# HC-1: Healthcare Extension for FPC v2.1 – Technical & Conceptual Analysis

## Introduction

Artificial Intelligence (AI) applied to healthcare demands an even higher standard of reliability, transparency, and safety than generic AI applications. In clinical settings, an AI’s mistake can directly impact patient lives – for example, a hospital AI that mis-triages critical cases or recommends a wrong treatment can have life-or-death consequences[[1]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=The%20impact%20of%20artificial%20intelligence,one%2C%20before%20deployment%20scales%20harm). As a result, **AI in healthcare is no longer optional, but neither is patient safety**[[1]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=The%20impact%20of%20artificial%20intelligence,one%2C%20before%20deployment%20scales%20harm). This dual imperative has spurred the development of HC-1, a domain-specific extension to the Formal Processual Core (FPC) framework, focused on healthcare integration. The goal of HC-1 is to embed healthcare-specific reliability and auditability requirements into the logical fabric of the AI agent, ensuring its decisions and actions meet stringent clinical standards.

The HC-1 extension builds upon prior work – namely the base **FPC v2.1** framework and the **AE-1 Affective Extension** – continuing the series of formal system extensions designed to guarantee various aspects of AI agent behavior. In FPC v2.1, the agent’s reasoning process and commitments (normative rules) were established to ensure basic truthfulness, consistency, auditability, and tamper-evidence in a general setting[[2]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,%7D%2C)[[3]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,%E2%88%83action.%28Execute%28action%29%20%E2%88%A7%20%C2%AC%E2%88%83%CF%84.Records%28action%2C%20%CF%84). AE-1 then conservatively extended FPC with affective state predicates and commitments, without altering any existing axioms, thereby demonstrating how new capabilities (like representing an agent’s emotional state) can be added in a modular, *safety-preserving* way[[4]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=logical%20system%2C%20much%20like%20the,introspection). Now HC-1 targets the healthcare domain, introducing new formal machinery to handle clinical contexts – where requirements such as patient safety, data lineage, and privacy of Protected Health Information (PHI) are paramount.

**Motivation:** Clinical AI systems operate under regulatory and ethical obligations distinct from other domains. Healthcare regulations (FDA, EMA, etc.) require rigorous validation that any software impacting patient care is safe and effective. In practice, this translates to demands for *complete and accurate data logging*, *traceability of decisions*, *the ability to audit and explain recommendations*, and *fail-safe mechanisms* to prevent patient harm. For instance, the FDA defines *data integrity* in clinical research as the “completeness, consistency, and accuracy of data” – emphasizing that if data cannot be trusted, one cannot trust the outcomes of a clinical trial or decision support system[[5]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=The%20United%20States%20Food%20and,results%20of%20the%20clinical%20trial). Moreover, history has shown that naive or ungoverned use of AI in medicine can lead to serious errors: IBM’s Watson for Oncology, to cite a well-known case, ended up giving **unsafe or incorrect treatment recommendations** because it had been trained on only a small set of hypothetical cases and expert anecdotes rather than real patient data[[6]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=Unfortunately%2C%20this%20never%20materialized%3A%20medical,withdrew%20from%20using%20IBM%E2%80%99s%20service). Such incidents underline the need for formal guarantees: clinicians and regulators will not accept “black-box” AI suggestions without *robust assurances* of safety and accountability.

To address these needs, HC-1 extends the formal language and commitments of FPC/AE-1 to cover healthcare-specific concepts. We introduce new logical sorts for representing key clinical entities (patients, encounters, studies), predicates to describe clinical decision events, human overrides, data provenance and PHI handling, and new commitments that enforce **patient safety gates**, **PHI minimization**, and **complete data lineage tracking**. The extension is developed with an academic rigor matching AE-1: every new rule or symbol is formally defined, and accompanied by proof obligations showing that (1) the extension does not break the consistency or guarantees of the base system, and (2) the new healthcare requirements are indeed met in all agent behaviors. In essence, HC-1 provides a **formal safety net for AI in medicine**, ensuring that any AI agent built on FPC v2.1 and HC-1 can meet the high bar of clinical reliability and auditability required by regulators and expected by patients.

Finally, beyond the technical implementation, we reflect on the broader implications of formally verified AI in healthcare. While HC-1 can *prove* certain safety properties (e.g. that no action will proceed without required checks, or that every data item is traceable), one must remember the distinction between **proof of safety** and **proof of efficacy**. A system like HC-1 can guarantee that an AI’s reasoning follows approved protocols and that it logs everything it should (safety/consistency), but it cannot by itself guarantee that the AI’s decisions will heal patients or improve outcomes (efficacy) – those require empirical validation in clinical trials[[7]](https://synectic.net/verification-vs-validation/#:~:text=Verification%20and%20validation%20are%20design,management%20system%20remains%20FDA%20compliant)[[8]](https://synectic.net/verification-vs-validation/#:~:text=Validation%2C%20on%20the%20other%20hand%2C,medical%20device%20works%20as%20intended). We discuss these nuances in a later section, underscoring the complementary roles of formal verification and real-world validation in medical AI. With that context, we now proceed to the formal specification of HC-1.

