WEKID™ Companion Paper — **Model Two Extensions (v1.3): Governing the Human in the Loop**

By James Madigan Judge

Version 1.3 (Companion Note — Model Two Extensions: overseer engagement, joint gating, calibration objective, decision-maker alignment, and scope boundary) July 2026

Document Control

Document	WEKID™ Companion Paper — Model Two Extensions (v1.3)
Companion to	WEKID™ Master Whitepaper, Version 1.3 (July 2026)
Author	James Madigan Judge · WEKID LLC
Status	Approved extensions for Version 1.3; Submitted as proposals to the WEKID community RFC process. The invariants fixed in §4.5 of the Master Whitepaper are unchanged.
External reference	Athey, Bryan & Gans, "The Allocation of Decision Authority to Human and Artificial Intelligence," AEA Papers and Proceedings 110 (2020), building on Aghion & Tirole (1997).

Executive Summary

Consistent with WEKID's status as an open framework, this paper is being presented through the community Request for Contribution (RFC) process and will be ratified only through that process as part of the Standard Cohort.

The proposal responds to independently published, peer-reviewed research on the allocation of decision authority between humans and AI — principally Athey, Bryan and Gans (2020), building on Aghion and Tirole (1997). That work corroborates the founding discipline of WEKID's AI Decision Authority Model (Model Two) — that capability and authority are different things — and, in doing so, isolates one determinant that the WEKID Master Whitepaper, Version 1.2.1 ("Version 1.2.1") does not yet formalize: the incentives and engagement of the human who holds or oversees a decision.

This note previews five additive extensions to Model Two, proposed for Version 1.3: (A) overseer engagement as a governed signal; (B) joint gating of autonomy on maturity and engagement; (C) an explicit calibration objective; (D) decision-maker alignment in the stakes floor; and (E) a stated non-adversarial-principal scope boundary. Together they take the determinant the external literature adds — the incentives and engagement of the human — and express it in WEKID's native terms, as governed, auditable signals that condition authority through the same Trust Bridge and governance cycle already defined. The framework's answer to "the better the AI, the less the human watches" is therefore not a rebuttal but an instrument: measure the watching, and let delegation depend on it.

Nothing here weakens the floor. The Trust Bridge still orders maturity before authority, the two determinants still resolve to the more conservative, and AL-4 decision classes remain reserved to accountable human judgment regardless of how mature the AI's contribution may be.

Introduction

The separation of capability from authority is a founding discipline of the WEKID Framework AI Decision Authority Model (Model Two): what a system is *able* to do and what it is *permitted* to decide are governed as two different things. That discipline finds independent corroboration in the economics of decision-authority allocation, most directly in Athey, Bryan and Gans (2020), who show within a formal principal–agent model that a statistically superior AI is not automatically the one to which a rational organization should delegate the decision. WEKID welcomes this result: it is an external, formal proof of a principle the framework already treats as normative.

That same literature also identifies one determinant that Version 1.2.1 does not yet formalize — the **incentives and engagement of the human who holds or oversees the decision**. Where an AI-enabled system is capable enough to be trusted, the human assigned to supervise it has a rational tendency to reduce effort, a phenomenon the source terms "falling asleep at the wheel." Because WEKID's Supervised (AL-2) and Augmented (AL-3) levels presuppose an engaged human, this determinant bears directly on the integrity of the authority ladder. The five extensions previewed below are additive and signal-driven; none alters the invariants fixed in Section 4.5 of the Version 1.2.1, the ordering enforced by the Trust Bridge, the conservatism of the two-determinant resolution, and the reservation of AL-4 decision classes to human authority.

Proposal: The Case for modification to the WEKID Standard

WEKID is designed to be tested. A governance standard that asks organizations to stake legal, safety, and reputational risk on its verdicts must be willing to meet the strongest external scrutiny and to change when that scrutiny is right. This companion paper addresses AI decision-authority, published by Athey, Bryan and Gans (2020), supplies exactly the scrutiny required to maintain a standard. It addresses a formal model, developed with no reference to WEKID, that both corroborates WEKID's core separation of capability from authority and exposes a determinant the framework had not yet made explicit. Treating such research as an input to be governed, rather than a critique to be deflected, is the point of the exercise.

That posture is only credible if the framework is open. WEKID is therefore stewarded as an open framework with a community Request for Contribution (RFC) process organized around three governance surfaces: a stable **Core** — the five epistemic layers and the two models, held constant so others can teach and build on them; a **Calibration Layer**, where domain working groups shape versioned, sector-specific risk profiles; and an open **Evidence Corpus** of documented AI-failure cases and reusable control patterns. Independently published research enters through this structure — as evidence in the corpus, and as proposals the Standards Council can triage to the appropriate surface, open for public comment, and issue as a versioned release. The five extensions previewed here are submitted on exactly those terms.

Keeping WEKID open is not incidental to its purpose; it is a precondition of the trust it seeks to govern. A standard intended to outlast any single vendor or organization can earn adoption only if its evolution is transparent, attributable, and community-reviewed rather than dictated unilaterally — which is why contributors keep authorship credit under a contributor license agreement, why recognition is public and attributed, and why the long-term intent is to move stewardship of the Core to a neutral body. This paper is both a worked example of that process — an external research result converted into concrete, auditable

proposals — and an invitation to challenge, refine, or extend those proposals through it. The current governance surfaces and RFC stages are described at https://wekid.org/community.html#surfaces.

