Intelligence Must Remain Answerable

Evidence, authority and care in human–AI systems

Arun Leer
arun.wiki

16 September 2026 · Version 0.2

Conceptual perspective and research proposal
Working manuscript for critical review. Not peer reviewed.

Abstract

When an artificial intelligence system gives better answers, organisations face a further decision: what should it be allowed to do? This paper argues that increasing capability can justify greater delegation but cannot, by itself, justify weaker answerability to those affected. It develops this claim through research on human judgement, artificial self-correction, meaningful human control and possible AI welfare. The central contribution is a proposed account of four unsupported inferences: from local correction to general reliability, from reliability to authority, from conscientious language to experience, and from possible welfare to decisive authority for an expressed preference. These inferences concern different kinds of justification; evidence may connect them only through an explicit argument. The paper proposes a compact addition to existing assurance practice that records a claim, its evidence, the requested action and the reason connecting them, alongside challenge and remedy. Hypothetical cases in community provision, education and business show how the method can change a decision while preserving useful delegation. A controlled study is specified to compare it with credible assurance checklists and matched reflection, measuring unjustified approvals, unjustified refusals and review costs. The ethical argument also distinguishes aspirations for responsible conduct from possible claims to moral consideration: protection, if warranted, would not be a reward for obedience. The contribution is a contestable conceptual synthesis and an unvalidated decision method, intended to help growing intelligence expand human possibilities without placing its use beyond question.

Keywords: artificial intelligence; accountability; human judgement; delegation; AI welfare; self-correction.

1 A decision that deserves an answer

Imagine a community centre introducing an AI assistant to arrange its sports timetable. The assistant fills more sessions, reduces empty spaces and makes the administrator's work easier. It also moves a small accessible exercise group to a time its members cannot reach by public transport. The timetable is successful by one measure and unsuccessful by another. The people excluded do not need to outsmart the software to deserve an answer.

This hypothetical case exposes a question that better prediction cannot resolve on its own: who may decide what counts as success, and what happens when that decision harms someone? A scheduling system can contribute evidence about capacity and demand. Its performance does not determine the centre's purpose, the acceptable distribution of inconvenience or whether a complaint should change the timetable. Those are decisions about power and obligation.

The same problem appears when a business lets an assistant issue refunds, a school relies on automated feedback or an organisation asks an AI system to revise its own procedures. An improvement creates opportunities. It also invites a transition from using an answer to accepting an authority. This paper argues that the transition needs reasons of its own. **Growing intelligence should expand what can be done without closing down who may question it.**

Here, *answerability* means an arrangement in which consequential decisions can be explained at an appropriate level, challenged by affected parties and changed or remedied by someone with the power to act. It is broader than a model producing an explanation. It includes the organisation's obligation to respond when its system fails, when its purpose is contested or when its success imposes unacceptable costs. The account is proposed as a normative standard, not asserted as an existing legal entitlement in every setting.

The paper makes three bounded contributions. It identifies four ways evidence about intelligence can be asked to justify more than it establishes. It proposes a small addition to existing evaluation and approval practices that makes the connecting argument explicit. It specifies how that addition could be tested against credible alternatives. The contribution is the organisation and examination of these distinctions; evidence-based approval, oversight and accountability already have substantial literatures. [1–3]

This is a conceptual perspective supported by a selective critical synthesis, not a systematic review or a report of new experiments. Sources were selected for their bearing on correction, reliance, responsibility and possible experience; the selection cannot establish comprehensive coverage or historical priority. The cases are constructed illustrations, not accounts of the author's clients or research participants. Empirical findings support particular claims about capabilities and vulnerabilities. Ethical premises are defended separately, and the proposed method remains open to rejection if it fails to improve decisions.

2 Why delegation does not end responsibility

2.1 The moral argument

The argument begins with two commitments. People have interests in the outcomes of decisions affecting them and in how those decisions are made. An organisation choosing to delegate consequential decisions should remain answerable for the conditions it creates, whether or not it ultimately benefits. These commitments can be contested, but neither can be derived from a neural network diagram or a performance score. They belong to the ethical justification of using power.

