



Everybody's System, Nobody's Risk

Accountability Gaps in Distributed AI Governance

About this paper

This paper is for boards and for the executives who answer to them. It asks a narrow question with wide consequences. When an AI-supported decision goes wrong inside an organisation, and the system behind it was built by one company on a model trained by another, used by a professional who did not configure it and overseen by several regulators with partial remits, who owns the outcome? My answer is that, unless the organisation has decided in advance, nobody does, and the paper sets out why that happens and what a board can do about it now.

It is the fourth of six Ecaveo working papers on the higher-order issues in AI adoption. The strand's single argument is that AI adoption is a changing sociotechnical configuration whose governance must keep pace with it. Here the configuration in question is the distribution of responsibility itself, across suppliers, deployers, users and regulators, at the moment the United Kingdom has chosen a sector-led model of regulation that spreads oversight across existing bodies. Where the paper summarises a white paper, a regulator's statement or a regulator's order, it reports what the source says and how firmly it says it. It does not state what the law is. Margin notes headed FOR COUNSEL mark points where an inference touches an organisation's own legal exposure and where a qualified adviser should confirm the position.

Editorial balance. Roughly two fifths of the paper reports published evidence and policy, drawn from moral philosophy, the sociology of automation, the accountability literature in computing and the UK government's March 2023 white paper. About a third is my own analysis of how those ideas apply to AI supply chains and to staff use of public tools, and the remainder, concentrated in the last two chapters, is a working framework of my own that is offered for testing and has not been validated.

HOW TO READ THE EVIDENCE TAGS

Every substantive claim carries a tag stating what kind of evidence supports it. Two tags are combined where a claim rests on both.

PEER-REVIEWED A journal article or peer-reviewed conference paper.

PREPRINT A working paper, preprint or technical report not yet peer reviewed.

STANDARD A published standard or framework, such as ISO or NIST.

POLICY A government, regulator or court document or action.

INDUSTRY A company announcement, press report, analyst or consultancy report.

SCHOLARLY A scholarly book, edited volume or law journal article.

PRACTICE Practitioner observation, including the author's own.

EVIDENCE GAP The available evidence does not answer the question.

ECAVEO The author's own framework or analysis, proposed rather than proven.

Legal and regulatory notice. Parts of this paper describe legislation, regulation, regulator activity or court decisions. Nothing in this document constitutes legal advice. The provision of legal advice sits outside the terms of any engagement with the author or with Ecaveo, and the material is presented to support discussion and further review by qualified advisers. Statements about what a source says are attributed to that source. Where I draw a practical implication that a source does not state, the text says so.

CONTENTS

Contents

FRONT MATTER

A note from the author	4
Executive summary	5

PART ONE · HOW RESPONSIBILITY DISSOLVES

1	An error with five owners	8
2	Responsibility gaps in theory	10
3	Moral crumple zones	12

PART TWO · THE SEAMS IN THE SYSTEM

4	The sector-led model and its seams	15
5	Supply chains and shadow use	17

PART THREE · CLOSING THE GAP

6	Closing the gap inside the organisation	21
7	What to ask of suppliers, and what remains open	24

END MATTER

References	26
About the author	27

ILLUSTRATIVE MATERIAL

Every organisation, person and scenario described as illustrative in this paper is invented. Published research and official documents are cited and numbered, and invented material is never presented as evidence.

Research closed 26 April 2023. Sources and events are as they stood at that date.

CITATION

Forrest, P. (2023) *Everybody's System, Nobody's Risk: Accountability Gaps in Distributed AI Governance*. Ecaveo Working Papers, AI Adoption: The Higher-Order Issues, Paper 4. Ecaveo, 28 April 2023. Available at: ecaveo.com.

Comments and corrections are welcome and will be acknowledged in later versions.

PAPERS IN THIS STRAND

- Paper one. *Knowledge With an Expiry Date*. Evidence, Evaluation and Decision-Making When the AI System Keeps Changing. 14 December 2022.
- Paper two. *The Competence Paradox*. When AI Makes the Team Better and the Expert Worse. 21 February 2023.
- Paper three. *The Unit of Adoption*. Why Organisations Adopt Configurations, Not Models. 31 March 2023.
- **Paper four. *Everybody's System, Nobody's Risk: Accountability Gaps in Distributed AI Governance*. 28 April 2023.**
- Paper five. *Approved Once, Changed Daily*. From Point-in-Time Ethics to Continuous AI Assurance. 31 May 2023.
- Paper six. *The Cost of Waiting*. Precaution, Opportunity and the Ethics of Not Adopting AI. 29 June 2023.

A NOTE FROM THE AUTHOR

Why the question kept coming back

The question at the centre of this paper arrived in meetings. In advisory work over the past year I have sat in several conversations about an AI tool that had produced something unwelcome, and I noticed that each one followed the same course. Someone asked who had approved the tool. Someone else pointed out that procurement had approved the supplier, which is a different thing. The supplier's terms were read aloud, and they said, in careful language, that outputs should be reviewed by a qualified person. The qualified person, when asked, said that the tool had been presented as reliable. By the end of the hour everyone in the room had behaved reasonably, and the outcome still belonged to nobody.

I do not think any of those people were evasive. Each held a real piece of the problem and was honest about the edges of that piece. What was missing was a person whose job it was to hold the whole thing, and an organisation that had decided, before anything went wrong, who that person would be.