1. Overseer Engagement as a Governed Signal

Version 1.2.1 already makes authority dynamic: declining Maturity Scores or repeated gating events trigger automatic demotion down the authority ladder (Sections 4.6, 5.6). Version 1.3 extends the monitored signal set from the AI alone to the **human review loop**. A grant of AL-2 or AL-3 is treated as valid only while the human oversight it assumes is demonstrably exercised — meaningful human control, instrumented rather than presumed. The governance cycle gains a second class of demotion trigger alongside epistemic drift: oversight decay.

Representative engagement signals (calibrated per deployment)

- **Review latency and dwell.** Time the human actually spends before releasing a staged action, relative to the action's complexity — near-zero dwell on consequential releases is a rubber-stamp signal.
- **Override and disagreement rate.** The frequency with which the human amends, defers, or rejects the AI's proposal; a rate collapsing toward zero as AI maturity rises is the operational signature of over-reliance.
- **Sampling depth under AL-3.** For human-on-the-loop regimes, the proportion and risk-stratification of autonomously executed actions actually reviewed after the fact, against the sampling policy of record.
- **Escalation responsiveness.** Whether flagged or low-confidence items are genuinely adjudicated or reflexively approved.

These signals are emitted as structured, timestamped evidence in exactly the manner Section 6.4 already prescribes for AI-side verdicts, so they are auditable and feed the same GRC stack. The design intent is precise: the objection that a capable AI lulls its supervisor into disengagement is answered not by argument but by *measurement* — an AL-3 grant that lapses the moment the human loop stops functioning.

2. Joint Gating of Autonomy on Maturity and Engagement

Version 1.2.1 gates delegation on demonstrated epistemic maturity: sustained, gate-free maturity admits AL-3 (Section 4.4, the maturity ceiling). Version 1.3 makes the highest delegable level conditional on *two* sustained conditions rather than one — maturity *and* demonstrated overseer engagement. This closes a specific failure the decision-authority literature predicts: an AI so capable that it is "too good to watch" would, under a maturity-only ceiling, graduate smoothly to higher autonomy at precisely the moment its human supervision is most likely to have hollowed out. Under joint gating, capability alone cannot purchase autonomy; the human loop must be shown to be intact. The precedence rules of Section 8.9 are extended accordingly: an engagement-decay condition resolves before the score bands, in the same way a failed layer already does, so that a high Maturity Score can never override a demonstrably absent overseer.

3. An Explicit Calibration Objective

The maturity thresholds that admit each Authority Level are organizationally tunable (Section 4.5), and the decision-matrix bands of Section 8.9 are illustrative defaults. Version 1.3 strengthens the calibration process by giving it an explicit **objective template** against which threshold choices can be derived and defended, not merely documented: the expected loss associated with each Authority Level, expressed as a function of demonstrated maturity, decision stakes, reversibility, and overseer engagement. Calibrating a threshold then

becomes the auditable act of choosing the level that minimizes expected loss for a decision class, given its evidence — so that a regulator or reviewer can answer not only "what Authority Level was assigned, and under what calibration?" but "under what objective, and why that threshold?" The formal apparatus of the decision-authority literature is one candidate source for this template. This extension changes nothing about the invariants; it makes the numbers defensible by derivation as well as by record.

4. Decision-Maker Alignment in the Stakes Floor

The Wisdom layer scores the alignment of the AI *output* with user intent, organizational policy, and societal constraints (Section 2.1). Version 1.2.1 does not, however, represent the alignment of the accountable *human* — the stakes floor attaches an accountable human at AL-4 but treats that human's presence, not their incentives, as the governed quantity. The decision-authority literature shows that the alignment between the deciding party and the organization materially changes the optimal allocation of authority. Version 1.3 therefore adds a decision-maker alignment-and-accountability factor to the stakes-floor logic of Section 4.4: the degree of the accountable human's alignment and the strength of their accountability may raise the required level of human authority, in the same conservative, floor-only manner as consequence severity and irreversibility already do. Consistent with the existing asymmetry, this factor may raise required human authority; it may never lower it, and it may never raise AI delegation above the maturity ceiling.

5. The Non-Adversarial Principal (a Stated Scope Boundary)

The decision-authority literature demonstrates that, under a pure profit objective, an organization may find it optimal to deploy an AI that is *deliberately* degraded (an "unreliable" AI held below feasible capability) or *deliberately biased* against the human (an "antagonistic" AI), in order to manipulate the human's effort. Version 1.3 does not adopt these remedies; it names the boundary that excludes them. WEKID governs for *warranted* trust on behalf of a non-adversarial principal — an organization that wants outputs that are factually sound, operationally safe, and appropriately judged, and that is accountable to users, regulators, and the public for them. An antagonistic or deliberately unreliable AI is precisely the engineered, manipulative behavior the Information and Wisdom layers exist to detect and gate; it optimizes one party's payoff at the expense of user welfare and system trust. Stating this as an explicit assumption converts a blind spot into a governed boundary condition and clarifies what WEKID is for.

