

Clarifying Agent Signals and Governance Implications

Roles of Agents vs. Human Decision-Makers

Semi-autonomous agents in our process are designed as *assistive tools* for analysis and preparation, not as decision-makers. They handle tasks like **content analysis**, **ambiguity identification**, **literal translation with preserved ambiguity**, **metadata tagging** (provenance, traceability), and **classification signals** (e.g. flagging health claims or cultural terms). These functions provide *epistemic support* to human editors by surfacing information and potential issues ¹. Crucially, final editorial judgments—such as approving or rejecting content—remain entirely with humans. The agents **do not** make outcome decisions, alter the defined scope of a proposal, or promote anything to a canonical status. This division aligns with best practices in AI-assisted decision support: the AI may inform or suggest, but *the human retains ultimate responsibility and control* ¹ ². By design, any suggestion or flag from an agent is advisory; it is not a “closure” or final determination.

Governance perspective: Reinforcing this role clarity is vital to prevent misunderstandings. Global AI ethics guidelines underscore that AI systems should never displace ultimate human accountability ². In our context, that means an agent’s output should be viewed as **insight for consideration**, not an automatic verdict. When everyone understands that agents assist with grunt work and pattern-spotting while humans author the *immutable decisions*, we reduce the risk of over-reliance on AI or mistaken assumptions that “the system decided.” This clarity sets the stage for interpreting the signals agents produce without conflating them with decisions.

Common Signals of Friction or Misalignment

In practice, certain **signals** in the human-agent interaction serve as red flags that our current workflow or understanding might be misaligned. These signals typically become visible through repeated patterns in how humans respond to or work around agent outputs. Key examples include:

- **Repeated Clarification Questions:** Team members frequently ask things like “*Do we need to act on this now?*” or “*Is this already a decision?*” in response to agent outputs. This indicates confusion about whether an agent’s flag or suggestion implies an immediate obligation or finality. Such questions have arisen in early pilots whenever an agent highlighted an issue, and humans weren’t sure if it was merely informational or demanded action. The recurrence of these questions is a signal that the agent’s intent (or the workflow status) isn’t clear enough from the output alone. It suggests users might be reading urgency or authority into agent signals that wasn’t intended.
- **Frequent Human Overrides of Agent Flags:** These occur when editors regularly ignore or reverse an agent’s classification or recommendation. For instance, an agent might flag a mild ambiguity or low-severity content issue that humans consistently decide is not worth changing. If “*technically*

correct but editorially premature” warnings are being overridden repeatedly, it signals a mis-tuning of the agent’s sensitivity or relevance. In other cases, an agent might classify something as a potential policy violation, but the human reviewer sees it in context and dismisses the flag as not problematic. A pattern of overrides or ignored flags means the signal is not aligning with editors’ judgment or needs. Industry guidance treats a **spike in override rates** as a sign that the automation isn’t well-calibrated to real needs ³ – either the agent is over-flagging inconsequential issues, or users lack trust in its accuracy ⁴ .