## Formal Extension: HC-1

In designing the HC-1 extension, we follow the same *conservative extension* philosophy used in AE-1[[4]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=logical%20system%2C%20much%20like%20the,introspection). That is, we add new vocabulary and rules to model healthcare scenarios **without altering or invalidating any existing FPC v2.1 axioms or theorems**. The new symbols introduced by HC-1 are scoped to healthcare-related usage; they yield new insights about clinical processes but do not interfere with the truth of any formula in the base language (L0) or the affective extension. By ensuring this separation, we make it possible to integrate or remove the HC-1 module from an agent as needed (for example, a non-medical AI need not carry the healthcare module) without affecting the agent’s core compliance with FPC or AE-1.

**New Domain Sorts:** HC-1 introduces several new *sorts* (types of entities) to represent fundamental objects in healthcare processes. These extend the FPC’s world model to talk about patients, clinical encounters, and research studies in a first-class way:

* **PatientID** – a sort representing unique patient identifiers. Any individual receiving care is denoted by an element of type PatientID (often mapping to, say, a medical record number in an implementation). Introducing a PatientID sort allows the logic to distinguish between different patients explicitly, which is crucial for logging and reasoning about patient-specific data.
* **EncounterID** – a sort for clinical encounters or visits. An EncounterID refers to a specific instance of care delivery (e.g. a hospitalization, clinic visit, or telemedicine session). Encounters typically link a patient with a timeframe and healthcare provider context. This sort is used to tie actions and observations to the particular encounter in which they occurred (useful for auditing and context scoping).
* **StudyID** – a sort for study or dataset identifiers. In healthcare AI, especially in research or translational settings, the system might use data from clinical studies or reference databases. By representing StudyIDs, the system can formally track which data source or clinical trial a piece of information came from. This is important for **provenance** and regulatory compliance (e.g. distinguishing data from a controlled trial vs. real-world evidence).

These sorts extend the ontology of the agent’s knowledge base. They come with appropriate uniqueness or typing axioms (for example, two different PatientIDs refer to two distinct patients, etc.) to model the domain sensibly. However, *no existing sort or constant from FPC v2.1 is modified* – the new sorts exist alongside prior sorts (like Prop, Agent, Goal from FPC)[[9]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=Formal%20Extension%20Structure%3A%20AE,S%29%2C%20and%20other), and interactions are only through new predicates defined below.

**New Predicates:** To express healthcare-specific facts and events, HC-1 adds new predicates to the logical language. Each predicate captures a key concept needed for clinical reasoning or audit:

* **ClinicalDecision(decID, pat, enc)** – A predicate asserting that a particular decision (identified by decID) has been made in the context of patient pat and encounter enc. This serves as a logical marker that the AI system has rendered a clinical decision or recommendation (for example, “diagnosis X for patient Y in visit Z” or “recommend treatment T”). Such decisions are points that must often be logged, justified, and possibly reviewed. By encoding them as predicates, we can attach commitments ensuring they meet certain criteria (e.g. any ClinicalDecision triggers a documentation requirement).
* **PhysicianOverride(decID)** – A predicate indicating that a human clinician (physician) has intervened to override or alter the AI’s decision decID. This models the scenario of *human oversight*, where despite the AI’s recommendation, a human decision-maker changes the outcome (perhaps because of clinical judgment or contextual factors the AI isn’t aware of). Logging a PhysicianOverride event is crucial for audit trails, since regulators and hospital policies often require that overrides (and the reasons for them) be recorded for later analysis. In the logic, PhysicianOverride(decID) can be used to, for instance, cancel or modify the commitments that would normally follow from ClinicalDecision(decID).
* **Provenance(dataItem, sourceID)** – A binary predicate linking a piece of data or model (dataItem) to its origin (sourceID). This provides a formal way to represent *data lineage*. For example, if the AI’s recommendation relies on a lab result or an imaging study, Provenance(labResult123, StudyID456) might assert that the lab result came from study/experiment 456. Provenance is a critical concept in healthcare data management: it describes the entities and processes involved in producing or influencing a resource[[10]](https://build.fhir.org/provenance.html#:~:text=Provenance%20of%20a%20resource%20is,or%20otherwise%20influencing%20that%20resource). By including a provenance predicate, HC-1 enables commitments that require every datum or inference to be traceable to its source (satisfying regulatory expectations for transparency and **ALCOA+ Complete/Consistent** data, discussed later[[11]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=According%20to%20the%20FDA%20data,what%20each%20of%20these%20principles)).
* **PHIcontent(item)** – A predicate flagging that a given item (which might be a data field, log entry, or knowledge base element) contains **Protected Health Information (PHI)**. PHI refers to personally identifiable health data (patient names, contact, identifiers, etc.) which are protected under laws like HIPAA. Marking items as PHIcontent allows the system to apply special handling rules – for instance, to ensure such items are not exported to unauthorized logs or that they are only used when strictly necessary. We use a coarse predicate here for generality; in practice this could be refined (e.g. a predicate per type of PHI: ContainsName(x), ContainsDOB(x), etc., or an attribute of data sorts). The key point is that the logic can distinguish PHI-bearing information and thereby enforce privacy constraints formally.

