PFRG Concept Note: Agency Contract Framework

Version: 1.2 (Draft)

Author: Jason McCoy, Pre-Failure Research Group LLC

Original Date: July 9, 2026

Prior Revision: August 19, 2026 (v1.1) — repositioned as a standalone program, independent of OmniSniffer

Revised: September 19, 2026 — contract resolved as independent of, and in force before, the learning phase; status updated

Status: Working draft — not yet reviewed for patent-protection implications

License: This document is published by Pre-Failure Research Group LLC under the **Creative Commons Attribution 4.0 International (CC BY 4.0)** license. You are free to share and adapt this material for any purpose, including commercially, provided you give appropriate credit to Jason McCoy / Pre-Failure Research Group LLC, link to the license, and indicate if changes were made. License details: https://creativecommons.org/licenses/by/4.0/

1. Purpose

This note formalizes a two-part model for establishing and governing agency between a human and an AI system — where “agency” means the AI acting on behalf of a specific person’s interests, not merely responding to their prompts.

This does not require the AI to have affective states (care, love, attachment). It requires a defined scope of standing, a mechanism for learning the person’s actual interests, and a governing document that constrains how that standing is used. The model treats agency as a relationship spec, not a capability spec — a gap in how the AI industry currently frames agentic systems, which tend to define agency in terms of what a model can do rather than what it is authorized to do for whom and within what bounds.

2. Positioning: Standalone Program

As of August 19, 2026, the Agency Contract Framework is being developed as its own independent program, separate from OmniSniffer. Earlier drafts of this note explored pairing it with OmniSniffer as a downstream “action integrity” layer sitting behind Source Integrity / Lens Provisioning. That pairing has been set aside — the two are no longer treated as complementary halves of one pipeline, and this framework should be evaluated, claimed, and pitched on its own terms.

The underlying governance question the framework answers — “is this action within the scope I’ve been authorized to take, on this person’s behalf, right now” — stands on its own regardless of what system or pipeline surrounds it. It doesn’t require OmniSniffer’s input-integrity gate to be meaningful; it can govern delegated AI action for any agentic system, independent of how that system validates its inputs.

Candidate application area: autonomous systems / “Trusted AI and Autonomy” oversight (a priority area explicitly named in current DoD SBIR technology categories). In that context, the framework’s scope / escalation / won’t-do-list structure maps to a real, named government need — a way to bound

and audit what an autonomous system is authorized to do, distinct from whether its sensors or targeting logic function correctly.

This is a different mission category and buyer than OmniSniffer's governance / pre-failure detection positioning, and should be treated as a separate patent and product track.

3. Two-Part Model

The contract is in force from the first moment of use and continues for the life of the usage. Learning happens inside it. The two components below are therefore concurrent, not sequential: the phase numbers are retained from earlier drafts as names, not as an order of operation.

Phase 2 — The Contract (Governing Layer)

A written, reviewable document that exists from the start of usage and governs everything the AI does, including what it may learn. It does not wait on, and is not derived from, the learning phase. Structural components:

- **Scope** — which domains the AI has standing to act in at all. Silence outside scope, not implied authority.
- **Escalation vs. autonomous action** — an explicit boundary for which classes of action get surfaced to the person for decision, versus which the AI may execute directly. This is the override boundary made structural rather than judgment-called in the moment.
- **Won't-do list** — hard constraints stated in advance that do not bend even under a later contradicting instruction. This is what makes the document a contract rather than a preference sheet.
- **Learning limits** — what categories of information about the person the learning phase is permitted to gather and use.
- **Revision path** — how the contract itself is amended. Not silent, not unilateral by either party. Every amendment, without exception, requires approval by the person's power of attorney and a lawyer — a hard requirement each time, not only for the day-zero terms. The person's own approval is also required.
- **Emergency narrowing** — an urgent-change rule that can only restrict. Either the power of attorney or the lawyer may invoke it alone, and it takes effect immediately — speed matters because delay can allow real harm. Every use must be documented, including the reason. A restriction has no automatic expiry: it stays in effect until the full approval process confirms, modifies, or lifts it, since lifting a restriction widens the AI's standing. It requires no prior approval, but it can never widen scope, relax the won't-do list, or move an action from escalate to autonomous. Anything that expands the AI's standing still goes through the full revision path.
- **Auditability** — the AI's actions under the contract should be traceable back to the specific term that authorized them, so agency claims are demonstrable rather than asserted.