- **Persistent Unresolved “Inbox” Signals:** Our system surfaces certain agent-generated alerts or tasks (for example, an “inbox” of ambiguous excerpts or flagged items) for humans to review. A signal of friction is when many of these items linger without resolution or are repeatedly deferred. If the same non-critical ambiguity remains highlighted run after run with no closure, it suggests that editors don’t find it important enough to address *yet* the system keeps surfacing it. This pattern might indicate that the criteria for surfacing such signals are too aggressive, or that guidance on how to disposition them (“acknowledge and move on” vs. “must resolve before closure”) is lacking. In short, unresolved signals point to a gap in either **actionability** or **policy clarity** – editors might be thinking “yes, it’s ambiguous, but so what?” without a clear rule on whether they can safely ignore it.
- **Out-of-Band Resolutions and Workarounds:** Sometimes humans step outside the official workflow to resolve an issue – for example, having a quick hallway conversation or a chat message to decide on a contentious point – and then come back and document a conclusion after the fact. If we notice decisions being made *around* the system (with the actual decision rationale only noted retrospectively, if at all), that’s a strong signal of friction. It means the current process or tooling wasn’t sufficient for that scenario: maybe the interface didn’t allow an easy way to discuss and record a nuanced decision, or perhaps the agent kept flagging something that the team decided to waive. While occasional out-of-band discussion is natural, a *pattern* of it (e.g. frequent offline fixes or unofficial “pre-decisions”) suggests the system’s expressiveness or the documentation of process is inadequate. Governance-wise, this behavior is explicitly discouraged – we have a norm that all editorial commitments should be recorded as closures with rationale, to maintain transparency and traceability. Therefore, seeing off-system decisions is a red flag that our documentation or tools need to better support *in-system* handling of those cases.
- **Agent Overstep or Scope Creep:** Another signal is when an agent output ventures beyond its intended role, prompting humans to intervene. For example, if an agent were to start “*approving*” a piece of content or altering an excerpt’s scope (something we’ve barred them from doing), editors would immediately override that as out-of-scope. Such incidents (even if rare) signal a **governance boundary being tested**. It could be caused by a misconfiguration or an overly eager agent feature. The key is that any time an agent blurs the line and appears to make a decision or change content on its own, humans notice and correct it – and that incident is a valuable signal that our safeguards or definitions need reinforcement. Fortunately, our agents have not been designed to do this, so direct oversteps are unlikely; more common is a subtle form, such as an agent *implying* a recommended outcome in its analysis. In those cases, the human may feel the agent is “leaning” toward a decision, which can be uncomfortable. Recognizing these subtle signals is important so we can dial back any unintended normative tone in agent outputs.

Each of these signals represents a form of **friction** between the human and the AI assistant. They tend to surface as repeated patterns or frequent questions in our operations. Importantly, these signals are *not*

about the AI being “wrong” in a binary sense; often the agent is factually or technically correct in its analysis. Rather, the friction comes from **mismatch in relevance, timing, or clarity** relative to human workflow and expectations. For example, an agent flagging an issue “too early” or without sufficient context can be technically accurate yet not helpful, leading humans to override it. The presence of these signals tells us something about our governance and where people feel the process isn’t smooth or clear.

What Governance Lessons These Signals Provide

Seeing the above signals prompts us to ask: *what do they indicate about our governance, and what should we do about it?* Each type of signal carries slightly different implications at the governance level, but broadly they are feedback on two main areas: **policy clarity** (documentation, training, guidelines) and **system design** (features, agent tuning, workflow structure).

- **Need for Clearer Documentation & Training:** When we observe *repeated clarification questions* or confusion about whether an agent’s output is actionable, it usually points to a documentation gap. The governance lesson is that our guidance to staff (or user interface cues) might not be explicit enough about what a given signal means. For instance, if users frequently ask “Is this signal blocking or just informative?”, we likely need to clarify in the handbook or UI labeling that “agent signals are for visibility only, and do not mandate immediate action unless explicitly noted.” Emphasizing the **informational (non-decisional) nature** of agent outputs in training materials can resolve the confusion. Similarly, if the concept of a “closure” (a human decision record) vs. an agent suggestion is not well understood, we must reinforce those definitions. This is a relatively straightforward governance fix: update the documentation, provide examples in onboarding, and ensure that every agent-generated notice has a short description of its purpose. Often, making roles and intentions more explicit is enough to eliminate the ambiguity that led to the signal. In short, when the root cause is human misunderstanding or lack of context, the solution is usually **better communication** rather than changing the system. We saw early on that some teammates interpreted “something happened” as “a decision was made” due to how an agent message was phrased – a problem solved by rewording the message and explaining it in the user guide. Clear policy and terminology go a long way toward aligning mental models.
- **Calibrating Agent Signals (Visibility vs. Noise):** Patterns of **overridden or ignored agent flags** teach us about the calibration of our agents’ outputs. If low-severity or exploratory signals are consistently overridden, governance should ask *why*: Are the agents flagging too much trivial content? Are the severity levels or confidence thresholds not tuned to what humans consider actionable? Or do humans perhaps lack trust in those signals’ accuracy? The response could go two ways: (a) **Documentation/Policy Adjustment** – for example, we might explicitly tell editors that “green-light” (low-severity) flags from the agent are purely informational and can be safely ignored unless they compound, just to affirm that ignoring them is acceptable. This would legitimize the current override behavior as consistent with policy (if indeed those signals are not important). Or (b) **System Change** – if we determine the agent is genuinely oversensitive, we might refine its algorithm or thresholds to raise fewer false-alarms, so that when it does flag something, editors know it truly needs attention. The key governance insight here is treating signals as **feedback on utility**. A long list of “technically correct but irrelevant” alerts is counterproductive; it can desensitize users (the *undertrust* problem, where people start ignoring the AI entirely) ⁴. Our goal is to ensure signals remain **signals** (useful visibility) and do not become noise. So governance might decide, for instance, to only surface ambiguity flags when they pertain to policy-sensitive text, rather than any minor