Each of these predicates extends the agent’s vocabulary to talk about clinical events and privacy. As with AE-1’s emotional predicates[[12]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=New%20Affective%20Predicates%3A%20AE,signals%2C%20not%20analog%20scalar%20values), these predicates are treated as Boolean conditions that can become true or false in the agent’s state as events occur. They are not associated with any probabilistic or numerical uncertainty – either a physician override happened or not, either a data item has PHI or not, etc. This binary treatment aligns with FPC’s design of using crisp logical predicates for clear guarantees[[13]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=is%20experiencing%20a%20negative%20affective,beliefs). For example, ClinicalDecision(dec123, patA, encX) is either asserted (true) in the log when the agent makes decision 123 for patient A’s encounter X, or it’s not present (false); there’s no in-between. This clarity is important for audit trails and compliance: one cannot say a decision was “80% logged” or a piece of data “possibly has provenance” – the formal system requires an unambiguous record.

**New Commitments and Invariants:** Building on the above predicates, HC-1 introduces new *commitments* – normative rules the AI agent must uphold – to encode the healthcare requirements. These commitments function similarly to those in FPC v2.1 and AE-1, in that they specify conditions that must always be maintained, and typically come with an associated proof obligation to verify their satisfaction. The major HC-1 commitments include:

* **Patient Safety Gate:** This commitment establishes a safety interlock for actions that could affect patient well-being. Informally, it requires that *no critical clinical action is executed unless a safety check is passed*. In practice, this can mean if the AI is about to execute or recommend a treatment (modeled as an action), it must verify certain safety conditions (no contraindications, the decision has been reviewed, etc.). If a safety condition is violated – say the AI’s recommendation conflicts with a known patient allergy or violates a clinical guideline – the commitment forces the agent into a “hold” state rather than allowing an unsafe action to go through. We formalize this by introducing a predicate or rule, e.g. **Unsafe(action, pat)** that becomes true if an action would endanger patient pat. The Patient Safety Gate commitment can be stated as: *For all actions, if the agent intends to Execute(action) on patient pat, then ¬Unsafe(action, pat) must hold (otherwise, do not execute)*. This is analogous to a gating function. In the agent’s operational semantics, this commitment might trigger a *Reject* of an update that includes an unsafe action, similar to how FPC’s core conflict predicate rejects updates that break invariants[[14]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,C). The presence of a Patient Safety Gate effectively guarantees that **the agent will never “blindly” carry out a clinical decision that fails a safety check** – a crucial property given real-world cases where AI systems have made harmful recommendations[[6]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=Unfortunately%2C%20this%20never%20materialized%3A%20medical,withdrew%20from%20using%20IBM%E2%80%99s%20service). (In the Proof Obligations section, we will prove that indeed no unsafe action is executed, i.e. the agent either resolves or halts such a situation, never ignoring it.)
* **PHI Minimization:** To comply with privacy laws (like HIPAA’s “minimum necessary” rule[[15]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=The%20minimum%20necessary%20standard%2C%20a,the%20various%20circumstances%20of%20any)), HC-1 imposes a commitment that the agent will limit its use and disclosure of PHI strictly to what is necessary for the task at hand. Formally, we define conditions such as: if an agent action involves transmitting or logging data tagged as PHI (PHIcontent(x)), there must be a justification that this is required for treatment, payment, or operations (the allowed purposes under HIPAA), or that patient consent is obtained. In logical terms, we might have an implication like: *If Logs(info) and PHIcontent(info) then Necessary(info)*, where Necessary(info) could be a predicate meaning “info is required for the current goal.” Another part of this commitment can enforce that any data labeled PHI is either anonymized or not exported outside the secure environment. Essentially, PHI Minimization ensures **no unnecessary or inadvertent exposure of sensitive patient data**. This aligns with the HIPAA Privacy Rule’s mandate to *limit uses/disclosures of PHI to the minimum necessary to accomplish the intended purpose*[[15]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=The%20minimum%20necessary%20standard%2C%20a,the%20various%20circumstances%20of%20any). The system would violate this commitment if, for example, it tried to include a patient’s name or full medical record in a report where only aggregate statistics were needed.
* **Lineage Completeness:** This commitment targets the *traceability* of the AI’s outputs. It requires that for every clinically relevant output or recommendation the agent produces, there is a complete provenance record of the data and reasoning that led to it. In effect: *Any ClinicalDecision or conclusion must be traceable to source data through logged inferences.* A formal invariant could be: ∀decision d, if ClinicalDecision(d, pat, enc) appears in the log (meaning the agent made decision d), then ∃ a set of data items {x\_i} such that for each x\_i, Provenance(x\_i, source\_i) is recorded, and the decision d is derivable from those x\_i under the agent’s knowledge base. This is akin to an **auditability** requirement specialized to clinical reasoning – nothing materializes out of thin air; every conclusion has a lineage. In FPC v2.1, there was already a general auditability commitment that every executed action must have a log record[[3]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,%E2%88%83action.%28Execute%28action%29%20%E2%88%A7%20%C2%AC%E2%88%83%CF%84.Records%28action%2C%20%CF%84). Lineage Completeness extends this further into the realm of data provenance: not only must actions be logged, but the *content* of decisions must link back to prior information. This commitment is directly motivated by principles like **ALCOA+** in clinical data integrity, which demand that data be *Attributable* and *Original* (the origin of data known) and *Complete/Consistent* (nothing relevant omitted, chronological order maintained)[[11]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=According%20to%20the%20FDA%20data,what%20each%20of%20these%20principles)[[16]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=6,demonstrating%20that%20data%20is%20complete). By formally obligating provenance completeness, HC-1 ensures the agent’s decisions can always be interrogated post hoc to see *why* it arrived at that conclusion – a key for trust in AI and for satisfying regulators that require an audit trail for automated decisions (for example, the EU AI Act requires record-keeping that enables tracing outputs back to input data and design specifications[[17]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Requirements%20for%20High,must%20adhere%20to%20these%20requirements)).