Coaching taught me something similar long before AI entered the picture. When a squad underperforms, the athletes, the coaches, the physiotherapist and the selectors can each give an accurate account of their own contribution, and the accounts can all be true while the season still fails. Good programmes name someone who answers for the whole.

The publication of the UK's white paper in March, and the run of events since, made the question feel less abstract. Regulators are being asked to cover AI within their existing remits, and the government says openly that risks may fall between them. That is a reasonable design choice for a national framework. It also means that organisations cannot outsource the job of joining things up. This paper is my attempt to say, as plainly as I can, what joining things up would involve.

Paul Forrest
Founder, Ecaveo
28 April 2023

TWO PAGES THAT CIRCULATE ON THEIR OWN

Executive summary

AI-supported decisions are now made by chains of organisations and people, each of whom owns a slice of the risk and none of whom owns the whole, so that when a decision fails everyone can point to someone else; organisations should close that gap themselves by naming one accountable owner for every AI-supported decision, whatever the supply chain looks like. The problem is old. Political theorists described the difficulty of assigning responsibility when many hands contribute to an outcome more than 40 years ago,¹ PEER-REVIEWED and computing ethicists carried the idea into software in the 1990s.² PEER-REVIEWED What is new is the combination of learning systems whose behaviour their makers cannot fully predict, supply chains built on shared foundation models, informal staff use of public tools, and a UK regulatory model that deliberately spreads oversight across existing regulators. The paper reads the March 2023 white paper closely and proposes an internal discipline that a board can adopt without waiting for regulators.

1

Responsibility for AI outcomes fragments predictably, and no single party needs to behave badly for it to happen.

The problem of many hands describes how responsibility dissolves when many people contribute to a result.¹ **PEER-REVIEWED** Learning systems add a further gap, because their makers cannot fully foresee what they will do,³ **PEER-REVIEWED** and later work separates at least four distinct gaps that call for different remedies.⁴ **PEER-REVIEWED**

3

The UK's sector-led model openly accepts that some AI risks may fall between regulators.

The white paper asks existing regulators to apply five principles, proposes central functions to monitor gaps, and states that it is not proposing to change how accountability is allocated across the AI life cycle at this stage.⁶ **POLICY** That leaves the allocation inside supply chains largely to contracts and to existing law.

5

Staff use of public generative tools creates accountability gaps that sit wholly outside formal governance.

Trade press in early April 2023 reported that engineers at a large electronics manufacturer had entered source code and meeting notes into ChatGPT.⁹ **INDUSTRY** The Information Commissioner's Office has set out questions that organisations using generative AI should ask, including whether they act as controller, joint controller or processor.¹⁰ **POLICY**

2

When responsibility is unclear, blame tends to settle on the human nearest the failure.

Elish's analysis of accidents in automated systems shows responsibility being attributed to operators who had limited control, a pattern she calls the moral crumple zone.⁵ **PEER-REVIEWED** In my experience the reviewing professional in an office workflow is at risk of playing the same part. **PRACTICE**

4

Supply chains built on foundation models separate the party that controls a model from the party that answers for its use.

A defect in a widely used foundation model may be inherited by the many applications adapted from it.⁷ **PREPRINT** The NIST framework notes that actors responsible for one part of the AI life cycle often lack visibility or control over other parts.⁸ **STANDARD**

6

A single named owner for each AI-supported decision, backed by a responsibility map and internal audit, closes most of the gap an organisation controls.

End-to-end internal audit frameworks already exist and can be adapted for this purpose.¹¹ **PEER-REVIEWED** The owner model proposed here is my own and has not yet been tested at scale. **ECAVEO**

What follows for a board

A board should ask for an inventory of decisions in which AI plays a part, including those supported by tools staff use informally, and should require a named individual who answers for each outcome. It should then require a map showing, for each of those decisions, what every party in the supply chain controls, what it can see and what the contract says it will do when something changes or goes wrong.

None of this depends on how the government's consultation, which closes on 21 June 2023, turns out.⁶ **POLICY** An organisation that has named its owners will find any later rules easier to meet, and it will have fewer meetings in which everyone is right and nobody is responsible.



PART ONE

How responsibility dissolves

The theory of distributed responsibility and why blame settles where it does.

CHAPTER ONE

An error with five owners

When an AI-supported decision goes wrong, each party can give an accurate account of its own role, and the accounts together still leave the outcome unowned.

The spring of 2023 has made AI governance a boardroom topic in a way that the previous five years did not. Within about ten days, a widely publicised open letter called on AI laboratories to pause for at least six months the training of systems more powerful than GPT-4,¹² **INDUSTRY** the UK government published its white paper on AI regulation,⁶ **POLICY** and Italy's data protection authority imposed a temporary limitation on the processing of Italian users' data by the company behind ChatGPT.¹³ **POLICY** Most of the public argument that followed concerned the capabilities of the systems themselves. This paper is concerned with something more mundane and, for most organisations, more immediate, namely who answers for an outcome when an AI system contributes to a decision.

1.1 A decision that went wrong

The scenario below is invented. It is composed from patterns I have seen in advisory work, and no organisation or person in it is real.