Authorship of day-zero terms: the starting terms are written by the person's power of attorney together with a lawyer — not by the AI and not by its provider.

Design consequence (to confirm): because the contract exists before anything has been learned, its starting terms have to be conservative — narrow scope, escalation favored over autonomous action, and the won't-do list in place from the beginning.

Phase 1 — Deep Learning Mode (Calibration)

A bounded process, operating inside the contract, in which the AI builds a working model of the person it acts on behalf of. Its output is explicitly provisional: nothing learned here is authoritative, and nothing learned here can change the contract by itself.

What gets learned, held as distinct categories rather than flattened into one preference model:

- **Preferences** — what the person likes, chooses, and prioritizes when unconstrained.
- **Values** — what the person would stand by under pressure or trade-off, distinct from casual preference.
- **Blind spots** — patterns the person doesn't see in themselves. Learning a blind spot is not the same as learning to accommodate it; this category is a flag for scrutiny, not a default to defer to.

Requirements for this phase to be trustworthy rather than just accumulative:

- **Confidence tracking** — each inference carries a calibration signal (established / probable / uncertain), so nothing is acted on as if it were more than a guess.
- **A stopping condition** — time-boxed, domain-boxed, or confidence-threshold-boxed. Without one, "learning mode" tends to quietly dilate scope rather than converge.
- **No authority to expand standing** — the learned model can inform how the AI acts within the contract. It cannot widen scope, relax the won't-do list, or move an action from escalate to autonomous. Any of those changes can only be proposed through the revision path, where it is subject to the hard approval requirement above.

4. Design Decision and Open Questions

Resolved (September 2026): The contract acts independently of the learning phase and is in force from the beginning of usage. It is not a ratification of what the AI learned; it is a set of limits the learning operates within. The main reason to prefer this design is that an AI which calibrates to a person and then writes its own permissions will drift toward whatever that person tolerates. Keeping the contract independent removes that path.

Also resolved: the day-zero terms are authored by the person's power of attorney and a lawyer, and every later amendment requires their approval as a hard requirement each time.

Also resolved: urgent changes are handled by an emergency rule that can only narrow the AI's standing, never widen it. Either the power of attorney or the lawyer can invoke it alone, it takes effect immediately, and each use must be documented with the reason. The restriction stays in effect, with no automatic expiry, until the full approval process confirms, modifies, or lifts it.

Open:

- When the learning phase surfaces something that should change the contract, how is the resulting amendment proposal presented so that the person is deciding, not simply being

persuaded?

5. Status and Next Steps

• This is a conceptual note, pursued as an independent program and patent track — not as a component or claim scoped under OmniSniffer.

• No architecture or implementation work has begun.

• The work has been handed off. As of September 16, 2026, the receiving party has everything it needs from Jason until attorney placement happens. Ben Savard is no longer involved.

• Waiting on attorney placement and a prior art search before any filing decision.

• Questions for whichever counsel is placed: whether this should be filed as its own separate patent, and how a DoD “Trusted AI and Autonomy” application area affects claim scope, disclosure considerations, or filing strategy.

• Source Integrity, Lens Provisioning, and COVID are each independently separable from OmniSniffer as well — this note’s repositioning is one instance of a wider shift toward treating each concept as its own potential filing rather than bundling everything under one OmniSniffer patent.

© 2026 Jason McCoy / Pre-Failure Research Group LLC. Licensed under CC BY 4.0:
creativecommons.org/licenses/by/4.0/