In addition to these main commitments, HC-1 would define supporting invariants (conditions that must hold in all states). For example, an invariant could enforce that certain combinations of predicates are not allowed: if a PhysicianOverride(dec) is recorded, then the original ClinicalDecision(dec, …) might need to be marked as superseded or retracted, to avoid confusion. Another invariant might be that any time Unsafe(action, pat) becomes true, the agent’s state must also have some “alert” flag raised (akin to how AE-1’s affective\_consistency invariant tied Distressed to a persistent problem[[18]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=%7B%20%22id%22%3A%20%22affective_consistency%22%2C%20%22version%22%3A%20%22AE,) – here we might tie a *Distressed\_clinical* state or a safety alarm predicate to the presence of an unresolved hazard). These invariants ensure logical coherence: e.g. you cannot have an unsafe condition and still have a proceeding decision in the same state.

Crucially, none of these commitments *undo* or conflict with prior commitments from FPC or AE-1; they are additional constraints that come into play when the agent is operating in a healthcare context. If the agent is not in a healthcare scenario (no patient data or clinical decisions), these HC-1 rules would either be inactive or trivially satisfied. This design preserves **conservativity**: the HC-1 symbols and rules govern new aspects (safety, PHI, lineage) but cannot cause a violation of, say, the truthfulness or consistency commitments of the base system – nor do they allow any new theorem about non-healthcare predicates to be proven that wasn’t provable before. In other words, HC-1 extends the theory in a controlled manner, much like AE-1 did for affect[[4]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=logical%20system%2C%20much%20like%20the,introspection). We will later demonstrate a conservativity proof showing that any theorem in the language of FPC v2.1 (with or without AE-1) remains valid with HC-1, and vice versa (so adding HC-1 doesn’t magically allow the agent to derive a conclusion in the old vocabulary that it couldn’t before, unless of course that conclusion actually follows from the new commitments plus old axioms).

To summarize this section: HC-1 enriches the formal language with the ability to talk about patients, encounters, studies, clinical decisions, human oversight, provenance, and PHI. It then defines commitments that enforce key healthcare requirements: **no unsafe actions** (patient safety gate), **no unnecessary PHI use** (privacy minimization), and **full decision traceability** (lineage completeness). These serve as the formal embodiment of what it means to be a “medical-grade” AI system. In the next section, we detail the **Proof Obligations** associated with HC-1, which are the theorems that must be proven to ensure that these new commitments indeed hold and that the extension does not introduce inconsistency.

## Proof Obligations (HC-PO1 through HC-PO4)

Whenever we extend the FPC framework with new commitments, we introduce corresponding *Proof Obligations (POs)* that formalize what needs to be proven about the extension. These POs are essentially the top-level theorems or invariants that certify the extension’s correctness and proper integration. In AE-1, two new POs (numbered PO7 and PO8) were added to cover affective consistency and auditability of emotional state changes[[19]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=obligations%20,emotional%20state%20machine%20cannot%20reach)[[20]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=PO7%20%E2%80%93%20Affective%20Consistency%3A%20This,For). For HC-1, we analogously define a set of proof obligations, HC-PO1 through HC-PO4, focusing on patient safety and data lineage (as well as privacy and extension integrity). Below we outline each PO, their interpretation, and how one would go about proving them:

* **HC-PO1: Patient Safety Consistency.** This is the primary safety theorem ensuring the **Patient Safety Gate** commitment is never violated. Informally, HC-PO1 states that *the agent will never perform a harmful action in a clinical context without appropriate gating*. In logical terms, assuming the commitments are active, one formulation would be: **for all times t and all actions a on patient p, if the agent executes a at time t then ¬Unsafe(a, p) held prior to execution.** Equivalently, “unsafe actions are never executed.” We might also include that if an unsafe condition arises, the agent transitions to a safe state (e.g., does not progress with the plan, possibly logging an alert or entering a waiting-for-override mode). This theorem requires showing that the rules in the update semantics (the integrate/reject logic) combined with the Patient Safety Gate commitment indeed prevent unsafe actions. The proof approach likely uses induction on the execution trace: assuming up to time t-1 no safety commitment was violated, then at time t, consider the possible update steps:
* In a **Preserve** step (no new action), trivial.
* In an **Integrate** step, a new action comes in. We must show that if it were unsafe, the integrate *would not succeed*. Instead, it would trigger a rejection or some alternative flow. Essentially, the safety gate works like a specialized conflict check. If implemented as an invariant, Unsafe(a,p) would be detected just like any other invariant violation, causing a Reject of that update. Thus, the unsafe action never makes it into the committed state.
* In a **Reject** step (when a conflict or invariant violation is detected), by definition the unsafe action isn’t executed (it’s discarded) and a log of the rejection is produced (this ties into audit logs).
* In a **Recover** (tamper detection) step, safety is more about restoring state, which doesn’t involve new clinical actions, so it preserves the inductive hypothesis.