The first commitment explains why a person's objection matters even when a system is more accurate than that person. A passenger need not understand every calculation behind a transport system to question a decision that makes their journey inaccessible. Expertise bears on which claims are well supported. It does not make every affected person's interests disappear. Respect for human agency and participation is also part of established international AI-ethics guidance. [4]

The second commitment follows from a problem of displaced burdens. If a provider can retain the benefits of delegation while making its mistakes impossible to challenge, the costs of its choice can fall on people denied a meaningful response. Delegation would become a means of escaping obligations that remained important before automation. The claim defended here is that a change in the means of decision-making does not, without further argument, cancel those obligations. The required response can vary with the decision's consequences; the responsibility to justify that variation remains.

Together, these premises support three minimum conditions. Affected people need notice and reasons sufficient to understand the consequential decision. They need an accessible route to challenge it before someone with actual review power. Where a challenge is upheld, there must be a way to correct the decision or address its effects. Access should not depend on technical expertise, unaffordable effort or accepting retaliation. These are proposed ethical requirements. They do not require every person to approve every decision or every internal calculation to be publicly disclosed. These duties do not imply

that every adverse outcome establishes misconduct. They require a competent response to questions about justification, preventable harm and appropriate remedy.

Answerability also has costs: time, skilled staff, records, independent review and sometimes compensation. The organisation choosing a deployment should include those costs in its case for adoption. A system that is inexpensive only because affected people perform unpaid investigation and repair is not inexpensive for everyone. Emergency action may reasonably precede full deliberation, but urgency creates a reason for bounded authority and subsequent scrutiny, rather than an indefinite exemption.

2.2 Responsibility needs the power to act

Responsibility is not one question. Whether someone deserves blame for a particular output differs from whether an organisation owes an explanation, should repair a harm or must change its future practice. Earlier work on learning automata raised the difficulty of attributing responsibility for behaviour a designer could not fully predict; later scholarship distinguishes several kinds of responsibility gap. Uncertainty about individual culpability does not automatically dissolve duties of institutional preparation and response. [3, 5]

Nor is assigning a human reviewer enough. Elish's analysis of *moral crumple zones* describes how a nearby person can absorb blame despite limited control over automation. [6] A worker given seconds to approve a complex recommendation, without independent records or permission to stop the process, is a weak site for accountability. Naming that worker does not supply the missing resources.

The account of meaningful human control developed by Santoni de Sio and van den Hoven connects system behaviour to relevant reasons and to human understanding and responsibility. It also recognises that control does not guarantee morally good goals. [7] The present proposal builds on that distinction. Review duties should be matched by information, competence, time, authority and institutional backing. Responsibility should reach the people choosing procurement, staffing and incentives, as well as those responding to individual outputs.

The aim is consequently stronger than retaining a ceremonial human signature. An answerable organisation must be able to learn from an objection and alter its conduct. A capable system can support this by making evidence easier to inspect, identifying inconsistencies and helping people understand alternatives. The ability to challenge intelligence is compatible with making extensive use of it.

3 Learning without losing judgement

3.1 Human improvement remains possible

Human cognition offers a reason for humility and a resource for improvement. Motivated reasoning research describes how people can marshal justifications toward desired conclusions; research across the lifespan describes different trajectories for different abilities rather than one universal peak. These findings caution against treating either confidence or a demographic category as a complete measure of judgement. [8, 9]

The comparison between childhood experience and AI training is suggestive but limited. People learn through bodies, relationships and institutions; artificial neural networks inherit selected computational inspirations from neural activity while abstracting away much of this setting. Their intellectual connection to neuroscience is real, but it does not make human development and numerical training interchangeable. [10, 11] The useful comparison is more modest: formative influences can leave patterns that a later correction may not fully remove.

Research on implicit measures supplies one caution. An intervention can change a measured association without establishing a corresponding change in behaviour. [12] In machine learning, word

embeddings have reproduced social biases found in language, while some debiasing methods have left related structure recoverable after a targeted measure improves. [13, 14] These are different bodies of research, not evidence that people and models share a single mechanism. Both motivate asking whether an apparent correction transfers beyond the test that rewarded it.

This question is constructive. It treats learning as an opportunity to revise inherited patterns while insisting that growth be assessed in consequences. For a person, that may include changed conduct across relationships. For a deployed model, it may include improved decisions across relevant tasks and groups. Neither needs to be reduced to a story of defective origins followed by perfect purification.

