Building Answerable AI

Why Automation Needs Owned Error

Luke F. Walton

Independent Researcher · lukefwalton.com

ORCID: 0009-0005-9263-1954

Working paper · June 2026

Working paper. This note states the central claims and framing of a paper in preparation; the full development is forthcoming alongside its companion papers. It is circulated to place the claims on record.

Companion papers: *The Decision No One Authored* (the special case; Walton 2026b), *The Invariant of Answerability* (the general claim; Walton 2026c), and *The Captured Oracle* (the live demonstration; Walton 2026a).

Abstract

What should a builder do under an obligation that survives every way of building around it? Companion papers argue that wherever an action reaches a party, an account is owed, and that no routing through machinery defeats the owing (Walton 2026b, 2026c). Three responses fail it: denial is refuted, pausing defers a reckoning it cannot perform, and to strand the account while building on is the captured channel under another name (Walton 2026a). This article defends the fourth response: construct, inside the owing, the relation that discharges it, and do so at speed. The thesis is that answerability is not only a constraint on automation; it is an enabling condition for automation that can compound without drifting. The note names the relation to be built, participational answerability; states the six conditions, drawn from and corroborated by recognized literatures, under which it self-organizes and the two-layer structure they compose; and specifies the artifact, framed automation, in which the conditions become architecture: the junctures at which an evaluative frame enters force are held by an answerable party, everything that merely executes a held frame is automated, influence that emerges from a learned representation rather than an authored frame is exposed rather than held, and the operation that would strand an account is made unavailable as an ordinary operation of the system. The comparative claim is stated with its warrant divided: the direction is supplied by a published asymmetry, that influence which is true

and owned loses nothing by being seen while a covert frame's advantage is concealment-dependent (Walton 2026a); the magnitude is an empirical commitment the article accepts and this note does not discharge. The claim concerns the build, not the field's wider hazards; whether those call for governing or slowing on their own grounds is a question this program does not join. A full treatment is in preparation; this note places the framing on record.

1. The fourth response

The owing is real, live, and route-indefeasible (Walton 2026a, 2026b, 2026c): wherever an action reaches a party, an account is owed, and no routing defeats it. Three responses to that fact fail. To deny it is refuted. To pause is to defer the reckoning without discharging it, and the verdict is scoped to exactly that: whether the field's wider hazards call for governing or slowing on their own grounds is a question this paper does not join; what no pause does, whatever else it buys, is construct the relation an owing requires. To strand-and-go — to externalize the account and keep moving — is the captured channel under another name. The fourth response is to construct, inside the owing, the relation that discharges it, and to do so at speed. "Construct inside the owing" carries no metaphysical ambition: it does not defeat or dissolve the owing, which cannot be defeated. It means installing and preserving a fit respondent — a party who holds the evaluative frame and can answer for it — where the route would otherwise leave none.

The claim is not that the answerable build knows the good in advance. In the domains at issue here, the good is often contested, contextual, and revised under pressure; the point is not calculator-like correctness. The more modest premise is that badness recurs: harm appears, standards slip, influence hides, explanations fail, and costs are sent downhill. A system that strands those failures may preserve margin, but it loses the only material from which correction is made. An answerable system does not become perfect. It keeps error attributable enough to be answered for, revised, and remembered.

Two questions follow: whether such a relation can be built, and what the built thing is. This note states the answers; the article in preparation argues them.

2. Participational answerability

Call the relation to be constructed participational answerability: a standing arrangement in which the parties who shape what is done remain, together, answerable for it, and in which the account can be reached and met rather than dispersed. What discharges a route-indefeasible owing is a fit respondent answering, not the account migrating to a convenient payer: a false settlement — the account marked paid by a party who cannot pay it. Distributed answerability is still answerabil-

ity; what converts distribution into the diffusion that strands an account is the absence, at some juncture, of any answerable hand at all.

The claim placed on record is that the arrangement self-organizes under six conditions, and that the conditions are engineerable. Shared stakes, because parties answer, and submit to being held, only where the outcome reaches them too; absent it, the shaping party answers to no one, since the cost falls elsewhere. Legibility, because a party can be held only to what is visible as someone's doing; absent it, the unauthored composite, with no one to address. Proximity, because the demand must be able to reach the respondent; absent it, the respondent put out of reach through a built door. Repeated play, because holding-to-account is a standing liability to be re-asked, and a practice in which that occasion recurs disciplines itself; absent it, the practice that would reliably discharge never forms, and every account waits on enforcement. Exit rights, because an account extracted from a party who could not have walked away is submission, not answering; absent them, the practice cannot tell endorsement from duress. And backstop enforcement, because an owing whose discharge depends on the wrongdoer's willingness is, against the strong, an owing with no path; absent it, the account falls on the weakest, the body tendered as the author. The first five compose a participation layer on which answerability polices itself; enforcement and recognition compose a backstop that holds where self-policing runs out, the supplement and not the source. The conditions are drawn from, and corroborated by, the literatures on commons governance, the evolution of cooperation, and non-domination (Ostrom 1990; Axelrod 1984; Pettit 1997); what is claimed here is their reading as the conditions of answerability in particular, and the two-layer structure they compose.