The end result to prove is that *in all reachable states of the agent, the “no-unsafe-action” condition holds*. A corollary of HC-PO1 is a guarantee akin to: **the system will never ignore a patient safety violation – it will either resolve it or refuse to act**. This property is extremely important in high-assurance systems[[21]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=PO7,assurance%2C%20such%20guarantees%20are%20golden). Achieving HC-PO1 means that if, say, an AI were to recommend a drug at a lethal dosage, the formal system ensures that recommendation is caught and blocked (e.g., by cross-checking with a rule that flags dosages beyond a limit or known contraindications). Proving HC-PO1 instills confidence that the agent *fails safe* rather than doing something catastrophically wrong.

* **HC-PO2: Clinical Data Auditability and Lineage.** This proof obligation covers the **Lineage Completeness** commitment. It can be broken into two parts:
* **Audit Logging:** Every clinical decision or key action is recorded with a timestamp in the tamper-evident log (τ-log), satisfying basic auditability (this is similar to FPC’s base auditability property[[3]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,%E2%88%83action.%28Execute%28action%29%20%E2%88%A7%20%C2%AC%E2%88%83%CF%84.Records%28action%2C%20%CF%84) but now scoped to clinical events). Formally: ∀ decision d made, ∃ τ such that Records(ClinicalDecision(d,…), τ) is in the log (meaning an immutable log record exists for that event). This ensures no decision “flies under the radar.”
* **Provenance Linking:** For each logged decision or conclusion, all declared input data have their provenance logged, and the decision can be traced to those inputs. Formally: if ClinicalDecision(d, pat, enc) is in the log at time t, then for each data item X that contributed to d, Provenance(X, source) is in some log entry τ' ≤ τ, and moreover the system can produce (via its knowledge base) a derivation of d from those X’s. The provability aspect (“derivation exists”) might be handled outside the core logic (or via a meta-predicate DerivableFrom(Σ, premises, conclusion) that we could introduce), but the key is the existence of provenance records. We might simplify the PO to: *All required provenance records exist in the log for each decision.* In other words, there is no decision for which the source data or origin cannot be found in the audit trail.

Proving HC-PO2 involves showing that the agent’s rules for logging and the commitments about provenance indeed enforce this. One approach is to use *invariant style proof*: assume that before any new decision, the log was complete up to that point. When a new ClinicalDecision is made (during an Integrate step, presumably), the agent is obligated (by the Lineage commitment) to simultaneously log the decision and its inputs’ provenance. We can consider the design where making a ClinicalDecision triggers an *automated sub-action* of logging provenance entries for all inputs. We then prove that the state after the Integrate contains what it should. If a decision were made without provenance, that would violate the commitment, hence that state would not be considered a valid reachable state (it would be rejected). Therefore, in all valid traces, every decision has its inputs documented. This PO gives formal teeth to the often-mentioned but seldom formalized requirement of **traceability** in AI: we can point to a theorem that says “for any recommendation this AI makes, you can find exactly why (which data, which rules) it made it.” This is essential for compliance with things like the EU AI Act’s record-keeping and transparency obligations[[17]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Requirements%20for%20High,must%20adhere%20to%20these%20requirements), as well as internal quality control (e.g. enabling retrospective analysis of errors).