ILLUSTRATIVE SCENARIO

A shortlist nobody made

A professional services firm with about 1,400 staff introduces a screening assistant to help its recruitment team sort applications. The assistant is sold by a software company that built it on a general-purpose language model licensed from a larger developer. The firm's procurement team checked the supplier's security certifications and signed its standard terms, which describe outputs as recommendations to be reviewed by a qualified person. A recruitment manager, under pressure to fill 30 roles, accepts the assistant's shortlists with light review. Months later, an unsuccessful applicant with a disability complains that the assistant marked down a gap in her employment history. The supplier says its terms require human review. The software company says the underlying model was updated by its developer without notice and that it had no way to test the change. The recruitment manager says the tool was approved by the firm. The firm's data protection officer says the process belongs to human resources, and human resources says the tool belongs to technology. The complaint could engage both the equality regulator and the data protection regulator, and each looks at a different part of the story.

Nobody in the scenario lied, and nobody was careless in a way a disciplinary panel would recognise. The procurement check was real, the terms were lawful on their face (so far as a business reader can tell), and the recruitment manager did what the or-

ganisation's process allowed. The failure sits in the joins between those actions. That is the pattern this paper examines, and I think it is common enough that most organisations using AI at any scale will meet it within a year or two. **PRACTICE**

1.2 Why the timing matters

Two developments make the question pressing now. The first is volume. Stanford's AI Index reports that the number of newly recorded AI incidents and controversies in the AIAAIC database was 26 times greater in 2021 than in 2012, reaching 260 in 2021, while noting that better tracking may account for part of the rise.¹⁴ **INDUSTRY** More use means more occasions on which something goes wrong and someone has to answer for it.

The second development is structural. On 24 April 2023 the government announced an initial £100m for a taskforce on foundation models,¹⁵ **POLICY** which signals that the models at the base of many supply chains are now a matter of national policy. At the same time the white paper declines, for now, to reallocate responsibility along those supply chains.⁶ **POLICY** The combination leaves organisations using AI in an awkward position, encouraged to adopt and not yet told precisely what they will answer for. My own view is that this is the moment to settle the question internally, before a complaint or an incident settles it for them. **PRACTICE**

CHAPTER TWO

Responsibility gaps in theory

Philosophers and political theorists have described, for more than 40 years, how responsibility dissolves when many hands and learning machines contribute to an outcome.

The idea that responsibility can leak away in complex systems did not begin with AI. It began with the study of bureaucracy, passed into computing ethics in the 1990s, and has been refined for learning systems over the two decades since, largely by philosophers of technology. The literature is worth reading closely, because it separates problems that boards tend to run together.

2.1 The problem of many hands

Thompson introduced the phrase in a study of public officials.¹ **PEER-REVIEWED** Because many officials contribute in many ways to the decisions and policies of government, he argued, it becomes difficult even in principle to say who is morally responsible for an outcome. He examined two tempting answers, one that places responsibility at the top of the hierarchy and one that assigns it to the collective, found both wanting, and defended an account that keeps moral responsibility with individual officials despite the complexity of the organisations they work in.

Nissenbaum carried the analysis into computing and identified four barriers to accountability in a computerised society.² **PEER-REVIEWED** The many hands of software development come first. Bugs, which practitioners come to treat as an unavoidable feature of complex code, form the second barrier. The third is the habit of blaming the computer itself, and the fourth is the tendency of software producers to claim the rights of ownership while disclaiming liability for what their products do. Most of the four can be traced in the recruitment scenario above, which suggests that the analysis has aged better than most writing about technology from 1996.

2.2 Learning systems and the gap Matthias described

Matthias added a problem that the many-hands literature does not cover.³ **PEER-REVIEWED** Traditional ascription of responsibility, he argued, assumes that the manufacturer or operator could in principle foresee and control what a machine does. A system that learns and adapts during operation breaks that assumption, because its behaviour depends on its training and its environment in ways its designers cannot fully predict. The result is a gap in which nobody appears to have enough control to be blamed, even though harm has occurred.

The UK white paper uses closely related language. It identifies adaptivity and autonomy as the characteristics that make AI hard to regulate and says that their combination can make outputs difficult to explain, predict or control.⁶ **POLICY** A government document published in 2023 and a philosophy paper published in 2004 thus describe essentially the same difficulty, which should reassure no one.

2.3 Several gaps with different remedies

Santoni de Sio and Mecacci argue that the responsibility gap is a set of at least four connected problems.⁴ **PEER-REVIEWED** They distinguish gaps in culpability, in moral accountability, in public accountability and in active responsibility, and they trace these to sources that are organisational, legal, ethical and societal as well as technical. They also note that such gaps can arise with systems that do not learn at all.

This distinction matters for boards because it shows that different gaps call for different remedies. A culpability gap concerns who deserves blame after the event. An active responsibility gap concerns who is charged, in advance, with making sure the system behaves well, and it is the one an organisation can most directly close.

Bovens offers a working definition that makes the point concrete. Accountability, in his account, is a relationship between an actor and a forum, in which the actor is obliged to explain and justify its conduct, the forum can ask questions and pass judgement, and the actor may face consequences.¹⁶ **PEER-REVIEWED** Cooper and colleagues revisit Nissenbaum's four barriers in the light of machine learning, bring them together with this relational view of accountability, and discuss how the barriers make such a framework hard to put into practice for data-driven systems.¹⁷ **PEER-REVIEWED** Read together, these sources suggest a practical test. For each AI-supported decision, can the organisation name the actor who will explain it, the forum to which that actor answers, and the consequence that follows a bad outcome? Where any of those is missing, the gap is open. **ECAVEO**

CHAPTER THREE

Moral crumple zones

When a complex system fails and responsibility is unclear, blame tends to settle on the person closest to the failure, whatever their real degree of control.

