The Architecture of Real Decisions

Why pension trustee boards risk becoming the place where decisions are signed rather than made, and what care-architected technology should do about it.

Knowa

The AI governance operating system for Boards and advisers · knowa.co

In a 2019 paper, Mark LaCour and colleagues deliberately programmed an on-screen calculator to give a group of undergraduates the wrong answers to certain problems.[1] Even when the calculator produced a patently absurd age result (114 years old for someone born in 1945 and assessed in 1994), a large minority of calculator users still showed no suspicion.

The setting is unremarkable. The stakes are low. Nobody is harmed. And precisely because of that, the finding is more unsettling than the dramatic case studies usually wheeled out to make this point. Two costs of disagreement were operating here. The first was practical, and it was essentially zero: the students could have re-done the arithmetic in seconds and caught the error. On that cost alone, deference makes no sense. The second was psychological, the cost of overruling a confident machine, present even when the machine is a pocket calculator and the question is a sum. That is the cost that produced the deference. The capacity for independent verification was intact. It was simply not exercised, because exercising it required overruling the machine.

That second cost is what this paper is about. It does not disappear when the machine becomes more sophisticated and the stakes become consequential. It compounds. This paper is about what happens when it does, in a setting where the integrity of millions of retirement outcomes depends on whether decisions are genuinely made by the people legally responsible for them. It is about pension trustees, the technology now sitting alongside them in the room, and what we at Knowa believe that technology should be designed to do.

SECTION 01

What the research has been telling us for fifty years

The cost the calculator study exposes has a name in the human-factors literature, and a half-century of research behind it. The findings are uncomfortable for any field that imagines its decisions are still being made by the human in the chair.

Automation complacency

When a system is reliable most of the time, the human supervising it stops genuinely supervising. Attention drifts. Verification becomes ceremonial. The failure mode is not that the human refuses to intervene; it is that the human no longer notices the conditions that should trigger intervention. Radiology has shown that even experienced clinicians are predictably swayed by confident but incorrect AI-flagged findings.[2] Similar vigilance and monitoring problems have long been studied in aviation, anaesthesia and other high-reliability settings.[3]

Automation bias

When a confident machine output meets an uncertain human, the human predictably defers under those conditions, even when the machine is wrong, and even when the human possesses information the machine does not.[4] This is not a failure of training. It is a feature of how human cognition handles asymmetric apparent competence. The more sophisticated the machine becomes, the higher the psychological cost of disagreeing with it, because disagreement now requires the human to claim a knowledge advantage they cannot easily articulate. The calculator study is the lowest-stakes demonstration of this cost: an arithmetic error the student could have caught by hand, accepted because the device was confident and the student was not. The same mechanism, scaled up, is what this paper is about.

The vanishing loop

“Human in the loop” is the most overworked phrase in modern technology governance. It implies that placing a person at the end of an automated pipeline restores judgement to the process. The evidence from fifty years of human-factors research is consistent: it does not, unless the loop is specifically designed so that the human can do the cognitive work the loop exists for. A human who has thirty seconds to review what a machine spent thirty minutes synthesising is not in the loop. They are signing the loop.

These findings sit within a long-running body of human-factors research. They have reshaped cockpit design, surgical checklists, nuclear control rooms, and ICU monitoring protocols. They have not yet meaningfully reshaped the design of the technology that sits in front of trustee boards.

SECTION 02

The trustee board as a system under stress

Consider the conditions under which a UK pension trustee board actually operates.

The board meets quarterly, sometimes more frequently for in-flight projects, usually for half a day. Its members are a mix of member-nominated trustees, employer-nominated trustees, and increasingly a professional or independent trustee. For lay trustees, the role is not their full-time occupation and was not in the job description for which they were originally hired. Even for professional trustee firms, the workload across a portfolio of schemes makes any single meeting a compressed exercise in absorption and response.

The materials for the meeting (investment review, covenant update, administration report, regulatory developments, actuarial papers, scheme actuary's note, ESG reporting, valuation papers in season) frequently exceed several hundred pages and arrive late in the week before the meeting.