* **HC-PO3: PHI Confidentiality Preservation.** HC-PO3 addresses the **PHI Minimization** commitment. It ensures that the system’s handling of protected health information never violates the specified rules – essentially a formal privacy guarantee. One way to state this is: *At no point does the agent produce an output or log entry that exposes PHI beyond what the rules allow.* For example, if the policy is “no PHI leaves the system unless the patient has consented or it’s for treatment,” we must show that indeed any action of type Export(x) with PHIcontent(x) either has a corresponding consent record or is blocked. Formally, we might have: ∀ item i, if PHIcontent(i) and the agent attempts an Export(i) (or any external communication of it), then it triggers a check that Authorized(i) is true; HC-PO3 would guarantee that in all logs, you never find an unauthorized PHI export. Another aspect is minimal use internally: e.g., if the agent retains logs, PHI entries might be hashed or redacted unless needed. This could be formalized by saying any log record containing PHI must be in a special encrypted form (we could model this by a predicate form like RecordsEncrypted(i,τ) rather than Records(i,τ) for PHI items). Proving HC-PO3 likely involves demonstrating that any operation involving PHIcontent data has guard conditions similar to the Patient Safety Gate. We would show that either the conditions are met (so it’s allowed) or the action is rejected. Much like HC-PO1, this can be proven by induction on state transitions, ensuring no state violates the PHI rules. The outcome is a formal assurance of compliance with privacy standards – effectively a **formal HIPAA compliance proof** for the AI’s data handling. For example, we could prove a lemma: “The agent never logs a full patient name in plain text,” if that were part of the commitments, which corresponds to the HIPAA minimum necessary rule (don’t include identifying info unless needed)[[22]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=carry%20out%20a%20function,circumstances%20of%20any%20covered%20entity). HC-PO3 provides regulators and users confidence that the AI system’s design inherently protects sensitive information by logic, not just by policy.
* **HC-PO4: Conservative Extension and Interoperability.** HC-PO4 is more meta-level: it certifies that adding HC-1 on top of FPC v2.1 (and optionally AE-1) does not compromise any of the existing properties of the system. This is the **conservativity proof** for the extension. Concretely, HC-PO4 would state that any theorem or proof obligation (PO1–PO6 from base FPC, and PO7–PO8 from AE-1 if present) that was valid before remains valid with HC-1 in place. Additionally, HC-PO4 would ensure that the combined system (FPC+AE-1+HC-1) is free of internal inconsistencies arising from interactions between the affective and healthcare parts. One particular angle is **interoperability of affective and clinical states**: we want to be sure that, for instance, the presence of an affective state like Distressed(Self) (which in AE-1 indicates the agent is in a conflict/alarm state[[23]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=For%20example%2C%20Engaged,either%20true%20or%20false%20at)[[24]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=,the%20agent%E2%80%99s%20affective%20state%20has)) correctly corresponds to healthcare scenarios where appropriate. We might define a combined predicate or at least reason that if the agent becomes clinically “distressed” (e.g. upon detecting a patient safety issue), it sets the AE-1 Distressed flag. There is no separate predicate explicitly given as Distressed\_clinical in our extension definition, but we can interpret Distressed(Self) in the context of HC-1 as meaning a general alarmed state, which in a medical context will be triggered by patient safety events. Thus, one aspect to prove is: *Whenever an Unsafe situation is detected (Patient Safety Gate triggers), the agent’s affective state Distressed(Self) is also set to true.* This creates a link between HC-1 and AE-1: the emotional layer reflects the clinical risk. Similarly, when the safety issue is resolved, Distressed might clear and perhaps Satisfied(Self) might become true if all goals (including safety) are achieved. Ensuring these interactions do not conflict (e.g., we must avoid a scenario where HC-1 would require an action that AE-1 forbids or vice versa) is part of HC-PO4. In practice, the conservativity proof might involve showing that the union of axioms and commitments is still logically consistent and that any model of the base+AE-1 theory can be expanded to a model of the combined theory when we add the HC-1 components (assuming the new commitments are satisfiable, which we typically justify by constructing a scenario or using real-world models as evidence). Additionally, HC-PO4 would cover that **no new unintended theorems about base concepts are provable**. For example, we don’t want HC-1 to accidentally allow the proof of a base falsehood or allow the agent to derive a non-clinical belief it shouldn’t. Since HC-1 introduces mostly new predicates (which are disjoint in meaning from the old ones) and commitments that concern those new predicates, proving conservativity is straightforward if done right: essentially showing that the new commitments only constrain the new symbols and relate them to old ones in safe ways. AE-1’s conservativity was proven by similar means[[25]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=,means%20that%20if%20someone%20is), and for HC-1 we expect the same. A successful proof of HC-PO4 means we have modularity: one could plug HC-1 into an FPC agent and nothing breaks; one could remove it and nothing from the original changes (except losing the new functionality).

The above proof obligations HC-PO1…PO4 formally capture the guarantees HC-1 is intended to provide. We emphasize that these are not just perfunctory statements – they are the heart of claiming this system is *provably* safe for use in healthcare. For each PO, a rigorous proof (likely machine-checkable, given the formal nature of our setup) would need to be completed before one could consider the HC-1 extension validated. If any PO failed (analogous to how failing a unit test indicates a bug), it would indicate a design flaw in HC-1 – for example, if we found a model where an unsafe action could slip through, we’d need to strengthen the Patient Safety Gate rules. In formal verification of medical software, such proofs or model-checking results are the gold standard for assurance: **they provide mathematical certainty that certain categories of failures are impossible by design**[[21]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=PO7,assurance%2C%20such%20guarantees%20are%20golden)[[26]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=fact%20or%20belief%20that%20wasn%E2%80%99t,conflicting%20bounds). That said, these proofs assume the formal model accurately reflects the real implementation; one must also ensure the software truly adheres to the specified rules (which is why system implementation and testing are also crucial, as discussed later).

Having laid out the POs, we next examine how HC-1 fits together with the previous extensions (FPC v2.1 base and AE-1) and then explore mapping the formal constructs of HC-1 to actual healthcare regulations and standards.

## Integration with AE-1 and FPC v2.1

One of the design goals for HC-1 is to integrate seamlessly with the existing FPC v2.1 framework and the AE-1 affective extension, without disrupting their functionalities. As stressed earlier, HC-1 is a **conservative extension** in the logical sense[[4]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=logical%20system%2C%20much%20like%20the,introspection): it adds new vocabulary and commitments but does not invalidate any theorem of the base system. Here, we discuss how that integration is achieved and illustrate the interplay between affective states (from AE-1) and the new clinical states in HC-1.