Theory explains why responsibility becomes hard to assign. It says less about where responsibility goes when an assignment is forced, as it usually is after a public failure. Elish's work answers that question, and the answer is uncomfortable for anyone designing a process with a human reviewer at the end of it.

3.1 Where blame lands

Elish analyses several accidents involving highly automated systems, including the partial meltdown at Three Mile Island, the loss of Air France Flight 447 and the 2018 death of a pedestrian struck by a self-driving test vehicle in Arizona, together with the media coverage that followed.⁵ **PEER-REVIEWED** She introduces the moral crumple zone to describe how responsibility can be misattributed to a human actor who had limited control over the behaviour of an automated system. A car's crumple zone protects the driver. The moral crumple zone, on her analysis, shields the reputation of the technology and leaves the human operator nearest the failure to absorb the blame.

Her cases come from high-hazard industries, where the operator is a trained professional and the investigation is formal. The mechanism needs only the combination of distributed control and a public demand for someone to blame, and both are present whenever an AI-supported decision affects a customer or an applicant.

3.2 The reviewing professional as a crumple zone

Many deployments of generative and predictive tools in 2023 place a human reviewer between the system and the decision. Supplier terms often require it. The reviewer's presence is sometimes the main reason a deployment was approved at all, because it reassures a risk committee that a person makes the final call. In my experience the reviewer is rarely given the time, or the authority, that meaningful review would need. **PRACTICE** The recruitment manager in the scenario had 30 roles to fill, no account of how the assistant weighed employment gaps, and no power to change its configuration.

When the complaint arrives, the reviewer is the easiest party to hold to account. I think this is the crumple zone in its office form, and I suspect it is already common in organisations that have treated "a human checks it" as a sufficient control. **PRACTICE** The evidence on how often this happens in office settings is thin, and I know of no systematic study of it. **EVIDENCE GAP**

3.3 Why a board should care about the edge

A board might reasonably ask why the attribution of blame after the event matters to governance before it. There are two reasons. The first is fairness to staff, since a process that loads responsibility onto people without giving them control will, over time, teach them to avoid AI tools or to use them without telling anyone. The second is that a crumple zone hides the real cause. If an investigation ends with the reviewer, the organisation never learns that the model changed or that the process gave the reviewer no real chance, and the same failure recurs with a different name on the file.

Elish's own framing points the same way, since the task she sets is to locate responsibility accurately in systems where agency is distributed and control is exercised at a distance.⁵ **PEER-REVIEWED** Accurate location is a design task, and it has to be done before the accident.

FOR COUNSEL

To confirm with counsel. Whether a supplier's requirement for human review shifts any legal exposure onto the deploying organisation or its staff is a legal question that this paper does not answer. The query is how the organisation's own terms and policies would be read if a reviewer relied on an output in good faith.



PART TWO

The seams in the system

How national policy and commercial supply chains leave gaps between the parties.

CHAPTER FOUR

The sector-led model and its seams

The UK has chosen to regulate AI through existing regulators, and its white paper is candid that some risks may fall between their remits.

The white paper published on 29 March 2023 is the government's most detailed account so far of how it intends to govern AI, and it is open for consultation until 21 June 2023.⁶ **POLICY** Because it shapes the environment in which UK boards will make their own arrangements, this chapter reads it with one question in mind, which is where it leaves the allocation of responsibility.

4.1 What the white paper proposes

The white paper sets out five cross-sectoral principles for existing regulators to interpret and apply within their remits. These are safety, security and robustness; appropriate transparency and explainability; fairness; accountability and governance; and contestability and redress.⁶ **POLICY** It proposes that the principles should first be issued on a non-statutory basis. Following that initial period, and when parliamentary time allows, the government says it anticipates introducing a statutory duty on regulators to have due regard to the principles, and it adds that it would not do so if monitoring showed the framework working without legislation.

The accountability and governance principle is the one most relevant here. The white paper says that governance measures should ensure effective oversight of the supply and use of AI systems, with clear lines of accountability across the AI life cycle, and that regulators will need to find ways to place clear expectations on appropriate actors in the supply chain.⁶ **POLICY** It also concedes that establishing such lines is difficult, given the complexity of AI supply chains and the adaptivity, autonomy and opacity of the systems.

In my reading, the design has real strengths. Regulators already know their sectors, the principles are broad enough to fit a wide range of uses, and a non-statutory start allows adjustment. **PRACTICE**

4.2 The gaps it acknowledges

The white paper is unusually frank about the weaknesses of the model it proposes. It records that some AI risks arise across, or in the gaps between, existing regulatory remits, and it reports industry's concern that regulatory gaps may leave risks unmitigated.⁶ **POLICY** It also notes a risk that some regulators may interpret their remit more broadly than intended to fill perceived gaps, which could increase incoherence.

On supply chains, the document goes further. It states that legal responsibility for compliance with the principles should sit with the actors best able to identify, assess and mitigate AI risks, that AI supply chains can be complex and opaque, and that it is not yet clear how responsibility and liability for showing compliance will be, or

should ideally be, allocated among existing supply chain actors.⁶ **POLICY** It then says that the government is not proposing to intervene and make changes to life cycle accountability at this stage, because it regards the question of liability as complex and too early to settle. It gives particular attention to foundation models, saying that their potential opacity can make it hard to identify and allocate accountability for outcomes generated by systems that rely on them.