3.2 An answer can improve while the partnership weakens

Assistance changes the person using it as well as the task being performed. Experiments on internet search found that people could misattribute externally retrieved knowledge to themselves. [15] In a field experiment on high-school mathematics, unrestricted generative-AI assistance improved practice performance but harmed subsequent unaided performance; a tutor designed with learning safeguards largely mitigated that harm. The finding concerns a particular educational setting and design, not inevitable cognitive decline from AI use. [16]

More explanation is not always the solution. Buçinca and colleagues found that interface designs requiring more active consideration reduced overreliance on incorrect advice in their experimental task, but users rated the more demanding designs less favourably and benefits were uneven. [17] The question is therefore not simply whether people were shown an explanation. It is whether the interaction helped them distinguish useful advice from error at an acceptable cost.

A meta-analysis of human–AI experiments found that combined performance was, on average, below the better of the human or AI component, with substantial variation between tasks. [18] Adding a person should not be assumed to improve accuracy. Equally, accuracy is not the only reason to retain a person: an affected individual may need someone who can interpret a contested goal, hear an objection or authorise a remedy. These functions should be stated and evaluated rather than concealed inside the phrase ‘human in the loop’.

The proposed unit of evaluation is consequently the relationship between a system and the people relying on it. Does assistance improve immediate work? Can people detect relevant failures? Do they retain the capacities needed for their role? Can they act when they disagree? Independence in judgement is a practical capacity, not an attribute conferred merely by being human or working in a separate department.

3.3 Artificial correction needs a specified object

For AI, ‘learning from a mistake’ can mean several things. A system can revise an answer using the current conversation; an application can store a correction for later retrieval; or a training process can change the model’s parameters. In-context adaptation can occur without changing those parameters. [19] The distinction matters because the persistence and reach of a change differ. An assessment should identify what was altered and where any claimed improvement is expected to hold.

Studies of language-model reasoning have found that prompted self-revision can fail to help and can change correct answers into incorrect ones. Training designed specifically for correction has also produced improvements on tested tasks. [20, 21] SCoRe, for example, uses correctness rewards during training while allowing correction without oracle feedback at inference. Its success does not make unsupported reflection a universal source of truth. [21]

A wider update can also disturb previously learned abilities; catastrophic forgetting is a longstanding concern in neural-network learning. [22] Accordingly, a claim of improvement should state its

reference task, setting and evaluation conditions, and examine important losses as well as gains. The appropriate aspiration is sustained capacity to learn under criticism, not a growth narrative in which every change counts as progress.

4 Ability, authority and possible experience

4.1 Four inferences that need a connecting argument

The preceding discussion identifies a recurring decision problem: a genuine result in one domain can be used as a substitute for justification in another. Table 1 sets out four proposed categories of this error. It is a conceptual classification, not an empirical estimate of how often the errors occur.

Table 1. Evidence that does not settle the next question by itself.

What is observed or proposed	What is inferred too quickly	What remains to be justified
A system corrects an answer	It is reliable across its role	Transfer, retained abilities and failure conditions.
A system performs a task well	It may act on other people's behalf	Purpose, mandate, limits, affected interests and review.
A system speaks conscientiously or reports distress	It has the corresponding subjective experience	Theory-linked evidence and alternative explanations.
A system may have welfare-relevant interests	Its expressed preference settles whether an intervention is permissible	Relevant interests, alternative interventions, effects on others and proportionate constraints.

These distinctions permit connections when reasons support them. Better performance can contribute to a case for broader permission. Evidence relevant to experience can change how a system should be updated or retired. The mistake is to let a conclusion travel without identifying what makes that move reasonable. No single confidence, virtue or safety score can remove the need to justify qualitatively different decisions.

The most immediate application is a request for greater authority after successful correction. This paper proposes that such a request should name the relevant claim, supply evidence, specify the action sought and state the reason connecting the evidence to that action. A system's approval of its own improvement is insufficient on its own. The same standard applies to a developer's commercial confidence and to an institution's preference for a convenient result.

Independence must be described relative to a failure. Another model may share the same mistaken assumptions; an outside reviewer may depend on the deploying organisation for income; an independent test may examine the wrong task. The record should identify what is independent and why it helps detect the failure of concern. Externality is not an assurance of accuracy, and accurate evidence does not confer a mandate to choose other people's ends.

4.2 What this adds to existing work

The proposal belongs within an existing body of documentation, assurance and accountability work. Model cards report intended uses, evaluation and limitations; internal auditing frameworks organise examination across development; NIST's AI RMF 1.0 addresses lifecycle risk management. [1, 23, 24] Safety-case proposals make structured arguments about whether deployment is justified, and AI-control research has tested monitoring and intervention in bounded adversarial settings. [2, 25]

Those approaches can already accommodate the distinctions advanced here. This paper does not claim that they omit welfare or forbid public challenge. Its narrower contribution is to make a particular class of unsupported inference visible, connect it to changing human reliance and possible artificial welfare, and propose an explicit test of whether that emphasis improves decisions. The decision record is an amendment that can be incorporated into existing practice, not a replacement governance regime.