phrasing nuance. In essence, frequent overrides tell us to either *better guide the humans on what to ignore or tweak the agent's output to be more relevant* – both routes aim to re-establish the right balance where signals are neither obligations nor annoyances, but helpful insights.

- **Actionability of Persistent Signals:** When certain agent flags or suggested tasks linger without resolution, they highlight a gray zone in our process. Governance should interpret a **persistent unresolved signal** as indicating that the issue is not important enough to warrant a human decision *under current policy*. This might mean our workflow lacks a proper “*acknowledge and waive*” mechanism. For example, if an excerpt is repeatedly flagged for potential inconsistency but editors collectively feel it's acceptable, we should have a way to record that judgment (even if it's a decision to “leave it as is for now”). An immutable closure of “decided not to change X” would clear the item from the pending list and provide rationale for future reference. If our system doesn't allow that, then the signal's persistence is telling us the governance framework needs a tweak: perhaps introducing a new closure type like “Noted – No Action Taken” or updating guidelines on how to handle non-issues. On the other hand, if the team believes the agent shouldn't be flagging it at all, that might call for a system adjustment (e.g., refining the agent's criteria). In both cases, the broader point is **making the status explicit**. We never want silent ignoring to be the norm; every surfaced issue should either lead to a change or an explicit decision of “no change”. Governance can address this by clarifying roles (e.g. assign someone to regularly triage lingering flags) or by adding options to formally close out such items. This ensures traceability – a core principle of our process is that outcomes are documented for auditing ⁵. A backlog of unattended signals is essentially undocumented decisions (decisions by omission), which we should avoid. Thus, persistent signals prod us to either downgrade them (if they truly don't matter) or force a documentation of the choice not to act.
- **Maintaining In-System Transparency:** Out-of-band resolutions are a serious signal that our *current governance tools might be insufficient*. Each time a decision happens outside the established artifact trail, we risk losing context and accountability. The fact that team members felt the need to go “offline” implies that either (1) the process for handling it in-system was too cumbersome or unclear, or (2) people were unsure whether they *could* use the system for that scenario (perhaps thinking it was outside scope). Governance response here is twofold: **cultural reinforcement** and **system improvement**. Culturally, we double down on the norm: no silent fixes, no decisions without documentation. If something *must* be discussed live (which is fine), the conclusion still needs to be captured in an artifact (like a meeting note attached to the proposal or a closure with rationale). We've been fairly good at this – whenever a side discussion occurred, the outcomes were later written down – but if it's happening often, we should question if the system can be made more accommodating. Maybe the agent or workflow could facilitate a “request discussion” tag or a way to flag an item as needing broader deliberation, so it stays within the loop. The governance takeaway is that **traceability and explicit context are paramount**: we want every significant action or non-action to be traceable in the logs or artifacts ⁵. Out-of-band signals indicate a crack in that traceability. Initially, the fix might be as simple as reminding everyone in training that “if it's not recorded, it didn't happen” in our process. Over time, if we see certain types of issues routinely handled outside, that justifies a deeper change, like new fields to capture discussion or new proposal types to cover those scenarios. Essentially, these signals urge us to ensure the system and documentation make it easy to do the *right* thing (document decisions) and hard to accidentally do the wrong thing (bypass the records).