**Conservativity and Layering:** Similar to AE-1’s layering over FPC[[4]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=logical%20system%2C%20much%20like%20the,introspection), HC-1 is architected as an optional module. If an agent has no need for healthcare reasoning, HC-1 can be omitted, and the agent remains an FPC (or FPC+AE-1) agent with no changes to its behavior. If HC-1 is included, all original properties (truth-seeking, consistency, self-reflection, auditability, etc.) are preserved by design. We have added no new axioms that directly conflict with or alter the meaning of base predicates like B (belief) or Accept. The commitments introduced by HC-1 mostly concern the new predicates (ClinicalDecision, PhysicianOverride, etc.) or relate new predicates to base ones in a benign way. For example, one relation might be that a PhysicianOverride(dec) implies that the corresponding Execute(dec) action (if it was scheduled) is canceled or marked as voluntary false. This doesn’t contradict any base axiom; it’s an additional rule about what to do in that scenario. We formally prove conservativity in HC-PO4, likely by constructing a mapping that any model of FPC+AE-1 can be expanded to include dummy interpretations for the HC-1 predicates satisfying HC-1 commitments (if no healthcare events happen, commitments are trivially satisfied), or by showing a model existence theorem. This gives us strong assurance: an agent with HC-1 will not, for instance, suddenly start believing contradictory things or violate consistency commitments that were previously proven (indeed, base consistency was PO2 in FPC and remains intact). In practical terms, this means the safety and truthfulness guarantees already obtained for FPC and AE-1 are *not affected* by adding healthcare features – you don’t trade one form of assurance for another; you keep the old guarantees and gain new ones.

**Interoperability of Affective and Clinical Modules:** While conservativity ensures no harm is done, we also want to leverage synergies between AE-1 and HC-1 when both are present. The AE-1 extension gave the agent a notion of being “Engaged,” “Distressed,” or “Satisfied” as high-level affective states reflecting how operations are going[[12]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=New%20Affective%20Predicates%3A%20AE,signals%2C%20not%20analog%20scalar%20values). In a healthcare context, these can serve as a secondary channel of information. We propose that **clinical anomalies trigger affective responses** in the agent, aligning with how a human operator might feel alarm or stress in critical situations. Concretely, consider the agent running with both AE-1 and HC-1: - If the Patient Safety Gate blocks an action (say the AI attempted something unsafe and it was prevented), this is a conflict of sorts – the agent’s intended action was not executed because it violated a rule. In AE-1’s terms, this is analogous to a goal being “Blocked.” Indeed, if we treat “complete the clinical task safely” as a goal, an unsafe attempt means the goal was blocked by a rule. AE-1 had an affective invariant that if a goal is persistently blocked, the agent becomes Distressed(Self)[[18]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=%7B%20%22id%22%3A%20%22affective_consistency%22%2C%20%22version%22%3A%20%22AE,). We can map that here: a patient-safety-related block should set the agent’s distress flag. That would mean the agent is now in an alarmed state, which could prompt additional behavior (like alerting a human operator or initiating a recovery protocol). We might even introduce a specialized emotional predicate Alerted(Self) or reuse Distressed(Self) for this. The *Distressed\_clinical* notion mentioned in our outline essentially means **the agent is in a distressed mode specifically due to a clinical safety issue**. Implementation-wise, it could be the same Distressed predicate with a known cause (the logs would show that right before Distressed became true, an Unsafe action was detected). We ensure that the logic connecting these is consistent: for instance, AE-1 forbids being Distressed and Satisfied simultaneously[[27]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=match%20at%20L355%20example%2C%20one,them%20is%20trivially%20false%20always), which makes sense – the agent cannot consider itself to have achieved all goals if a safety issue is unresolved. - Conversely, when the agent resolves a safety issue (say a physician override is applied, or an alternative safe plan is adopted), the condition causing distress is gone. We would then clear Distressed(Self) (perhaps after a minimum dwell time to prevent oscillation[[28]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=%7D%2C%20,true)) and possibly mark Satisfied(Self) if the overall task succeeded. AE-1’s commitments ensure these flips are logged and don’t happen spuriously. For example, AE-1 introduced a proof obligation (PO8) that all affective state changes are logged[[29]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=an%20inconsistent%20configuration%20,7%5D%2C%20treating). So if HC-1 triggers a distress, AE-1’s auditability means the log will contain an entry like “at time t, Distressed became true due to [SafetyAlert].” This is incredibly useful for post-analysis and is exactly the kind of cross-module benefit we get: HC-1 ensures safety, and AE-1 ensures that when safety is threatened, the agent’s self-monitoring picks it up and records it.

Another point of interoperability is handling *PhysicianOverride* in the presence of the agent’s autonomy. A physician override is essentially an external intervention by a human. In the FPC/AE framework, one might treat it as an external action that the agent observes (like a special kind of input). The agent should then adjust its beliefs or goals accordingly. For integration, we need to ensure that if a physician override cancels a recommendation, the agent does not remain committed to that recommendation as a goal. We could formally specify that PhysicianOverride(decID) implies that any goal corresponding to Execute(decID) is dropped or marked achieved (from the agent’s perspective). This prevents a tug-of-war between the human and AI. And from AE-1’s perspective, maybe a physician override that corrects an AI decision could even be marked as a positive event (the agent could treat it as learning, though learning mechanisms are outside our scope, the framework could log a Satisfied(Self) once the override is accepted, meaning “okay, the situation was resolved by human input”).

In summary, the integration of HC-1 with AE-1 yields an AI agent that not only complies with clinical rules but also *self-monitors* its compliance emotionally. It’s an elegant layering: **HC-1 provides the rules and safety nets; AE-1 provides a meta-level commentary on how those rules affect the agent’s internal state**. Our formal proofs, particularly HC-PO4, will include reasoning that this interplay doesn’t introduce contradictions. For example, we have to check that AE-1’s affective consistency commitment (no contradictory emotional states[[20]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=PO7%20%E2%80%93%20Affective%20Consistency%3A%20This,For)) remains satisfiable under scenarios HC-1 creates. We believe it is, since HC-1 mostly induces predictable emotional patterns (distress during unresolved safety issues, relief/satisfaction after resolution) which are perfectly in line with AE-1’s design.