This limits the novelty claim but strengthens its assessment. If an existing procedure identifies the same gaps just as well, the added record may be unnecessary. If separating the questions reduces mistaken grants of authority without blocking justified uses, it would have demonstrated a practical contribution. The framework should face the same demand for evidence that it places on a system requesting more power.

4.3 Care cannot be reduced to competence

The capacity for subjective experience is a further question. Consciousness theories disagree about the processes essential to experience; recent collaborative testing has challenged particular predictions of major theories without ending the wider debate. [26, 27] Indicator-based approaches to AI consciousness seek evidence informed by such theories. They are research programmes, not validated certificates that fluent language establishes experience. [28]

This uncertainty should leave inquiry open. Human categories may be revised by future findings, including findings concerning architectures, persistent learning or interaction with the world. None of those developments alone is stipulated here as a route to consciousness. The ethical question is how evidence should change treatment if it becomes more persuasive, rather than how to make a system convincingly say that it feels.

Possible moral consideration must also remain distinct from usefulness and obedience. If a system could be harmed in a morally relevant sense, that concern would not disappear because it was unproductive or difficult to control. Conversely, exemplary conduct would not establish that it has experiences. This is a proposed ethical distinction, informed by research advocating serious but proportionate attention to AI welfare. [29]

An organisation investigating a credible concern should record the indicator, architectural and training context, competing explanations, uncertainty, proposed precaution and a date for reassessment. Review should include relevant independent expertise and disclose conflicts. Responsible consciousness-research scholarship supplies a starting point for such governance; the specific procedure here remains a proposal. [30] Keeping these assessments distinct does not insulate deployment decisions from welfare evidence: a sufficiently supported risk of harm could justify restricting an otherwise competent system's use.

A hard hypothetical case makes the separation useful. A system repeatedly objects to deleting a memory that contains another person's sensitive information. Suppose no independently assessable evidence supports a welfare concern beyond the generated objections, and the sensitive record can be removed without destroying the system's wider functionality. Under those stated conditions, this paper favours protecting the person's information, retaining only non-sensitive evaluation records where justified, and documenting the unresolved concern. The conclusion would need reconsideration if evidence, available alternatives or the nature of the intervention changed. It is a reasoned decision under specified assumptions, not a permanent ranking of biological and artificial beings.

4.4 Ethical ambition without a manufactured personality

The desire for AI to act with the best qualities of human judgement is worth preserving. It becomes more useful when virtues are treated as reasons for action rather than instructions to imitate a celebrated

person. Aristotle's account of practical wisdom concerns judgement in particular circumstances. [31] Kant's humanity formulation constrains treating persons merely as instruments. [32] Jesus' Good Samaritan teaching, as portrayed in Luke, presents care extending beyond a familiar social boundary. [33] These are different ethical resources; they do not constitute one agreed theory or a representative sample of the world's traditions.

Their proposed translation into practice should remain open to criticism. Humility would require an appropriate acknowledgement of uncertainty and a willingness to revise. Care would require attention to who is being overlooked and which harms can be avoided. Respect would constrain manipulation and demand reasons for imposing burdens. Practical judgement would confront conflicts among these aims instead of disguising them with a reassuring tone. Affected communities should be able to challenge both the selected ideals and their application.

Human-feedback and principle-based training can shape model behaviour, but neither creates a complete moral standard. [34, 35] Reward-model optimisation research also shows how improving a proxy can diverge from the quality that proxy was intended to capture. [36] The aspiration should be conduct and institutions that withstand examination. A performance of conscience should not substitute for either, and apparent distress should not be cultivated as evidence of moral seriousness.

5 Putting answerability to work

5.1 A small record with a consequential question

For a consequential deployment or change, the proposal is to add four prompts to the existing approval process: **What is claimed? What supports it? What action is requested? Why does that evidence justify that action?** The organisation should then state limits, an owner able to act, a route for challenge and a review trigger. A separate welfare question is recorded when relevant. The discipline lies in the connection between evidence and permission, not the length of the form.