The investment consultant has done the analysis, framed the choices, and recommended the action. The scheme actuary has set out the assumptions. The legal adviser has drafted the resolution. The trustee board's job, as a matter of law, is to bring independent judgement to those recommendations. The operating model around them, as a matter of practice, exerts considerable ratification pressure: conscientious people, given long packs and short windows, tend to confirm rather than challenge unless the architecture around them makes challenge cheap.

This is not a moral failure on the part of trustees. It is a structural one. The legal architecture of trustee duty assumes a board that is genuinely deliberating: fiduciary obligation, the prudent person standard, sections 35 and 36 of the Pensions Act 1995, the trustee knowledge and understanding requirements in the Pensions Act 2004, and the regulatory expectations set out in The Pensions Regulator's General Code (which is not itself the law but which courts and tribunals must take into account when assessing how those duties are met).[5] The operational reality assumes a board that is genuinely overworked. The gap between the two has been absorbed for years by a tacit cultural agreement: the advisers produce defensible recommendations, the board documents that it considered them, and the audit trail looks adequate. Real decisions, in the demanding sense the law intends, happen unevenly.

The legal architecture assumes a board that is genuinely deliberating. The operational reality assumes a board that is genuinely overworked. The gap between the two has been absorbed for years by a tacit cultural agreement.

Into this system, AI is already arriving. Not because anyone has chosen to deploy AI for trustee work, but because consumer AI is now ambient: ChatGPT in the browser, Copilot embedded in Office, Gemini surfacing in Google Workspace, generic LLMs one prompt away from any laptop in any meeting room. The choice being made, often without anyone realising it is being made, is not "shall we use AI?" but "shall we use AI tools that were never designed for this work?" Tools that hold no scheme-specific permissions. Tools that produce confident summaries without citations. Tools whose data-handling policies were written for general consumer use, not for fiduciary records subject to data protection, professional privilege and regulatory scrutiny.

None of it is the result of an explicit decision. It is the result of the quietest possible failure of governance: the failure to notice that a decision was being made at all.

We are already seeing this in our market conversations: a trustee pasting a consultant's recommendation into ChatGPT to ask whether it looks reasonable, an adviser running a draft manager paper through Copilot to tighten the language, a scheme secretary asking a generic LLM to summarise a Pensions Regulator update before the meeting. None of it is governed. None of it was designed for the work.

This is the calculator study at trustee scale. The verification the legal architecture assumes is occurring is, in practice, displaced by trust in a system that was never built for the purpose.

What a purpose-built alternative for trustee work has to do is the opposite by default: scheme-level permissions, full audit trails, citations back to source, and a controlled environment confidential papers do not leave. That is the precondition for the platform being trusted with the material at all.

Knowa Q is built specifically for UK trustee boards. Precisely because we have committed to it as a category, we owe a clear-eyed account of what the category could become if it is designed badly. The category is being defined right now, and if it is not defined deliberately, the default will be set by the consumer chatbots already in the room.

SECTION 03

Two futures for the trustee technology stack

Every capable tool placed inside a stressed decision system pulls in one of two directions. It either reduces the cost of doing the thing the system is meant to do, or it reduces the cost of appearing to do that thing. These are not the same outcome, and the design choices that produce them are largely invisible to the buyer.

The smooth-ratification future

This future is already partially arriving, by default, through the consumer chatbots described above. A purpose-built platform that follows their lead would entrench it. In this version, AI for trustee boards is optimised for what feels efficient. The trustee arrives at the meeting under-prepared. They ask the platform whether the consultant's recommendation looks reasonable. A poorly designed platform would synthesise the SIP, the fund's recent performance, the peer comparison, and the regulatory backdrop, and produce a confident, well-cited affirmation. The board approves. The minutes record a thoughtful exchange. The audit trail is, in fact, more comprehensive than it used to be.

Nothing visible has gone wrong. The board feels more informed. The consultant's recommendations are now backed by what looks like a second source. The sponsor's covenant exposure looks well managed. And yet the actual cognitive content of the meeting, the moment at which an independent trustee mind engages with a recommendation and tests it, has been hollowed out further than before.

The friction that used to occasionally produce a real question has been smoothed away. The board has been moved one step further from being the place where decisions are made. And, quietly, a piece of software has been allowed to behave as though it were giving advice, which is precisely the line a trustee technology platform must never cross.

This is the dark mirror of trustee technology: sophistication-laundering for decisions that were never really debated. It is the calculator on the desk, scaled to industrial governance.