The relationship is best seen graphically. WEKID keeps two independent dials where the profit-maximizing model effectively has one: it can hold AI capability at maximum for decision *support* while withholding decision *authority* through a lower Authority Level or an AL-4 reservation. It thus reaches the literature's legitimate goal — keeping the human engaged and deciding what matters — without the cost of a worse or antagonistic machine. Figure 2 maps the four types of the source taxonomy onto the WEKID ladder and marks the region WEKID rules out by design.

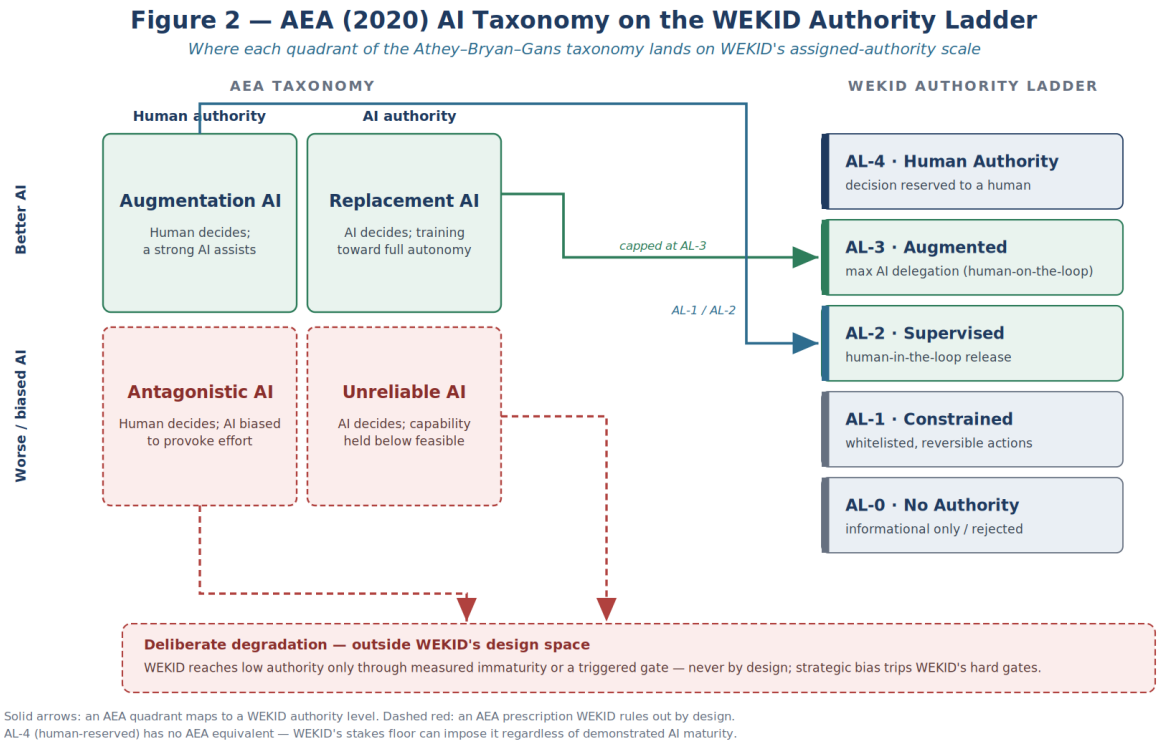

Figure 2. The Athey–Bryan–Gans (2020) AI taxonomy mapped onto the WEKID Authority Ladder. The "better AI" quadrants (Replacement, Augmentation) map onto delegable levels; the deliberately degraded quadrants (Unreliable, Antagonistic) fall outside WEKID's design space. AL-4 has no counterpart in the source model.

Call for Contribution

The five extensions in this paper were proposals, and settled within the framework. Each passed through the WEKID RFC stages — proposed, triaged to a governance surface, reviewed by domain experts, opened for public comment, decided by the Standards Council, and issued as a versioned release. Practitioners, researchers, and domain working groups are invited to contribute failure cases, calibration evidence, and critique, and to test whether these extensions hold under conditions the author has not foreseen. Contribution details and the current governance surfaces are at https://wekid.org/community.html#surfaces.

References

James Madigan Judge, *WEKID™ Master Whitepaper — A Hierarchical Epistemic Framework and Operational Model for Evaluating and Governing Enterprise Artificial Intelligence Outputs*, Version 1.2.1, WEKID LLC (July 2026). The canonical statement of the WEKID framework, cited throughout this companion by section.

Susan C. Athey, Kevin A. Bryan & Joshua S. Gans, "The Allocation of Decision Authority to Human and Artificial Intelligence," AEA Papers and Proceedings 110 (May 2020): 80–84.

Philippe Aghion & Jean Tirole, "Formal and Real Authority in Organizations," Journal of Political Economy 105 (1997): 1–29.

WEKID Community Contribution Framework — Governance Surfaces and RFC Process, WEKID LLC. https://wekid.org/community.html#surfaces