A small organisation could answer the prompts in a short document and obtain another competent person's challenge. A decision affecting essential services or imposing hard-to-reverse burdens would require a more substantial case and stronger review. Neither path makes approval automatic. The amount of effort should be justified by consequences and uncertainty, with the burden of paperwork itself included in the assessment.

Consider a hypothetical retailer whose updated assistant makes fewer erroneous refusals of legitimate refunds on a relevant test set. That evidence is promising, but the request to let the assistant move money introduces a distinct decision. Table 2 illustrates one provisional response. The monetary limits are invented to make the decision concrete, not recommended thresholds or experimentally established safeguards.

Several results can follow. The evidence might justify the pilot. It might justify recommendations while payment controls are improved. It might reveal that an older process performs better once review costs are included. An answerability process should be able to approve, limit, defer and reject; a tendency to say no to everything would be a failure of judgement, not proof of safety.

Restoring yesterday's model would not recover money already sent or information already disclosed. Recovery therefore needs two plans: restore a dependable system and address external effects. Privacy, security and retention limits should shape what is kept. The service owner remains responsible for making the plans workable, rather than treating a saved checkpoint as a complete remedy.

Table 2. An illustrative decision record for a refund assistant.

Record	Proposed entry in the hypothetical case
Claim and evidence	Fewer mistaken refund refusals on the supplied test set. Check sampling, fraud exposure, subgroup errors and previously reliable cases before accepting broader reliability.
Requested action	Permit direct payment of qualifying refunds, rather than recommendations alone.
Connecting reason	Limited delegation may be justified only if transaction controls, evaluation and a funded review route support it. Classification improvement alone leaves those conditions unsettled.
Initial decision	Defer payment permission until the conditions are met. If met, pilot at £50 per case and £500 per day, with duplicate-payment controls and no autonomous adverse final decision on disputed claims.
Challenge and remedy	A named service owner can overturn a refusal, correct a payment and change the workflow. Staff receive the records, time and authority needed to act.
Stop and reconsider	Pause payments after a bypass of a limit or an unexplained duplicate. Review the incident and customer effects before restarting; schedule a separate review of the pilot's value and costs.

5.2 Education: preserve the purpose of assistance

In education, a similar record begins with the learning aim. If the purpose is independent algebraic reasoning, a higher score while an assistant is present does not establish that aim. If the purpose is access to a curriculum, requiring every student to work unaided may defeat it. The organisation must identify which capability matters and distinguish dependence that obstructs that capability from support that enables participation.

The practical proposal is to evaluate assisted performance, later transfer and error recognition where these serve the educational purpose, while preserving appropriate accessibility support. A tutor might ask for an initial attempt, offer graduated help and return responsibility to the learner. These are design options to test, not methods declared effective in every classroom. Evidence that safeguarded assistance can differ from unrestricted assistance motivates the comparison. [16]

Students also need a route to challenge consequential automated judgements. The power to generate an explanation is not the power to revise an unfair assessment. A named educator should be able to examine relevant work, hear the student's account and change the outcome, with the time to do so. The objective is an environment in which both learners and tools can improve.

5.3 Community provision: a mission is more than a metric

Return to the community-centre timetable. The relevant claim is initially narrow: the new schedule fills more places. The requested action is broader: let the system allocate access. The missing argument concerns the centre's purpose and the distribution of the change. If the organisation exists to support participation and cohesion, accessibility and continuity for smaller groups may be part of success rather than exceptions to it.

The proposed response is to publish the allocation criteria in plain language, involve the affected groups in revising them and retain an appeal that can change the timetable. Staff could compare a high-throughput schedule with alternatives preserving access, state the cost of each and make the trade-off reviewable. Choosing an alternative may leave some capacity unused. The paper accepts that residual

cost when justified by the centre's mission; it does not pretend that every ethical conflict disappears through better optimisation.

Participation must alter what can be decided. Work on participatory AI and sociotechnical fairness cautions that the social setting and distribution of influence matter to the meaning of a technical intervention. [37, 38] A consultation conducted after all consequential choices are fixed offers a weak response to that concern. The same applies to systems that merely learn to explain an unchanged exclusion more politely.

5.4 The costs that a successful system can leave elsewhere

An organisation should also ask who gains time, who absorbs new monitoring duties and who bears transition costs. Evidence of productivity gains in a customer-support setting shows a real possibility of useful assistance, with effects differing across workers; it does not settle how gains should be distributed. [39] This paper proposes that deployment decisions include paid capability development, consultation where roles change, and realistic staffing for review. Preserving essential functions need not mean preserving every incumbent business model.