The architected-judgement future

In this version, the platform is designed around a different question. Not "how do we make it easier for the board to confirm?" but "how do we make it cheaper for the board to challenge?"

The two questions sound similar. They produce radically different products.

A platform that reduces the cost of challenge looks like this. It surfaces where the board's previous decisions are inconsistent with what is now being proposed. It identifies the quiet drift in fund characteristics that the headline performance number disguises. It points to the question the board asked last September that was answered with a promise to follow up, and never was. It makes the second, uncomfortable question, the one most trustees lack the time or context to formulate, almost free to ask.

It treats the trustee's scarce attention as the precious constraint to be optimised for. Not the consultant's preferred conclusion as the answer to be ratified.

This is the future Knowa is building toward. The product stages it implies are set out further on. Each exists to make the architected-judgement future cheaper to live in, and the smooth-ratification future harder to slip into by accident.

The two questions sound similar. They produce radically different products. One reduces the cost of confirming. The other reduces the cost of challenging.

SECTION 04

Care as the architecture of clarity

This is where a phrase we have used internally and somewhat poetically becomes a methodology we can defend technically.

Care, as we mean it at Knowa, is not the absence of friction. A frictionless system for trustee decisions would produce smooth ratification, and we have just argued that this is precisely the failure mode the field cannot afford. Care is the deliberate preservation of the friction that creates judgement, alongside the deliberate removal of the friction that wastes attention.

Most enterprise software is designed on the opposite assumption. It treats friction as undifferentiated waste, and efficiency as undifferentiated good. This works well for systems whose purpose is to execute defined workflows quickly. It fails dangerously for systems whose purpose is to support genuine deliberation under trustee duty, because in that setting some of the friction is the work.

Care-architected technology distinguishes between four kinds of friction:

- ***Wasted friction.*** Locating a document, reconciling versions, remembering which manager the consultant referenced three meetings ago, retyping context that already exists somewhere. This friction should be removed without hesitation. It produces nothing but exhaustion.
- ***Productive friction.*** The pause that surfaces an inconsistency. The flag that points to an unanswered question. The prompt that asks whether the rationale being offered for this manager is the same rationale that was rejected for a different manager last year. This friction should be preserved, and indeed engineered for, because it is the medium through which independent judgement actually occurs.
- ***Cosmetic friction.*** The performance of deliberation: lengthy minutes, ceremonial votes, polished memoranda that document a debate that did not happen. This is the most dangerous category, because it is indistinguishable from real friction in any audit and is increasingly easy for AI to manufacture at industrial scale.

- ***Constitutive friction.*** The slow, deliberate cadence of meetings, papers, minutes and reviews that the law has built around trustee decisions. This friction exists because the decisions matter and the people making them are not full-time experts. Technology that compresses this cadence in the name of efficiency is not making trustee work easier; it is making trustee work less recognisable as trustee work.

A care-architected platform is one whose default behaviours are calibrated against this taxonomy. It removes wasted friction aggressively. It engineers productive friction deliberately. It refuses to manufacture cosmetic friction even when asked. And it respects constitutive friction even when faster alternatives exist.

SECTION 05

What this means for Knowa

Knowa Q is built specifically for UK trustee boards, and the standard for the category is being set, in effect, as it ships. It would be dishonest to write this paper as a description of a finished product. It is more accurately a description of the standard against which we are willing to be measured, and of the design tensions we have committed to taking seriously.

Knowa is used by more than a thousand UK trustee boards, representing over £300 billion in assets under governance. The platform serves both lay trustee boards and professional trustee firms, where the shape of the problem is similar but the scale is amplified: a professional trustee carrying ten or twenty appointments needs the architecture of judgement to work consistently across every one of them, with no scheme's record blurring into another's.

The vision runs in three stages. One is live today. One is currently in development. One is on the horizon.

Knowa (Core), live today

The first stage gives the board a single source of truth for its work, in three parts that share one governance record.

The **Collaboration Space** is the work between meetings: every discussion, vote and decision threaded to the paper or policy it relates to, so the discussion itself becomes part of the record.

Meeting Flow is the work of the meeting itself: the pack assembled cohesively from Knowa and external content, trustees reading and annotating in advance, the discussion captured live, minutes drafted within the hour with actions already in the tracker.