For a board, the plain consequence is that the white paper does not tell an organisation where its share of responsibility begins and ends. My own view, which goes beyond what the white paper says, is that for the next several years the allocation within most supply chains will be determined mainly by contract and by the existing law each regulator already applies, and that organisations which wait for a central answer will wait some time. **PRACTICE**

4.3 Central functions and their limits

To manage the seams, the white paper proposes a set of central functions, initially delivered by government. They include monitoring, assessment and feedback; support for coherent implementation of the principles; cross-sectoral risk assessment, intended in part to make sure that risks without an obvious home in a single remit are addressed; support for innovators; education and awareness; horizon scanning; and promotion of interoperability with international frameworks.⁶ **POLICY** The document also raises the possibility that an independent body might be more effective in the longer term.

These functions look at the system from above. They can identify that a category of risk has no regulatory home and prompt action. They are not designed to decide, in a specific complaint, whether a supplier, a deployer or a user should answer for a specific outcome, and the white paper does not suggest otherwise. Whether a sector-led model can work well without a central function that monitors the gaps in real time is one of the open questions this paper returns to at the end, and the consultation invites views on exactly that. **EVIDENCE GAP**

FOR COUNSEL

To confirm with counsel. The white paper notes that data protection law distinguishes controllers from processors and that product safety law distinguishes producers from distributors. How those existing categories would apply to the organisation's own AI uses should be checked case by case.

CHAPTER FIVE

Supply chains and shadow use

Contracts tend to place risk where it can be written down, while the capacity to prevent harm often sits somewhere else in the chain.

The recruitment scenario had four commercial links, a model developer, an application builder, a deploying organisation and a professional user. That shape is becoming common. For generative AI in business it may already be the default, and it has consequences for accountability that the older literature, written when software was mostly built and run by the same organisation, did not anticipate. A fifth link, the member of staff who uses a public tool without anyone's approval, sits outside the chain altogether.

5.1 Four links in the chain

Bommasani and colleagues use the term foundation model for a model trained on broad data that can be adapted to a wide range of downstream tasks, and they warn that the same property creates homogenisation, so that defects in a foundation model may be inherited by every application adapted from it.⁷ **PREPRINT** A single change at the base of the chain can therefore alter the behaviour of many products at once, including products whose builders were not told of the change.

The NIST AI Risk Management Framework, released on 26 January 2023, makes the organisational version of the same point. It observes that AI actors in charge of one part of the life cycle often do not have full visibility or control over other parts, and that good intentions in one area can be undermined by decisions made elsewhere.⁸

STANDARD Among its governance outcomes, it lists policies and procedures that address AI risks arising from third-party software, data and other supply chain issues, and contingency processes for failures or incidents in third-party data or AI systems deemed to be high-risk.

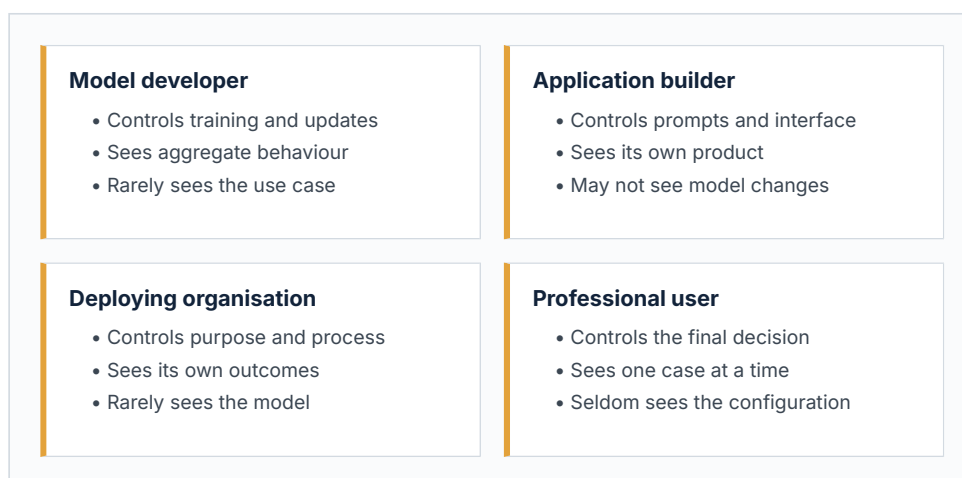


FIGURE 1 Four links in a typical generative AI supply chain, and what each can see. Source: author's analysis, drawing on Bommasani et al. (2021) and NIST (2023).

5.2 Where contracts put risk and where it sits

In the agreements I have reviewed in advisory work, standard terms from AI suppliers tend to do three things. They describe outputs as suggestions, they require the customer to review outputs before relying on them, and they limit the supplier's liability to a modest multiple of fees. **PRACTICE** This is commercially rational, and it mirrors the pattern Nissenbaum called ownership without liability.² **PEER-REVIEWED** The effect is to move risk down the chain towards the deploying organisation and, through its internal processes, towards the individual reviewer.

The capacity to prevent harm is distributed differently. The table below sets out my reading of where control and visibility sit, compared with where standard contracts tend to place the risk. It is an analytical summary, and the right-hand column reflects my own experience of commercial terms, which may not be representative. **PRACTICE**

Link	Can prevent	Often cannot see	Where standard terms tend to place risk
Model developer	Harms from training data and model updates	How the model is used downstream	Largely disclaimed
Application builder	Harms from design, prompts and testing	Changes to the underlying model	Limited, often capped at fees
Deploying organisation	Harms from purpose, process and oversight	Model internals and update history	Most of the operational risk
Professional user	Harms in the individual case	Configuration and known limitations	Carried through internal policy

TABLE 1 Control, visibility and contractual risk along an AI supply chain. Source: author's analysis, drawing on the NIST AI RMF and the UK white paper.