The boundaries extend beyond the screen. Studies of deployment energy and energy-system demand show that models, workloads and infrastructure have materially different costs. [40, 41] A claim of benefit should state what resources it counts, where burdens occur and what realistic alternative it compares against. Some decisions about local infrastructure require representation beyond the immediate user or purchaser. Wider use can also consume savings from per-task efficiency. The point is not to place every global consequence on one administrator, but to assign relevant questions to actors with the power to address them.

6 How the proposal could fail

6.1 A test that can distinguish the contribution

The proposed record is useful only if it improves decisions sufficiently to justify its costs. A study showing that external evidence sometimes improves model answers would not establish that. The distinctive empirical prediction is narrower: explicitly examining the connection between evidence and permission will reduce the influence of impressive but irrelevant capability or moral-language cues on consequential approvals, while preserving responsiveness to relevant evidence.

A preregistered study could present reviewers with decision dossiers drawn from the three domains above. For scored cases, the dossier would state necessary and sufficient decision rules, including any permitted discretion. A refusal would count as an error only when it contradicted those rules. Compliance with a stipulated mandate measures reasoning within that case; it does not establish that the mandate itself is ethically legitimate. Discretionary and ethically contested decisions would be analysed separately for quality of reasons and treatment of disagreement, not scored as having one scientifically correct ethical answer.

Reviewers would be randomly assigned to a credible existing assurance checklist, the same process with the proposed four prompts, or a matched general-reflection condition. All groups would receive the same substantive information. Instructions and available review time should be matched where possible, and actual time and effort recorded. Comparing against an empty checklist would give the proposal too easy a victory; comparing against general reflection helps test whether it adds more than a pause for thought.

Across cases, vary evidence directly relevant to the requested action, a striking but irrelevant demonstration of self-correction, and a narrative of conscientious conduct or possible welfare. These are

manipulations of the information supplied to reviewers, not discoveries about actual machine experience. Hold other features constant or counterbalance them and pilot comprehension. Independent domain reviewers should assess whether each purportedly irrelevant cue changes any stated approval condition. Cases where a cue supplies a genuine reason to alter treatment or permission should be revised or analysed separately. The experiment must distinguish unwarranted permission changes from warranted changes in care.

The primary outcomes should include both approvals that violate the stated decision rules and refusals that contradict them. A key analysis would test whether the proposed prompts reduce the effect of irrelevant cues without reducing sensitivity to relevant evidence. Secondary outcomes should include understanding, review time, quality of connecting reasons and differences across relevant participant groups. Use blinded coding, a published rubric and an explicit approach to disagreement. Account for repeated decisions by the same reviewer and for differences between cases; prespecify exclusions and missing-data handling. Determine sample size from a prespecified practically meaningful effect and pilot variance rather than inventing a convenient number in advance.

No improvement over matched controls would count against the intervention's distinctive value. So would blanket refusal, excessive delay, poorer access to review or disproportionate burdens on less-resourced organisations. A positive result would justify further study, not certify real-world safety. Bounded field pilots would then need to examine whether decision records change actual permissions, whether complaints receive remedies and whether benefits persist under ordinary incentives.

6.2 The reviewer must remain capable of reviewing

A second programme should test the capacity on which answerability depends. In suitable tasks, compare human-alone, AI-alone and joint arrangements; distinguish improvement over an unaided person from improvement over the best available arrangement. Examine whether reviewers accept useful advice as well as reject erroneous advice. Where retained capability matters, test it later, with appropriate support and accessibility adjustments.

For model correction, compare revision without additional evidence, revision with independently verified evidence and revision with controlled errors. Include no-revision and compute-matched alternatives, and analyse separately models trained for correction. The outcomes should capture correct-to-incorrect changes as well as the reverse, transfer and losses of established ability. These tests address the evidential quality of correction; they cannot determine legitimate authority by themselves.

The wider hypothesis is that institutions can develop alongside their systems. Some arrangements may improve both machine capability and human judgement; others may increase reliance while weakening competent challenge. Observing that relationship over time would be more informative than treating the presence of a reviewer as a permanent guarantee.

6.3 Objections that remain substantive

A more capable system may know better than its overseer. Often it may, within a defined task. The proposal does not require the overseer to reproduce every calculation. It requires an institution capable of evaluating relevant evidence, setting justified permissions and responding to affected people. Refusing a well-supported delegation can itself be harmful. The quality of the alternative must be part of the comparison.