3. Structural inexpressibility

The engineering claim is sharper than buildability, and it is the note's central stake. The nearest constructive neighbor, meaningful human control, names design conditions under which control over an automated system and the attribution of responsibility for it are preserved (Santoni de Sio and van den Hoven 2018); everything in that is conceded, and the delta is the direction of the claim: where the neighbors secure responsibility as a constraint design must satisfy — the system kept responsive to the relevant human reasons and traceable to a capable party — the pattern stated here makes the unauthored move structurally inexpressible at the interface. The claim is not metaphysical or total-system impossibility. It is that the strandable operation is not available as an ordinary authored operation of the system: the interface offers no ordinary operation by which an unauthored synthesis could be entered, the way a form with no signature line cannot take an unsigned submission as an ordinary entry. The constraint is paid once, at design time, by the shape of what can be expressed, not per decision by a monitor who might miss. Safety engineering has long held the same ordering as the hierarchy of hazard controls: elimination and engineering controls

outrank administrative controls and protective equipment precisely because a hazard designed out needs no monitor to catch it (NIOSH). And the determined actor who goes around the interface concedes the mechanism: going around is itself an authored, legible act, a default overridden as someone's doing and recorded as such. Circumvention does not evade authorship; it performs it in the open.

4. Framed automation

The artifact the program specifies is framed automation, and it can be stated as a pattern in one sentence: hold the junctures, automate the execution, and let failure write the interface. The held points are the five junctures the first companion identified as the sites where authorship of an evaluative frame is most easily lost (Walton 2026b). The ends a system serves are a stated, owned commitment, someone's to revise. The standards its outputs meet are authored tolerances, never inferred from what has lately been let through. Verification is owned: the measurements run at machine speed, but a party answers for what the tests test. Acceptance is the signature line: an output enters action as someone's acceptance, or it does not enter action. And the final form is answerable as released: what reaches an affected party carries its owner with it.

Everything that merely executes a frame already held may be automated; the stretch between two held points originates no frame of its own. A stretch that begins to set what counts is no longer between junctures; it is one, and the holding moves to it, so long as the setting is an authored act a party can hold. Not all setting is. Ownership is stratified. What is fixed by an authored judgment is held at a juncture. What is fixed by an editorial selection, which sources the system may draw on at all, is owned by attribution: a named party chose the boundary, and the choice is contestable as a judgment. But what is fixed by behavior emerging from a learned representation, which evidence the system surfaces for a given phrasing and which it leaves aside, has no sentence to author and no juncture for the holding to move to; it is testable, tunable, and partially correctable, but not ownable the way a standard in a file is. What answerability asks of it is not holding but exposure: the system surfaces what it passed over, so that a party who never complained can contest what the representation buried. The retrieval shadow names the case. The cited list shows what was used; answerability wants, beside it, an account of what was not.

Speed lives in the automated layer: the pattern does not ration automation; it places it, and what it asks of the build is not more attention but differently positioned attention, the owned acts few, dense, and recurring where frames enter force, and absent everywhere else. Autonomy, on this pattern, is not a quantity a system has more or less of; it is a placement: everything between the junctures is exactly as autonomous as engineering can make it; the junctures themselves, never. The position that fully autonomous agents should not be developed (Mitchell et al. 2025) is thereby

neither contested nor accepted in its own terms. Framed automation is what "not fully autonomous" looks like when built rather than banned.

Nor is the set of inexpressible operations decreed from above, which would be the arbitrary power the conditions exist to rule out. It is discovered: each failure in which an account stranded is a demonstration of a move that should not have been expressible, and the response is to close that move at the interface so it cannot recur as an ordinary operation. The inexpressible set is the accumulated record of where answerability broke, revisable as the practice teaches more. The failures that teach are caught and discharged, not consumed; the affected party's account is met, and the design signal is the residue, never an unpaid debt the loop feeds on. And discovery can run ahead of injury: staged failure, against no one, generates the signal synthetically.

