Does AI Have Any Humanity? Article 7 of 7

The decisions we still own

By Dr Peter Fields

THE 2-MINUTE SUMMARY

In 1983, Bainbridge described an irony of automation: removing routine work can leave people responsible for difficult interventions while depriving them of the practice needed to intervene well. Today's AI gives that old problem a new reach.

Human involvement is worth little if the person involved cannot understand or alter what happens. Responsibility must also reach those who build systems, choose their purposes and set the conditions under which others use them. The last person to approve a result is not automatically the person best placed to prevent the failure.

We needn't defend human error to defend human agency. Properly tested assistance may improve decisions and make previously difficult tasks accessible. The relevant comparison includes both the benefits and the errors introduced by relying on it. People affected by a decision also need ways to challenge the evidence and, where appropriate, the criteria used.

There is a longer-term question about learning. Better performance with a tool does not establish that its user has learnt to perform independently. Schools and employers need to decide where building judgement is part of the task, and provide for it when routine work changes.

This series has found real value in AI's achievements without treating them as proof of an experiencing subject. AI can contribute to creative work, offer useful support and widen our practical freedom. Whether it has interests of its own remains a separate, unresolved question. Our responsibility for the uses we choose is already here.

An old warning for a new machine

The difficult part of automation can be the work it leaves behind.

Bainbridge made that point in 1983, writing about industrial systems. Once routine operations have been automated, people may be expected to monitor them and take over during a failure. Yet recognising the failure and recovering from it can require skills that ordinary work previously kept alive. The operator is asked to become an expert precisely when the system stops doing the job.[1]

This was not a study of chatbots. Its relevance is the problem it identifies: supervision takes knowledge, practice and suitable working conditions. Calling somebody a supervisor does not provide any of them.

We now routinely hear that a human will check an AI's work. That promise can describe a serious arrangement. It can also conceal the absence of one. A reviewer needs access to the evidence, enough time to assess it and permission to reject the recommendation. Without those things, an instruction to exercise judgement is liable to become an instruction to accept responsibility.

That is where this series ends: with the decisions people and institutions continue to own, whatever we conclude about the machinery.

Responsibility starts before the result arrives

Elish calls attention to the *moral crumple zone*: the visible human operator can absorb blame for a failure despite having limited control over the larger system that produced it.[2] The expression is memorable because it captures a familiar unfairness. The person nearest the event is often easier to identify than the people who made it likely.

Responsibility should follow what each participant knew, could influence and was required to do. That includes decisions made well before anyone reviews an output.

A model provider can affect training, testing and the information supplied about limitations. An application developer chooses how the model is connected to records or tools and what users are shown. An organisation adopting the application determines its purpose, staffing and review procedures. An operator may then have obligations to check a particular result or report a problem. These roles can overlap. None disappears simply because another participant also made a mistake.

This is an argument about the allocation of responsibility, not a statement that every participant has the same legal liability. The obligations must fit the role. A person cannot reasonably be expected to correct a defect they have neither the information to recognise nor the authority to remedy. Equally, limited control is no excuse for ignoring an obvious danger within one's power to address.

Floridi's account of artificial agency helps separate a related issue. Systems can act and produce consequential results without our first settling whether they understand as people do.[3] His stronger description of this as agency without intelligence depends on what we mean by intelligence. We need not adopt it to recognise that successful action leaves another question unanswered: who will give reasons, reconsider the decision and put matters right?

A generated apology does not perform those duties. The service has to make correction possible.

Geek Box: When is human oversight meaningful?

Information: can the reviewer examine relevant evidence, omissions and uncertainty, rather than only a recommendation?

Competence and time: can they assess it properly under the conditions in which they actually work?

Authority: can they pause or change the action, reject a result and obtain help?

A response to problems: will a challenge receive attention, and can repeated failures lead to changes in the system?

These are practical tests I propose, informed by Bainbridge and Elish.[1][2] They do not require individual approval of every decision. Some activities need supervision of the process as a whole. What matters is whether the chosen form of control works.