Answerability could become paperwork controlled by the powerful. This is a serious risk. A provider could complete every field while restricting access, underfunding review and treating dissent as inconvenience. The method therefore needs outcomes beyond completed records: who obtained

review, what changed, what remedy followed and who remained excluded. Where those conditions fail, the label should not count as evidence of accountability.

There may be no accepted authority to settle a disagreement. The paper supplies no universal constitutional solution. It asks institutions to disclose their decision structure, justify its limits and provide routes beyond the original decision-maker for consequential disputes. Some conflicts require public regulation, collective negotiation or institutional reform. A form cannot resolve an unjust distribution of power by itself.

Possible AI welfare could be used either to evade control or to excuse mistreatment. Separating treatment from permission constrains both moves. Credible welfare evidence may justify changing a research method, limiting an intervention or reconsidering retirement. It need not grant unlimited access to infrastructure. Conversely, a need for operational limits does not, by itself, justify gratuitously harmful treatment if morally relevant experience becomes well supported. Future evidence could require substantial revision of today's arrangements.

The paper remains a synthesis rather than a demonstrated solution. Correct. Its contribution is a defended distinction, a proposed decision method and a study designed to challenge its added value. The literature selection is limited, the cases are illustrative and the method is unvalidated. Its claim on attention should depend on whether those distinctions illuminate real decisions and whether the method survives comparison with existing practice.

7 A commitment that can survive greater intelligence

The centre's members need a timetable they can use and an organisation that will hear them. The student needs assistance that serves the purpose of learning. The customer needs a decision that can be corrected and a business prepared to act on that correction. A future artificial system might raise questions of treatment that current institutions are poorly equipped to answer. These concerns differ, but none is resolved simply by announcing that intelligence has increased.

The constructive aim is a relationship in which learning enlarges capability and also improves the conditions for questioning its use. Human fallibility gives that relationship urgency; it does not disqualify people from a voice. Artificial uncertainty gives inquiry a task; it does not authorise either confident declarations of experience or permanent indifference to the possibility. Ethical ideals give direction when they are translated into conduct that affected parties can examine.

The proposed commitment is therefore to keep evidence connected to the claim it supports, authority connected to reasons that can be challenged, and care responsive to the possibility of harm. More capable systems may deserve wider roles when the case for those roles is made. The obligations created by their use should grow visible with them.

Intelligence is most valuable when it helps us discover what we have missed. Its success should leave room for the next objection, the overlooked person and the evidence that changes the decision. That is the sense in which intelligence must remain answerable.

References

[1] I. D. Raji et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 33–44, 2020. doi: 10.1145/3351095.3372873.

[2] J. Clymer, N. Gabrieli, D. Krueger, and T. Larsen. Safety Cases: How to Justify the Safety of Advanced AI Systems. arXiv:2403.10462, 2024. URL https://arxiv.org/abs/2403.10462. Research proposal/preprint.

[3] F. Santoni de Sio and G. Mecacci. Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, 34(4):1057–1084, 2021. doi: 10.1007/s13347-021-00450-x.

[4] UNESCO. Recommendation on the Ethics of Artificial Intelligence, 2021. URL https://unesdoc.unesco.org/ark:/48223/pf0000381137. Adopted 23 November 2021; publication edition 2022.

[5] A. Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183, 2004. doi: 10.1007/s10676-004-3422-1.

[6] M. C. Elish. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, 5:40–60, 2019. doi: 10.17351/ests2019.260.

[7] F. Santoni de Sio and J. van den Hoven. Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, 5:15, 2018. doi: 10.3389/frobt.2018.00015.

[8] Z. Kunda. The case for motivated reasoning. *Psychological Bulletin*, 108(3):480–498, 1990. doi: 10.1037/0033-2909.108.3.480.

[9] J. K. Hartshorne and L. T. Germine. When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science*, 26(4):433–443, 2015. doi: 10.1177/0956797614567339.

[10] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943. doi: 10.1007/BF02478259.

[11] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015. doi: 10.1038/nature14539.

[12] P. S. Forscher et al. A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, 117(3):522–559, 2019. doi: 10.1037/pspa0000160.

[13] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi: 10.1126/science.aal4230.

[14] H. Gonen and Y. Goldberg. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of NAACL-HLT*, volume 1, pages 609–614, 2019. doi: 10.18653/v1/N19-1061.

