring and relationship management. The commoditization of routine analysis elevated, rather than eliminated, human work.

THE UPS REVELATION

Perhaps no example better illustrates the messy reality of expertise transformation than UPS's On-Road Integrated Optimization and Navigation (ORION) system. This case provides critical insights into transformation thresholds and the complexity of human-AI collaboration. The algorithm could calculate optimal delivery routes across millions of variables, promising significant efficiency gains.

Initial results proved catastrophic, demonstrating why transformation thresholds involve more than technical capability. Driver compliance languished below 30% as veterans with 15-20 years of route knowledge rebelled against mathematically optimal paths that ignored human reality.

The breakthrough came from recognizing that driver expertise contained irreplaceable value — precisely the kind of tacit knowledge that creates implementation challenges. Drivers knew which customers had aggressive dogs, where construction delays were likely, and which dock managers insisted on specific delivery windows despite official policies. Rather than mandate compliance, UPS developed “experiential algorithms” that learned from driver deviations.



This transformation required three distinct implementation phases:

1. **ORION 1.0 (2008–2010)** — technology-centered approach that failed due to driver resistance
2. **ORION 2.0 (2011–2013)** — process-centered approach that still faced significant resistance
3. **ORION 3.0 (2013–2016)** — people-centered approach that achieved success through human-AI integration

Academic analysis decomposed the performance improvements into three components: pure algorithmic optimization achieved 5%-8% reduction in miles driven; improved driver compliance added 7%-10% efficiency gains, and human-AI synergy (bidirectional learning between drivers and algorithms) contributed an additional 5%-7% improvement.[11] The combined system achieved 17%-25% total improvement, but this success came only after recognizing that human-AI collaboration requires careful design. The initial phases demonstrated how human-in-the-loop systems can underperform when implementation ignores human factors.

THE ECONOMICS OF TRANSFORMATION

Understanding when expertise shifts from scarcity to abundance requires careful attention to economic thresholds. Goldman Sachs estimates that current enterprise AI implementation costs range from $50,000 to $500,000 for initial deployment.[12] But these figures tell only part of the story.

Independent research from academic institutions reveals that adaptation costs typically run two to three times the technology investment, encompassing workforce retraining, process redesign, change management initiatives, and productivity dips during transition.[13] Organizations currently allocate 2%-5% of revenue to AI initiatives, reaching 10% in technology-intensive sectors.[14]

Critical performance thresholds determine economic viability. The 95% accuracy threshold frequently cited in AI adoption represents not arbitrary performance targets but moments when algorithmic consistency begins to surpass human variability in economically meaningful ways. This threshold often falls slightly below peak human performance because it represents the point where AI’s consistency advantages offset human performance peaks, creating net economic value despite not exceeding the best human practitioners.

The cost dynamics follow predictable patterns. Research from academic institutions tracking AI development costs demonstrates how rapidly capabilities democratize: what cost thousands of dollars per million tokens in early models now costs $0.01 per 1,000 tokens, making AI analysis more economical than human review for many tasks.[15]

CATEGORIES OF TRANSFORMATION

Through systematic analysis of transformation thresholds, four distinct patterns of expertise evolution emerge:

1. **Commoditized capabilities represent expertise where AI has definitively crossed performance thresholds.** Basic legal document review, routine medical imaging, and standard financial analysis increasingly fall into this category. These domains share characteristics: rule-based

processes, objectively measurable outcomes, standardized procedures, and abundant training data. The strategic imperative involves systematic transition planning rather than swift action alone. Organizations must develop internal capabilities to capture value from commoditized expertise while avoiding premium costs for capabilities competitors access at commodity rates.

2. **Augmentation opportunities encompass domains where human-AI collaboration multiplies effectiveness.** Complex medical diagnosis exemplifies this category; AI processes vast research libraries while physicians provide contextual interpretation and patient relationship management. Success requires designing interfaces that maximize both computational power and human insight. The goal isn't replacing human judgment but amplifying it through algorithmic support.