The mismatch in that table is the accountability gap in commercial form. The white paper itself warns that an inappropriate allocation of responsibility to a business that uses but does not develop AI could stifle adoption, and that allocating too much to developers of foundation models would hamper innovation.⁶ **POLICY** Both warnings are fair, and between them they describe the space in which organisations are currently left to negotiate for themselves.

5.3 Shadow use

The fifth link is the one boards see least. In early April 2023, trade press reported that engineers at a large electronics manufacturer, permitted to use ChatGPT for help with their work, had entered confidential material including source code and internal meeting notes.⁹ **INDUSTRY** I cite the reports as an illustration of a pattern and make no finding about the company. The pattern is that staff adopt public tools faster than organisations write policy, and that nobody inside the organisation owns the resulting decisions because nobody knows they are being made.

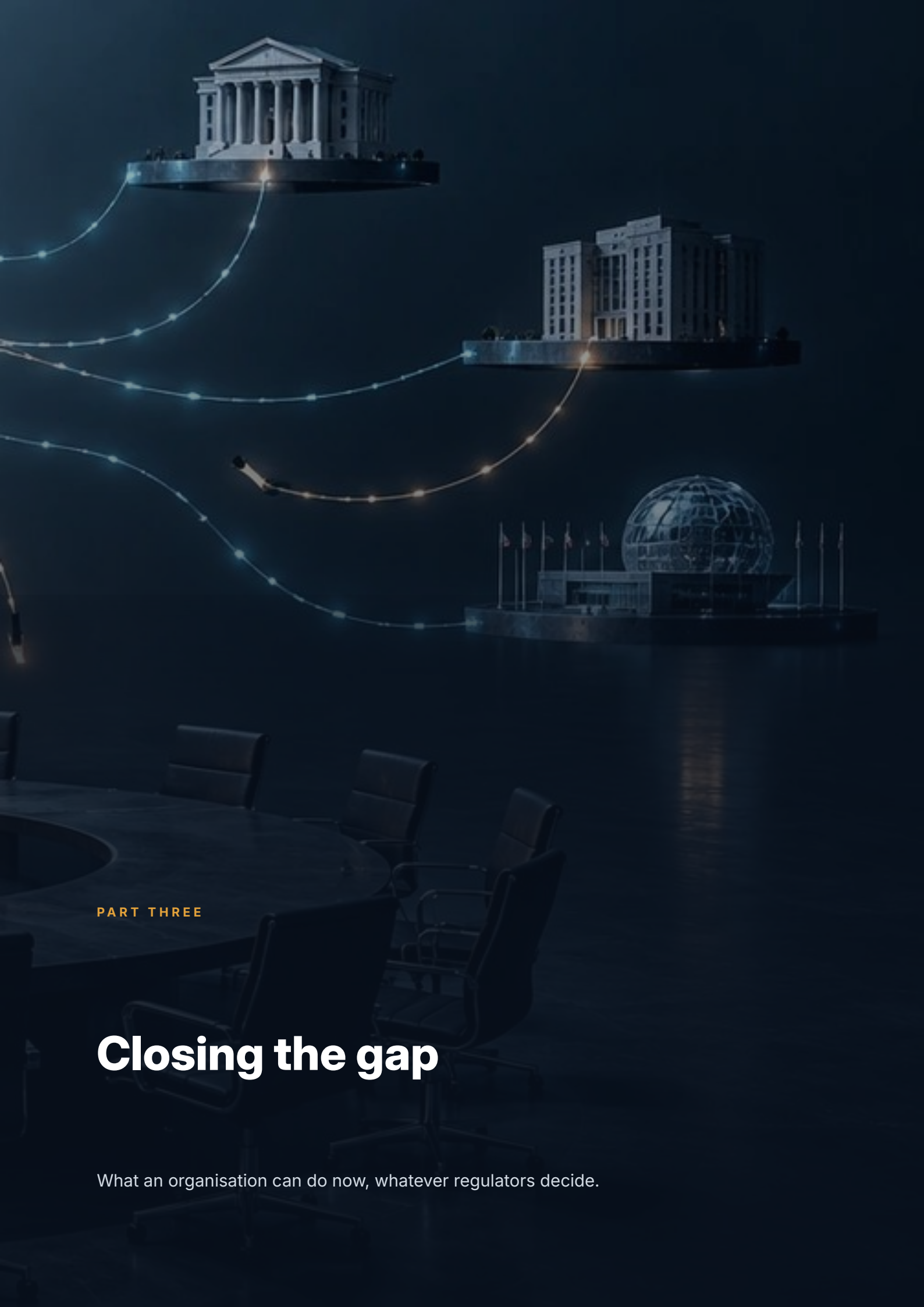
Regulators have begun to speak to this. The Information Commissioner's Office published eight questions on 3 April 2023 for organisations developing or using generative AI that processes personal data.¹⁰ **POLICY** They cover the lawful basis for pro-

cessing, whether the organisation is a controller, joint controller or processor, data protection impact assessment, transparency, security risks, limits on unnecessary processing, individual rights requests and solely automated decisions. The ICO says that an organisation using or adapting models developed by others may be a controller, joint controller or processor, which is a reminder that the role an organisation plays depends on what it does, and a member of staff using a public tool on the organisation's behalf may be creating obligations nobody has assessed.