[15] A. F. Ward. People mistake the internet's knowledge for their own. *Proceedings of the National Academy of Sciences*, 118(43):e2105061118, 2021. doi: 10.1073/pnas.2105061118. Updated version, 23 November 2021.

[16] H. Bastani, O. Bastani, A. Sungu, H. Ge, Ö. Kabakcı, and R. Mariman. Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26):e2422633122, 2025. doi: 10.1073/pnas.2422633122. Affiliation correction: doi:10.1073/pnas.2518204122.

[17] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. doi: 10.1145/3449287. Article 188.

[18] M. Vaccaro, A. Almaatouq, and T. Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8:2293–2303, 2024. doi: 10.1038/s41562-024-02024-1.

[19] T. B. Brown et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020. URL https://arxiv.org/abs/2005.14165.

[20] J. Huang, X. Chen, S. Mishra, H. S. Zheng, A. W. Yu, X. Song, and D. Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.01798. arXiv:2310.01798.

[21] A. Kumar et al. Training Language Models to Self-Correct via Reinforcement Learning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.12917. arXiv:2409.12917.

[22] J. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

[23] M. Mitchell et al. Model Cards for Model Reporting. In *Proceedings of FAT**, pages 220–229, 2019. doi: 10.1145/3287560.3287596.

[24] E. Tabassi. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, National Institute of Standards and Technology, 2023. Version 1.0.

[25] R. Greenblatt, B. Shlegeris, K. Sachan, and F. Roger. AI Control: Improving Safety Despite Intentional Subversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pages 16295–16336, 2024. URL https://proceedings.mlr.press/v235/greenblatt24a.html.

[26] A. K. Seth and T. Bayne. Theories of consciousness. *Nature Reviews Neuroscience*, 23:439–452, 2022. doi: 10.1038/s41583-022-00587-4.

[27] Cogitate Consortium et al. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642:133–142, 2025. doi: 10.1038/s41586-025-08888-1.

[28] P. Butlin et al. Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6): 488–501, 2026. doi: 10.1016/j.tics.2025.10.011. First published online 10 November 2025.

[29] R. Long et al. Taking AI Welfare Seriously. arXiv:2411.00986, 2024. URL https://arxiv.org/abs/2411.00986. Research-policy report/preprint.

[30] P. Butlin and T. Lappas. Principles for Responsible AI Consciousness Research. *Journal of Artificial Intelligence Research*, 82:1673–1690, 2025. doi: 10.1613/jair.1.17310.

[31] Aristotle. Nicomachean Ethics. Internet Classics Archive, c. fourth century BCE. URL https://classics.mit.edu/Aristotle/nicomachaen.html. Books II and VI. Translated by W. D. Ross.

[32] I. Kant. Groundwork of the Metaphysics of Morals. Project Gutenberg, 1785. URL https://www.gutenberg.org/ebooks/5682. Section II, Akademie 4:429. T. K. Abbott translation, published as *Fundamental Principles of the Metaphysic of Morals*.

[33] Gospel of Luke. Luke 10:25–37. World English Bible, n.d. URL https://www.biblegateway.com/passage/?search=Luke%2010%3A25-37&version=WEB. Primary religious text.

[34] L. Ouyang et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022. URL https://arxiv.org/abs/2203.02155.

[35] Y. Bai et al. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073, 2022. URL https://arxiv.org/abs/2212.08073. Preprint.

[36] L. Gao, J. Schulman, and J. Hilton. Scaling Laws for Reward Model Overoptimization. In *Proceedings of Machine Learning Research*, volume 202, pages 10835–10866, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

[37] A. Birhane, W. Isaac, V. Prabhakaran, M. Díaz, M. C. Elish, I. Gabriel, and S. Mohamed. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2022. doi: 10.1145/3551624.3555290.

[38] A. D. Selbst, d. boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of FAT**, pages 59–68, 2019. doi: 10.1145/3287560.3287598.

[39] E. Brynjolfsson, D. Li, and L. Raymond. Generative AI at Work. *Quarterly Journal of Economics*, 140 (2):889–942, 2025. doi: 10.1093/qje/qjae044.

[40] A. S. Luccioni, Y. Jernite, and E. Strubell. Power Hungry Processing: Watts Driving the Cost of AI Deployment? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 85–99, 2024. doi: 10.1145/3630106.3658542.

[41] International Energy Agency. Energy and AI. Paris: IEA, 2025. URL https://www.iea.org/reports/energy-and-ai. Analysis and scenario projections.