3. **Transformation candidates include expertise requiring fundamental reconceptualization to remain relevant.** Project management illustrates this evolution: traditional scheduling expertise becomes less valuable while orchestrating human-AI teams grows critical. Financial analysis shifts from spreadsheet manipulation to interpreting AI-generated scenarios. These capabilities don't disappear; they morph into forms that previous practitioners might not recognize.

4. **Resilient differentiators comprise capabilities where human judgment, creativity, and relationship building create value that resists commoditization.** Complex negotiations, cultural leadership, and strategic vision exemplify domains where success depends on trust, ambiguity navigation, and contextual understanding emerging from lived experience. Yet even these must evolve: yesterday's differentiator becomes tomorrow's commodity as AI capabilities expand.

LEARNING FROM CORPORATE MORTALITY

The consequences of misreading expertise transformation are severe. Among Fortune 500 companies from 60 years ago, nine out of 10 have disappeared through bankruptcy, merger, or irrelevance. Kodak's century of photographic expertise became worthless when digital cameras commoditized image capture. Blockbuster's retail expertise became useless when streaming commoditized content delivery. These weren't failures of execution but fundamental misunderstandings of how expertise commoditization restructures entire industries.

The pattern repeats with disturbing regularity. Tower Records dominated music retail through deep genre expertise and curated selections. When digital distribution commoditized access to music, the company's expertise became a liability rather than an asset. Borders Books invested heavily in retail expertise while Amazon commoditized book distribution. These examples illustrate the challenge of identifying transformation thresholds before they reshape competitive dynamics. The organizations that succeeded were those that recognized threshold approaches early and repositioned their expertise portfolios accordingly.

THE PATH FORWARD

Organizations navigating transformation thresholds must embrace four strategic imperatives:

1. **Develop threshold monitoring systems.** Traditional expertise developed over careers; commoditized expertise evolves over quarters. Organizations need systematic approaches to identifying when AI capabilities approach performance, economic, and adoption thresholds in their specific domains. This requires continuous capability assessment rather than periodic strategic planning.

2. **Design for symbiosis.** Stop asking whether AI will replace specific roles. Instead, reimagine how humans and AI can combine to create capabilities neither possesses alone. UPS's experience demonstrates that the highest returns come from bidirectional learning systems in which humans and algorithms continuously improve each other — but only when implementation addresses human factors rather than assuming automatic collaboration.

3. **Embrace strategic ambiguity.** In environments of rapid expertise commoditization, maintaining flexibility matters more than perfecting plans. Organizations need what I call "adaptive sensing" in my intelligent enterprise framework: the ability to recognize when capabilities approach transformation thresholds and pivot accordingly. This requires cultural comfort with uncertainty and a willingness to abandon successful strategies before they become obsolete.

4. **Cultivate cognitive diversity.** As AI commoditizes analytical capabilities, uniquely human perspectives become more valuable. But diversity alone isn't sufficient. Teams need frameworks that allow professionals from different backgrounds to collaborate effectively in guiding AI systems. The most successful organizations combine cognitive diversity with operational coherence.

FRAMEWORK IMPLEMENTATION: TRANSFORMATION THRESHOLD MATRIX

Building on the intelligent enterprise approach, I propose the Transformation Threshold Matrix as a practical tool for identifying and navigating expertise transformation. This framework systematically monitors three threshold dimensions:

1. **Technical capability monitoring** — tracking AI performance against domain-specific benchmarks
2. **Economic viability assessment** — monitoring cost trajectories and implementation economics
3. **Organizational-readiness evaluation** — assessing internal capacity for expertise transition

Organizations can apply this matrix by:

- Mapping current expertise portfolios against threshold proximity
- Developing trigger-based transition strategies
- Creating cross-functional threshold monitoring teams
- Implementing continuous capability reassessment protocols

This systematic approach helps organizations move beyond reactive responses to proactive threshold management.

CONCLUSION: MASTERING TRANSFORMATION THRESHOLDS

The expertise revolution is fundamentally about recognizing and navigating transformation thresholds. Organizations that master threshold identification and strategic transition will define the next era of competitive advantage.