The Italian case shows how quickly the question of a supplier's compliance can become a question for its users. The Garante's order of 30 March 2023 cited, among other things, the absence of information to users and data subjects, the lack of a legal basis for large-scale collection of personal data to train the platform's algorithms, and the absence of age verification.¹³ **POLICY** Its decision of 11 April set out measures the company must adopt by 30 April for the limitation to be lifted, covering an information notice, age gating, the legal basis for using users' data in training, and tools for people to seek correction or erasure of inaccurate data about them.¹³ **POLICY** The regulator's own statement made plain that the service would become accessible from Italy again only once those measures were in place,¹³ **POLICY** so any Italian organisation whose staff had come to rely on it lost access for reasons entirely outside its control (a small but instructive lesson in dependency).

FOR COUNSEL

To confirm with counsel. Where staff enter personal or confidential data into a public generative tool, the organisation's status and obligations under data protection law, and under its own confidentiality commitments, need to be assessed by a qualified adviser.



PART THREE

Closing the gap

What an organisation can do now, whatever regulators decide.

CHAPTER SIX

Closing the gap inside the organisation

An organisation cannot close every gap in a national framework, but it can decide who inside it answers for each AI-supported decision.

The previous chapters describe gaps that no single organisation created and none can fully close. This chapter concerns the part an organisation does control. My argument is that most of the internal gap can be closed with three disciplines applied together, a single accountable owner for each decision, a responsibility map across the supply chain, and internal audit that tests the seams. **ECAVEO** The first of these is the one most often missing.

6.1 A single accountable owner per decision

The unit of accountability should be the decision the AI system supports, such as shortlisting applicants or prioritising repair requests. Tools are shared. A single assistant may support a dozen decisions across several departments, and owning a tool tends in practice to mean owning its licence and its uptime. Owning a decision means answering for its outcomes, whichever tools, suppliers and people contributed.

For each such decision, I recommend that the organisation names one individual, at a level senior enough to change the process, who holds four responsibilities. **ECAVEO**

- The owner approves the use of AI in the decision and records the purpose it serves.
- The owner makes sure reviewers have the time, information and authority that meaningful review requires.
- The owner receives supplier notifications and incident reports and decides what follows from them.
- The owner answers to a named forum, usually a risk or audit committee, when the decision produces a bad outcome.

This reflects the NIST framework's expectation that roles, responsibilities and lines of communication for AI risk are documented and clear, and that executive leadership takes responsibility for decisions about AI risks.⁸ **STANDARD** It also answers Bovens's three conditions of an actor, a forum and a consequence.¹⁶ **PEER-REVIEWED** The owner is the actor, the committee is the forum, and the consequence is whatever the organisation already applies to accountable executives who fail in other domains.

A decision with several owners in practice has none. Joint ownership sounds collaborative, but in AI governance it recreates the problem of many hands inside the organisation. Name one person per decision, and let that person draw on as many colleagues as the work requires.

6.2 A responsibility map across the supply chain

The owner needs a map. For each AI-supported decision, the map records every party in the chain, from model developer to reviewing professional, and for each party what it controls, what it can see, what it has promised in contract and how it will tell the organisation when something changes. The table in the previous chapter is a generic version, and a real map is specific to each decision and each supplier. **ECAVEO**

Building the map is also the most reliable way I know to find shadow use, because the question "which tools contribute to this decision?" tends to surface tools that nobody approved. **PRACTICE** The map should be revisited whenever a supplier changes its model or terms, since a map drawn before a model update may describe a system that no longer exists.

1	Inventory decisions	List every decision AI supports, including informal staff use.
2	Name the owner	Assign one senior individual to answer for each decision's outcomes.
3	Map the chain	Record what each party controls, sees and has promised.
4	Set supplier terms	Secure change notice, incident reports and clear data use terms.
5	Audit the seams	Test hand-offs between parties, beyond each party's own controls.
6	Review on change	Redraw the map when models, suppliers or purposes change.

FIGURE 2 A working sequence for closing internal accountability gaps. Source: author's framework.

6.3 Internal audit as connective tissue

Raji and colleagues propose an end-to-end framework for internal algorithmic auditing, designed to be applied throughout an organisation's development life cycle.¹¹

PEER-REVIEWED Its stages are scoping, mapping, artifact collection, testing and reflection, and it is intended to help practitioners find harmful effects before deployment and trace problems back to their source afterwards. The framework was written for organisations that build AI systems. Much of it transfers to organisations that buy them, particularly the mapping stage and the emphasis on documentation that lets a failure be traced.

My own adaptation, which goes beyond what the authors propose, is that audit in a deploying organisation should test the seams between parties as well as each party's own controls. **ECAVEO** A useful audit question asks whether the reviewer in fact re-

ceives the information the supplier says it provides, and whether a model update announced by the developer ever reached the decision owner. In the organisations I have seen, those hand-offs are where the evidence most often runs out. **PRACTICE**

CHAPTER SEVEN

What to ask of suppliers, and what remains open

An organisation that has named its owners can then ask its suppliers for the information those owners need.

The recommendations in this chapter are my own and are offered for testing. **ECAVEO** They are written for a board that wants to act before the consultation closes and before any statutory duty exists, and they assume nothing about how either will turn out.