- **Reinforcing Agent Boundaries:** If we ever catch an agent output that oversteps (even implicitly), it's a signal to *immediately shore up the guardrails*. Governance documentation already specifies that agents cannot decide or publish anything; seeing any hint of that (e.g., phrasing that sounds like a decision) is actionable. The response is straightforward: adjust the agent's template or training data to remove the authoritative tone, and remind users that any such output is not valid. We haven't had agents truly "go rogue," and we structurally prevent it by not giving them authority. But this remains a watch-point signal. It indicates the health of our human-in-the-loop design. In a well-governed system, **no one blames the agent for a bad outcome** ²; instead we blame the process or ourselves for misusing the tool. By treating any agent overstep as a system misconfiguration rather than "the AI's fault," we keep accountability where it belongs. So this signal, when present, tells us to refine either the system or the users' understanding (or both) to realign with the principle that agents assist and humans decide.

Overall, these signals collectively help us gauge whether our current governance (policies, documentation, and system rules) is sufficient. They make visible the points of **tension, confusion, or inefficiency**. Each instance of friction is an opportunity to ask: is this just a misunderstanding that better documentation can fix, or is it revealing a deeper design flaw in the workflow? The **governance mindset** is to treat signals as neither mandates nor nuisances, but as *valuable visibility into system dynamics*. In fact, the reason we instrumented agents to flag things in the first place was to gain visibility on potential issues early – we just must ensure that visibility doesn't get misconstrued as automated governance. A useful analogy is how an airplane cockpit issues many alerts and indicators: pilots are trained to interpret which require immediate action and which are just informational. Likewise, our team should know that an agent's "caution" signal is like a yellow light, not a red light. The signals prompt human judgment; they do not replace it. Keeping that distinction clear is itself a governance success.

Balancing Documentation-Only Fixes vs. Deeper Changes

A core goal of monitoring these signals is to **sharpen our judgment on intervention level**. When we see a friction signal, we ask: *Can this be addressed by clarifying our guidelines and expectations (a documentation or training fix)? Or does it require modifying the system or governance architecture?* We want to avoid knee-jerk system changes for issues that a memo or slight policy tweak could solve, but we also don't want to paper over a systemic flaw with repeated reminders if the root cause persists.

Generally, our approach is to attempt a **documentation-only clarification first** in many cases. Documentation is easy to update and can often resolve misunderstandings without disrupting the workflow. For example, if users misinterpret an ambiguity flag as an error that must be fixed immediately, a quick update to the guidelines (stating "Ambiguity flags are FYI and need action only if X or Y condition holds") might clear up the issue. This kind of clarification can immediately reduce anxiety and friction, as everyone aligns on what the signal *means*. It's low cost and preserves the existing system behavior while making sure people use it correctly. In governance terms, this falls under continuous training and communication – ensuring that humans have "**proper understanding of the AI system's limitations and outputs**" ⁴ so they can appropriately calibrate their trust and actions. Many times, what looks like a system problem is actually an education problem. Thus, our first response to a troubling signal is often to double-check: *did we explain this properly?* If not, we improve the explanation.