What if the machine is better?

People overlook evidence, favour familiar stories and get tired. Protecting our right to repeat those errors would be a poor defence of humanity.

Goddard, Roudsari and Wyatt's review of automation bias gives us a more useful comparison. Automated advice can improve performance overall while introducing mistakes through inappropriate reliance. Their evidence concerns older decision-support systems, including aviation and healthcare, rather than today's conversational models. It nevertheless shows why benefits and newly introduced errors must be assessed together.[4]

One bad AI-assisted decision does not prove that unaided people would have done better. A favourable average, however, can conceal repeated failures affecting particular groups. We need to know who gains and who bears the errors, as well as whether the total score improves.

Nor must a person reproduce every calculation to retain control. We rely on tools whose workings we could not recreate ourselves. For some uses, the appropriate job is to test performance, set limits and monitor failures, with a clear means of stopping an unsafe process.

The person affected by a decision needs more than reassurance that this has been done. They may need to correct missing information or challenge the rule that made it decisive. An appeal about an inaccurate record differs from an objection to using that record at all. A fair process must recognise the distinction.

The person operating the software is not always the person whose freedom is at stake. Faster processing for an organisation can coexist with less opportunity for its customers, employees or applicants to be heard. Personal advice about writing better prompts cannot remedy a procedure imposed on people who never get to use the tool.

Geek Box: A specific UK legal example

The UK GDPR now distinguishes decisions involving meaningful human involvement from those based solely on automated processing. Its provisions concern decisions with legal or similarly significant effects on an individual, using personal data. Safeguards for significant solely automated decisions include opportunities to make representations, obtain human intervention and contest the decision.[5]

These provisions, introduced by section 80 of the Data (Use and Access) Act 2025, took effect on 5 February 2026. Their application depends on the statutory conditions; they are not a general rule covering every AI output.

They illustrate a useful principle beyond that particular legal setting: the presence of a person somewhere in a process is different from meaningful participation in the decision.

The work that teaches us to judge

An experienced writer may spot the weakness in an AI draft at once. A beginner may read the same passage without knowing what needs checking. Both have received assistance. They have not received the same benefit.

Bastani and colleagues tested two GPT-4-based tools during four mathematics study sessions at a high school in Turkey. Both improved assisted practice scores. Students using the more open-ended tool

subsequently did worse than the control group in unaided tests. A version designed around teacher-supplied material and hints largely avoided that disadvantage, but did not demonstrate better unaided performance than the control group.[6]

The finding does not establish a general decline in intelligence. It exposes a distinction we can easily overlook when a completed task looks successful: getting the answer and learning how to reach it can come apart. The design of the help changes the result.

For a one-off administrative job, mastering the underlying task may be unnecessary. For a student or apprentice expected to recognise future errors, acquiring that judgement can be the main purpose. Removing effort indiscriminately confuses those cases.

Nobody needs to struggle merely to earn the right to learn. A good explanation or timely hint can make difficult work accessible. The question is whether the learner still does enough of the relevant thinking to develop it, with feedback that shows where they went wrong.

Employers face the same practical issue. If AI reduces junior work, how will future senior staff acquire the experience on which reliable judgement depends? The answer need not be to preserve every old task. It does need to be more substantial than hoping people will learn by checking outputs. Replacement training requires time, funding and evidence that it works.

Bainbridge's warning returns here in a different form. We should be wary of eliminating the work that teaches judgement and then treating judgement as a resource that will always be available on demand.[1]

Assistance can give agency back

Human agency does not mean doing everything alone. Help with reading, writing or communicating may allow someone to pursue a purpose they could not otherwise carry out. Requiring unaided performance as the price of independence would exclude people assistance ought to serve.

The practical test is whose purpose the assistance serves and whether the person has suitable ways to understand, correct or refuse what is proposed. That will look different for different people. Sometimes a trusted representative is needed. Dependence or difficulty explaining a choice does not diminish anyone's dignity.