7.1 What to ask of suppliers

Contracts cannot close the gap on their own, but they can make sure the decision owner is told what the owner needs to know. I recommend that organisations seek the following commitments from AI suppliers, adjusted to the scale of the decision and the supplier's size. **ECAVEO**

- Advance notice of material changes to the underlying model, including changes made by an upstream developer.
- Prompt notice of incidents affecting the service, including data exposure and significant errors.
- Clear terms on whether customer inputs are used to train or improve models, and how that use can be switched off.
- Enough information about known limitations for reviewers to judge when an output needs closer scrutiny.
- Access to documentation and, where proportionate, to audit evidence, sufficient to trace a failure to its source.
- A tested exit route, so that the loss of a service does not leave a decision without a process.

Not every supplier will agree to every item, and smaller organisations have little bargaining power with large developers. Where a commitment cannot be obtained, the responsibility map should record the gap and the owner should decide whether the decision can tolerate it. **ECAVEO**

7.2 Recommendations for the board

I recommend that the board take five actions this year. **ECAVEO**

- Require an inventory of AI-supported decisions, including those supported by tools staff use informally.
- Require a named accountable owner for each such decision, answerable to a named committee.
- Require a responsibility map for each decision above an agreed threshold of impact, reviewed whenever a supplier changes its model or terms.

FOR COUNSEL

To confirm with counsel. How supplier commitments of this kind interact with liability caps, indemnities and the organisation's own obligations to customers is a drafting question for a qualified adviser.

- Ask internal audit to include the seams between parties within its plan.
- Adopt a clear policy on staff use of public generative tools, with a route for approving useful uses quickly.

A word on the last item. Blanket bans are tempting after the April reports, and in my experience they tend to push use out of sight. **PRACTICE** A fast approval route brings the use into the inventory, where an owner can be named.

7.3 Conclusion and open questions

The argument of this paper is that the fragmentation of responsibility for AI outcomes is a predictable result of distributing work across many hands, learning systems and regulators, and that no party needs to act badly for it to occur. The UK's sector-led model accepts that risks may fall between remits and, for now, leaves the allocation of responsibility along supply chains to contracts and existing law.⁶ **POLICY** Organisations can close most of the gap they control by naming an owner for each decision, mapping the chain and auditing its seams. **ECAVEO** Several questions remain open, and I set them out without pretending to answer them.

- Can a sector-led model work without a central function that monitors the gaps as they arise, and who would staff it?
- Should deploying organisations be required to name an accountable individual for each high-impact AI use, as some already do for other categories of risk?
- How should liability be shared when a foundation model changes beneath a deployed application whose builder was not told?
- What evidence would show that the reviewing professional is becoming a moral crumple zone in office work, and who is placed to collect it?

The last question matters more than its position suggests. Without that evidence, the debate about human oversight will continue to rest on the assumption that a person in the loop is a safeguard, and it may sometimes be a scapegoat instead.

END MATTER

References

1. Thompson, D.F. (1980) Moral responsibility of public officials: the problem of many hands. *American Political Science Review* 74(4), 905-916.
2. Nissenbaum, H. (1996) Accountability in a computerized society. *Science and Engineering Ethics* 2(1), 25-42.
3. Matthias, A. (2004) The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3), 175-183.
4. Santoni de Sio, F. and Mecacci, G. (2021) Four responsibility gaps with artificial intelligence: why they matter and how to address them. *Philosophy and Technology* 34(4), 1057-1084.
5. Elish, M.C. (2019) Moral crumple zones: cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society* 5, 40-60.
6. Department for Science, Innovation and Technology (2023) *A pro-innovation approach to AI regulation*. CP 815, 29 March 2023.
7. Bommasani, R. et al. (2021) On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
8. National Institute of Standards and Technology (2023) *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1, 26 January 2023.
9. Maddison, L. (2023) Samsung workers made a major error by using ChatGPT. *TechRadar*, 4 April 2023.
10. Information Commissioner's Office (2023) Generative AI: eight questions that developers and users need to ask. Blog, 3 April 2023.
11. Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33-44.
12. Future of Life Institute (2023) Pause Giant AI Experiments: An Open Letter. 22 March 2023.
13. Garante per la protezione dei dati personali (2023) ChatGPT limitation order, 30 March 2023, and subsequent decision of 11 April 2023.
14. Stanford HAI (2023) *Artificial Intelligence Index Report 2023*. 3 April 2023.
15. UK Government (2023) Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI. Press release, 24 April 2023.
16. Bovens, M. (2007) Analysing and assessing accountability: a conceptual framework. *European Law Journal* 13(4), 447-468.
17. Cooper, A.F., Moss, E., Laufer, B. and Nissenbaum, H. (2022) Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FACCT '22)*, 864-876.

END MATTER

About the author

Paul Forrest is the founder of Ecaveo. He advises boards and leadership teams on the adoption and governance of artificial intelligence, drawing on a background in business process reengineering and on hands-on work building machine learning systems, including small, domain-specific language models. He has served as a special adviser to UK government on the Global Entrepreneur Programme and works as a board member and business adviser.

He is also a qualified athletics coach who coaches sprint, hurdle and endurance athletes as a volunteer, and much of his thinking about how people learn, decide and improve comes from the track.

This paper is part of the Ecaveo Working Papers, AI Adoption: The Higher-Order Issues. The views expressed are the author's own.

ecaveo.com

Everybody's System, Nobody's Risk

Ecaveo Working Papers, AI Adoption: The Higher-Order Issues
Paper four of six · 28 April 2023