However, **documentation has its limits**. If we've clarified and reminded and the friction still persists (or if the signal reveals something fundamentally misaligned with our goals), that's when a deeper governance or

system change is justified. Some examples of deeper changes include: adjusting the agent's logic or thresholds, redesigning a part of the workflow, adding a new artifact type or status (to better categorize an outcome), or even changing team processes. The threshold for jumping to a system change is typically when a signal indicates a **structural issue or recurring inefficiency**. For instance, if out-of-band decisions keep happening despite training, maybe the system genuinely lacks a needed collaboration feature – so we might introduce a formal “discussion” step or tool. Or if humans keep overriding a particular type of agent flag even after we told them when to pay attention vs. ignore, perhaps the agent's model is not well-suited to our context and needs retraining or retuning. Essentially, when a signal reveals that “*we're asking humans to constantly work around X*”, it's a good sign X should be fixed at the source. High override rates or low trust in certain outputs, if not solved by clarifications, suggest the **cost of leaving the system as-is is higher than the cost of improving it** ³. At that point, governance would green-light a more operational change.

It's important to emphasize that we are **avoiding premature operationalization**. We deliberately aren't rushing to build dashboards of override metrics or fancy tagging of every signal at this stage. We don't want to over-engineer a solution for something that might be solved by aligning understanding. The focus is first on *what to look for and why it matters*, not on implementing real-time monitors for those signals. In practice, we notice these signals through normal team interactions and retrospectives (e.g., repeated questions in meetings, patterns in weekly reports, etc.). That qualitative awareness is enough to inform governance decisions for now. If a pattern becomes pronounced, we can then consider lightweight instrumentation – but the current task is to refine our judgment, not build a software feature.

In summary, **we treat agent-related signals as a form of feedback** on our governance model's effectiveness. Minor confusion or one-off incidents usually warrant a documentation update or a team discussion to clarify roles. Such *documentation-only clarifications* often prove sufficient in smoothing the workflow (and are preferable as a first step due to their low risk). However, when signals persist or point to a deeper mismatch (for example, a feature that doesn't fit the way humans need to work, or a risk that isn't being mitigated by current policy), that's when we escalate to considering a *deeper change* in the system or governance structure. The art is in correctly diagnosing which category a signal falls into. By cataloguing these signals and analyzing their root causes, we aim to develop an intuition for this: **Are we dealing with a communication breakdown or a design flaw?**

Our goal is to resolve friction at the lowest effective level. Often, clearer guidelines and expectations will suffice; when they don't, the signals will continue, effectively *justifying a governance intervention*. This iterative approach ensures we neither ignore important warning signs nor overreact to them. It keeps the human-editor-AI-agent partnership productive and accountable, without adding unnecessary complexity prematurely. In doing so, we uphold the principle that AI agents serve to *augment human judgment, under transparent oversight*, and we ensure that any needed course-corrections – whether in words or in code – are made before small frictions grow into serious breakdowns.

Throughout this process, remembering *why* we have these agents and governance in the first place helps guide decisions: to improve efficiency and consistency while maintaining human responsibility and editorial integrity. Any signal that suggests we're drifting from that balance is, at its heart, a sign to pause and realign either our understanding or our system with those founding principles. By heeding these signals with the appropriate level of response, we can continuously refine the system's reliability and the team's trust in it, without conflating the tools with the decision-makers.

1 Owing Decisions: AI Decision-Support and the Attributability-Gap - PMC

<https://pmc.ncbi.nlm.nih.gov/articles/PMC11189344/>

2 Ethics of Artificial Intelligence | UNESCO

<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>

3 When Bots Make Mistakes: A Practical Guide to Safe Human-in-the-Loop AI for SMBs | Artificial Intelligence | MyMobileLyfe | AI Consulting and Digital Marketing

<https://www.mymobilelyfe.com/artificial-intelligence/when-bots-make-mistakes-a-practical-guide-to-safe-human%E2%80%91in%E2%80%91the%E2%80%91loop-ai-for-smbs/>

4 Human-AI Interaction Models for Workforces | Gloat

<https://gloat.com/blog/human-ai-interaction-models/>

5 Knowledge Based Agents in AI What They Are and How They Work

<https://key-g.com/el/blog/knowledge-based-agents-in-ai-what-they-are-and-how-they-work/>