The **Compliance Modules** are the work of evidence: risk register, compliance register, action tracker, ESOG, Smart Vault, conflicts. Each module connected to the same record, so a control logged once surfaces wherever it is needed: the meeting pack, the audit trail, Knowa Q.

The board should never spend cognitive bandwidth locating the SIP, reconciling versions of the trust deed and rules, or wondering whether a deed of amendment was ever executed. That work disappears into the architecture, leaving the human attention free for the work only the human can do. Nothing gets lost. Nothing gets duplicated. The record survives the turnover.

Knowa (Core)nected, currently in development

The second stage gives the board one view across every obligation. Knowa (Core) connects to the live Modules (risk register, action tracker, conflicts, CPD, compliance calendar, trustee training) and each is wired into the same underlying source of truth. A decision in the Collaboration Space creates an action. The action updates a risk. The risk cites the paper that flagged it. The paper lands in the next meeting pack. The board no longer holds its obligations in the secretary's notebook or in last quarter's spreadsheet. It sees every obligation in one place, and every obligation sees the board. Compliance stops being a project. It becomes the default.

For professional trustee firms, this anchoring works at portfolio scale. Standards of evidence, rationale and follow-through become consistent across every appointment. A new colleague joining the firm inherits not a folder but a fully navigable history of how each scheme thinks. The firm's reputation, which is built on the consistency of judgement across appointments, has the architecture it needs to be defensible at the level of any single trustee meeting.

Knowa (Core)mand, the horizon

The third stage is where governance work begins to trigger itself. When the data is live, governance tasks fire on their own dates: a review cycle starts on time with the right documentation already assembled, a risk threshold alerts the right trustee, a regulatory update drafts the action it implies, routed to the right person with the rationale attached. Administration is automated. Judgement is supported, never removed. The board's time goes to the things only humans can do. The platform handles the rest.

Knowa Q, the intelligence layer that runs across all three

Knowa Q is the most consequential design surface in the platform, and it is what makes each of the three stages cohere. It is the current generation of an intelligence layer that has run through Knowa from the start, earlier expressed in machine learning and natural language processing, now expressed in the more capable form that today's models permit. It is also where the boundary between augmentation and advice has to be drawn most carefully. Knowa does not give advice. It is not a fiduciary, it is not an FCA-authorised investment adviser, and it is not a consultant. The decisions belong to the trustees. The advice belongs to the advisers. Knowa Q exists to make both of those roles work better, by ensuring that what each party knows about the scheme, the record and the relevant context is no longer constrained by what they happened to read on the train.

In practice, this means Knowa Q answers the question that was actually asked, against the record. If a trustee asks what the SIP says about manager replacement, Knowa Q surfaces the relevant sections with citations. If they ask when the board last considered a similar manager change, Knowa Q points to the meeting minutes where that conversation happened, the rationale recorded at the time, and any follow-up actions that were or were not closed. If they ask which areas of the General Code are relevant to a particular decision, Knowa Q answers from the scheme documents and the regulatory record.

Where Knowa Q draws the line is at recommendation. It will surface an inconsistency between two papers in the record, and not tell the board which is right. It will show that an action assigned three meetings ago has not been closed, and not opine on whether closing it should change the present decision. It will retrieve the SIP section a consultant's recommendation appears to depend on, and not assess whether the consultant's reading of that section is sound. Those are judgements, and judgements belong to the trustees and their advisers. Knowa Q ensures that when the human forms a judgement, the record they form it against is fully accessible, fully searchable and fully cited, not constrained by what they happened to read on the train.

An illustration. A trustee asks: "Have we considered this manager before?" A poorly designed AI might answer "the recommendation appears reasonable." Knowa Q answers: the board considered a similar replacement in September 2023; the stated concern was fee transparency; the follow-up action assigned to the consultant was not marked complete; here are the relevant minutes, the original SIP section, and the action item in the tracker. The board still decides. It does so with the record in front of it, not behind it.

Knowa Q answers questions about the record. The judgements belong to the trustees, grounded in their scheme's own history.

SECTION 06

Augmentation, not substitution: the adviser relationship

It is worth being explicit about what Knowa is not, because the temptation in any AI-adjacent product category is to let the boundary blur.