JPMorgan didn't eliminate lawyers when COIN automated document review; it elevated them to higher-value work. UPS didn't replace drivers when ORION optimized routes; it enhanced their capabilities through algorithmic partnership — but only after recognizing that human-AI collaboration relies on systematic design, not automatic synergy. In each case, the commoditization of routine expertise created space for distinctly human contributions: relationship building, creative problem-solving, and navigating ambiguity.

The intelligent enterprise of the future won't be one where machines replace humans; it will be where human judgment finds its highest expression, guided by systematic threshold monitoring toward decisions no algorithm could make alone. In this new landscape, competitive advantage won't flow from hoarding scarce expertise but from orchestrating abundant intelligence (both human and artificial) in combinations that continuously evolve.

The question facing every organization is not whether transformation thresholds will reshape their industry; that outcome is mathematically inevitable given current trajectories. The question is whether they'll develop an enduring capability to identify these thresholds before competitors and lead the transformation rather than react to it.

Success will belong to organizations that are wise enough to recognize that in an age of abundant AI, human expertise becomes more valuable, not less, but only when guided by frameworks that help us identify what expertise means in the first place.

REFERENCES

[1] Byrum, Joseph. "Progress Toward the Intelligent Enterprise." *MIT Sloan Management Review*, 11 February 2021.

[2] Autor, David H. "Why Are There Still So Many Jobs? The History and Future of Workplace Automation." *Journal of Economic Perspectives*, Vol. 29, No. 3, Summer 2015.

[3] "Franz Edelman Award for Achievement in Advanced Analytics, Operations Research, and Management Science." INFORMS, accessed 2025.

[4] "The Cost of Sequencing a Human Genome." National Human Genome Research Institute, accessed 2025.

[5] Liang, Percy, et al. "Holistic Evaluation of Language Models." arXiv preprint, 1 October 2023.

[6] David, Paul A. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *American Economic Review*, Vol. 80, No. 2, May 1990.

[7] Esteva, Andre, et al. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature*, Vol. 542, January 2017.

[8] Polanyi, Michael. *The Tacit Dimension*. Routledge & Kegan Paul, 1967.

[9] Leonard, Dorothy, and Syliva Sensiper. "The Role of Tacit Knowledge in Group Innovation." *California Management Review*, Vol. 40, No. 3, April 1998.

[10] Son, Hugh. "JPMorgan Software Does in Seconds What Took Lawyers 360,000 Hours." Bloomberg, 28 February 2017.

[11] Toth, Paolo, and Daniele Vigo (eds.). *Vehicle Routing: Problems, Methods, and Applications, Second Edition*. Society for Industrial and Applied Mathematics (SIAM), 2014.

[12] Goldman Sachs. "Gen AI: Too Much Spend, Too Little Benefit?" *Top of Mind*, No. 129, 25 June 2024.

[13] "The True Cost of Artificial Intelligence: Beyond the Hype." Pure Storage, 2 April 2025.

[14] "Worldwide AI and Generative AI Spending Guide." IDC, 2025.

[15] "Artificial Intelligence Index Report 2024." Stanford Institute for Human-Centered AI (HAI), 2024.

About the author

Joseph Byrum is an accomplished executive leader, innovator, and cross-domain strategist with a proven track record of success across multiple industries. With a diverse background spanning biotech, finance, and data science, he has earned over 50 patents that have collectively generated more than US $1 billion in revenue. Dr. Byrum's groundbreaking contributions have been recognized with prestigious honors, including the INFORMS Franz Edelman Prize and the ANA Genius Award. His vision of the "intelligent enterprise" blends his scientific expertise with business acumen to help Fortune 500 companies transform their operations through his signature approach: "Unlearn, Transform, Reinvent." Dr. Byrum earned a PhD in genetics from Iowa State University, USA, and an MBA from the Stephen M. Ross School of Business, University of Michigan, USA. He can be reached at www.josephbyrum.com.

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

CUTTER
AN ARTHUR D. LITTLE COMMUNITY

For more content,
visit www.cutter.com