Inexpressibility and exposure divide one labor by what the interface can reach: the strandable move it can decline to offer is foreclosed at design time; the representational influence it cannot decline, because that influence is a property of how evidence is surfaced and not an operation the interface exposes, is instead made visible enough to be contested. The first removes a move from what can be done; the second accounts for influence that remains even when no forbidden move was made.

5. The rate claim, stated not argued

The comparative claim is stated here and argued in the article. Its premise is the first section's: error is not exceptional, so the comparison that matters is never erring against not erring but what becomes of error as it accumulates. That accountability rightly structured enables sound practice rather than taxing it has a long pedigree (O'Neill 2002); the mechanism claimed here is the program's own. Under efficiency pressure, operating points migrate toward the boundary of acceptable performance (Rasmussen 1997), and a route that drops the name, the record, and the interruptible person is cheaper at every transaction; the cost avoided compounds into margin. The stranded build compounds that margin and, beneath it, unregistered error, since with no author there is no operative difference between the frame being wrong and the frame having shifted. Drift, in this paper's sense, is error with no owner; the named study is the normalization of deviance, in which intact returns were read as confirmation until the record itself had quietly set a limit no one authored (Vaughan 1996). The answerable build compounds corrections: an error with an owner is locatable, answerable, and revisable, and each discharged failure is a revision the frame keeps. The warrant divides exactly. The direction is supplied by *The Captured Oracle*'s published asymmetry, that owned influence loses nothing by being seen while a covert frame's advantage is concealment-dependent (Walton 2026a); the magnitude, how fast correction compounds into capability, is an empirical commitment the article accepts and this note does not discharge. The first may win locally. Only the second can improve under accumulated error.

6. What the full paper argues

What this note places on record is the framing: the fourth response, the relation and its six conditions, the two-layer structure, structural inexpressibility as the engineering turn, framed automation as the pattern, and the comparative claim with its warrant divided. What it withholds for the article is the argument: the conditions argument in full, with the structural reason each condition is necessary for answerability in particular; the convergence result, the mechanism by which corrections compound, tested against an evidence class drawn from organizational learning, high-reliability practice, and software delivery; the development of the exposure layer, including what counts as adequate exposure of a representational frame and the interface that would carry it; the placement against the accountability and autonomy literatures at length; and the threshold at which the argument leaves the builder. The article in preparation supersedes this note; the note exists so that the framing, the names, and the claims carry a date.

Disclosures

Competing interests. The author is the founder of Surmado, Inc., which builds managed AI systems of the kind the pattern describes; the criterion this paper states binds the author's own layer. The paper is written in a personal capacity, without employer involvement and without funding.

Generative AI. Generative AI systems (ChatGPT 5.5, OpenAI; Claude Opus 4.8 and Claude Fable 5, Anthropic; Gemini 3.5, Google) were used for drafting assistance, editorial review, and reference checking under the author's direction. The claims, the argument, and the final text are the author's, and the author takes full responsibility for the work.

References

Axelrod, Robert. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Mitchell, Margaret, et al. 2025. "Fully Autonomous AI Agents Should Not Be Developed." arXiv:2502.02649.

NIOSH (National Institute for Occupational Safety and Health, CDC). *Hierarchy of Controls*. https://www.cdc.gov/niosh/hierarchy-of-controls/ (accessed June 2026).

O'Neill, Onora. 2002. *A Question of Trust: The BBC Reith Lectures 2002*. Cambridge: Cambridge University Press.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Pettit, Philip. 1997. *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press.

Rasmussen, Jens. 1997. "Risk Management in a Dynamic Society: A Modelling Problem." *Safety Science* 27(2–3).

Santoni de Sio, Filippo, and Jeroen van den Hoven. 2018. "Meaningful Human Control over Autonomous Systems: A Philosophical Account." *Frontiers in Robotics and AI* 5: 15.

Vaughan, Diane. 1996. *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*. Chicago: University of Chicago Press.

Walton, Luke F. 2026a. "The Captured Oracle: Authorship and Agency in the Ethics of Answer-Engine Optimization." Preprint. Zenodo. https://doi.org/10.5281/zenodo.20676328.

Walton, Luke F. 2026b. "The Decision No One Authored: The Answerability Gap in Generative AI." Preprint, v1.4. Zenodo. https://doi.org/10.5281/zenodo.20622946.

Walton, Luke F. 2026c. "The Invariant of Answerability." Working paper. Zenodo. https://doi.org/10.5281/zenodo.20606493.