Knowa does not advise. It does not recommend managers, evaluate covenant, opine on strategy, or substitute for the judgement of an investment consultant, scheme actuary, legal adviser or professional trustee. The legal architecture of regulated advice exists for good reasons, and the people qualified to give that advice have spent careers earning the right to do so. Replacing them with software is neither the goal nor a coherent ambition.

What Knowa does is augment the relationship between the board and the people who advise it. The consultant arrives at the meeting better prepared, because the platform has surfaced the questions the board is likely to raise and the inconsistencies in the record that deserve a response. The board arrives better prepared, because the platform has made the consultant's analysis legible against the scheme's full history rather than against whatever the board happens to remember. The conversation between them, which is the actual mechanism by which trustee judgement is exercised, becomes denser, more grounded, and more genuinely deliberative. The smooth-ratification future is one in which AI replaces the substantive content of that dialogue with a polished synthesis. The architected-judgement future is one in which AI raises its floor.

Advisers, in our experience, recognise this distinction immediately. A platform that pretends to give advice threatens their role. A platform that makes their advice land in a better-prepared room makes their role easier and more valuable. The same is true of professional trustees, in-house counsel, scheme secretaries and administrators. Augmentation expands the relevance of every party who already brings genuine expertise to the table. Substitution would diminish them, and the members whose retirements depend on those relationships would be worse off for it.

Knowa does not advise. It raises the floor of the conversation between those who do advise and those who must decide.

SECTION 07

The standard we owe the people who are not in the room

Trustee duty exists for the members. They are the people whose retirement security depends on whether the board made a real decision or a ceremonial one. They are not in the meeting. They will never read the minutes. They have no view of which questions were asked and which were not. Their interests are entirely structural, encoded in the design of the system that decides on their behalf.

This is why the design choices in trustee technology are quietly ethical choices. A platform that optimises only for the comfort of the board or the legal cover of the sponsor is serving constituencies that are already well represented in the room. A platform that optimises for the integrity of the decision, and for the quality of the conversation through which advisers and trustees reach it together, is serving the constituency that is not.

That is the test we apply to ourselves; and, we believe, the test the next generation of governance technology will be measured against, in regulation, in litigation, and in the slower verdict of which members ended up adequately served by the systems built in their name.

Care, in this sense, is not a sentiment. It is a structural commitment to the people the system exists to protect, expressed in the architecture of the tools that decide for them. It is what we mean when we say that clarity is not the absence of complexity, but the presence of a deliberate hand.

It is what we are building.

Knowa

The AI governance operating system for Boards and advisers · knowa.co

REFERENCES & FURTHER READING

Sources cited in this paper

A short bibliography for readers who want to follow the human-factors and legal references in the text. The numbering corresponds to the superscripts in the body of the paper.

Human-factors research

1. LaCour, M., Cantú, N. G. & Davis, T. (2019). When calculators lie: A demonstration of uncritical calculator usage among college students and factors that improve performance. *PLOS ONE*, 14(10): e0223736.
2. Dratsch, T. et al. (2023). Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology*, 307(4): e222176.
3. Weinger, M. B. & Englund, C. E. (1990). Ergonomic and human factors affecting anesthetic vigilance and monitoring performance in the operating room environment. *Anesthesiology*, 73(5): 995–1021.
4. Parasuraman, R. & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3): 381–410. The standard review of automation complacency and automation bias as related but distinct phenomena.

UK pensions law and regulation

5. Pensions Act 1995, sections 35 (investment principles) and 36 (choosing investments); Pensions Act 2004, sections 247–248 (trustee knowledge and understanding); The Pensions Regulator (2024), *General Code of Practice*, in particular the modules on the role of the governing body, knowledge and understanding, meetings and decision-making, systems of governance and internal controls. The General Code is not itself a statement of law, but courts and tribunals must take its relevant terms into account when assessing how legal duties are met.

Background reading

- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6): 775–779. The original articulation of the “human in the loop” problem.
- Goddard, K., Roudsari, A. & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1): 121–127.
- The Pensions Regulator (2024). *General Code of Practice*, in particular the modules on governing body, knowledge and understanding, decision-making, and risk management. Available at thepensionsregulator.gov.uk.