Vallor argues for preserving the shared activities of questioning, exchanging reasons and forming commitments. She also recognises that AI could support those activities.[7] I find that constructive possibility more persuasive than a general dismissal of generated ideas as reflections of the past.

An assistant can expose a weakness in an argument or make an unfamiliar alternative easier to understand. The resulting challenge may enlarge freedom more than an agreeable answer would. There is no requirement that the machine share our experience for its contribution to be valuable.

We must still choose the purpose. A system can help compare the effects of competing policies, but better predictions cannot settle every dispute about what should matter. Converting a disputed aim into a numerical target may make it easier to administer while making the underlying disagreement harder to see. People need routes to question the purpose as well as the performance.

What remains of the humanity question

We began with the impression of someone behind the screen. A familiar voice and a remembered detail make conversation feel personal. Looking more closely revealed several processes and services assembled into that encounter. Neither its unity of presentation nor its distributed machinery settled whether a subject was present.

The strongest conclusion of this series is not that AI's achievements are hollow. It is that their value can be recognised more accurately once we stop asking each achievement to prove everything else.

There are grounds for attributing forms of task understanding where a system can apply what it has learnt, cope with relevant changes and give us reason to trust its performance. There are creative contributions worth judging on their merits. Supportive responses can provide real relief. These are substantive achievements even when they do not establish felt understanding, an author's lived experience or reciprocal affection.

Whether an AI could have morally relevant interests of its own remains open. Convincing conversation cannot carry that argument alone. Further evidence may change the position, and the possibility deserves investigation rather than a verdict chosen for convenience.

The word *humanity* also asks something of us. A considerate response can be embedded in an unfair service; an impersonal calculation can help somebody receive treatment they would otherwise miss. We have to assess the arrangement and its effects, including the people who get no opportunity to speak through the interface.

I would welcome AI that helps us understand more, express what we mean and act on better reasons. That prospect deserves ambition. It also requires institutions willing to explain their choices, change failed procedures and listen to those affected. The humanity of our use of AI will depend on whether people retain that influence, long after the novelty of the conversation has worn off.

Sources and notes

1. Lisanne Bainbridge, Ironies of automation, *Automatica* 19(6), 775–779 (1983). Examines the difficulties of leaving people to supervise automated processes and intervene during failures, particularly when maintaining the necessary skills becomes harder.

2. Madeleine Clare Elish, Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction, *Engaging Science, Technology, and Society* 5, 40–60 (2019). Explains how responsibility can fall on a visible human operator despite control being distributed across a larger system.

3. Luciano Floridi, AI as Agency Without Intelligence: on ChatGPT, Large Language Models, and Other Generative Models, *Philosophy & Technology* 36, article 15 (2023). Separates successful artificial action from human understanding and directs attention to the consequences of designing and using these systems.

4. Kate Goddard, Abdul Roudsari and Jeremy C. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, *Journal of the American Medical Informatics Association* 19, 121–127 (2012). Reviews inappropriate reliance on automated advice, alongside its benefits, and examines the effects of users, tasks and system design.

5. Data (Use and Access) Act 2025, section 80, inserting UK GDPR Articles 22A–22D; SI 2026/82, regulation 2(j). The provisions concerning meaningful human involvement and safeguards for significant solely automated decisions took effect on 5 February 2026. They provide a specific legal example of the distinction between nominal and effective involvement.

6. Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı and Rei Mariman, Generative AI without guardrails can harm learning: Evidence from high school mathematics, *Proceedings of the National Academy of Sciences* 122(26), e2422633122 (2025). Compares two GPT-4-based forms of assistance with a control group, separating performance during supported practice from subsequent unaided assessment.

7. Shannon Vallor, In a Mirror, Dimly: Why AI Can't Tell Our Stories and Why We Must, *Journal of Media Ethics* 40(4), 139–150 (2025). Argues for preserving the shared activities of reasoning, forming purposes and telling our own stories, including through tools designed to support those activities.