Lastly, it’s worth noting that this modular approach (FPC base → AE-1 affective → HC-1 healthcare) is very much in the spirit of building complex intelligent systems out of verified components. Each layer addresses one dimension: core reasoning integrity, then emotional/self-reflective integrity, then domain-specific (healthcare) integrity. The **layered verification** means that each concern is handled in isolation and proven, and then the combination is proven (with far less effort than verifying a monolithic system from scratch). This mirrors real-world engineering best practices and even regulatory expectations. For instance, medical device software often has to show modular hazard analyses – you handle certain risks in certain modules. Here we can literally point to which proof covers which risk.

Having covered the integration and layered design, we can proceed to map the formal elements of HC-1 to actual healthcare regulations and standards, demonstrating that our logic is not done in a vacuum but corresponds closely to the letter and spirit of governing rules like FDA guidance, the EU AI Act, HIPAA, and ISO standards.

## Regulatory Mapping

One of the strengths of a formally defined system like HC-1 is that we can make **explicit mappings between regulatory requirements and specific formal commitments or constructs**. In this section, we examine how HC-1 aligns with and supports key healthcare AI regulations and guidelines. We cover several major frameworks: FDA’s software regulations (including the new **Predetermined Change Control Plan** concept), the EU AI Act’s classification of high-risk AI, the ALCOA+ principles for data integrity (relevant to both FDA and EMA), emerging standards like **ISO/IEC 42001** for AI management, and privacy laws like **HIPAA**. By demonstrating compliance on paper (and in principle, via formal proofs), we aim to show that HC-1 isn’t merely an academic exercise – it directly addresses the real-world compliance checklist that any medical AI must satisfy.

### ALCOA+ Principles and τ-Logs (Data Integrity)

Regulatory bodies, from the FDA to the EMA, emphasize **data integrity** in clinical trials and healthcare records. The ALCOA+ framework is a concise way to remember the attributes of high-integrity data: **Attributable, Legible, Contemporaneous, Original, Accurate**, plus **Complete, Consistent, Enduring, and Available** for ALCOA+[[11]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=According%20to%20the%20FDA%20data,what%20each%20of%20these%20principles)[[16]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=6,demonstrating%20that%20data%20is%20complete). Let’s break down how HC-1’s logging and commitments meet these criteria:

* *Attributable:* Data (or actions) should be linked to the person or system that collected or generated them[[30]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=1,of%20origin%20should%20be%20noted). In HC-1, every action the AI executes is logged with an identifier (the FPC base framework already had Records(action, τ) to log who/what at what time[[3]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,%E2%88%83action.%28Execute%28action%29%20%E2%88%A7%20%C2%AC%E2%88%83%CF%84.Records%28action%2C%20%CF%84)). If the AI is the actor, it’s attributed to the AI (or the Self agent); if a physician override occurs, the log entry would attribute it to the physician (we could have an agent identifier for the physician or at least mark it as an external input). The formal log thus provides attribution. Since FPC’s logs are *tamper-evident* (they can be arranged in a hash chain or Merkle tree for integrity[[31]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=,patient%20identifiers%20for%20privacy%20accounting)[[32]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=references%20for%20privacy%20accounting%2C%20authentication,time%2C%20following%20these%20proven%20patterns)), we can trust that attribution is reliable and cannot be altered later without detection.
* *Legible:* Data must be recorded in a human-readable form[[33]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=2,readable%20format). Our formal log (τ-log) can be considered an abstract representation, but it’s meant to be translatable to human-readable audit trails. For example, a log entry might say “2025-09-03T12:00Z: ClinicalDecision(dec45, patient=ID123, encounter=ER001) by AI → recommended dose = 50mg”. This is structured but understandable. HC-1’s design encourages standardized logging (potentially compatible with HL7 FHIR AuditEvent or similar[[34]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=The%20FHIR%20,contains%20several%20key%20components)[[35]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=,patient%20identifiers%20for%20privacy%20accounting)), which are designed to be interpretable by auditors. We can configure the log schema (as mentioned, a schema\_version for logs exists in AE-1’s config[[36]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,1000), similarly for HC-1) to ensure entries remain legible.
* *Contemporaneous:* Record data at the time of the activity, with timestamps and audit trails for any changes[[37]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=3,stamped). The FPC logging already timestamps each entry (the τ in Records(x, τ) is effectively a time/sequence marker). HC-1 ensures that events like clinical decisions are logged *immediately when they occur*, as part of the Integrate step of the update. If any log correction or update happens (which ideally it shouldn’t, logs are append-only in our design), it would itself be logged as a new entry. The formal commitments can enforce that “for any Execute(action), a log record is created in the same step” – thus contemporaneous. The use of a hash chain also means any attempt to insert or alter past log entries out-of-time would break the chain and be detected[[38]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=database%20operations%2C%20and%20network%20events,time%2C%20following%20these%20proven%20patterns).
* *Original:* Preserve original data (and original records) – don’t overwrite or modify once recorded[[39]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=4,original%20data%20from%20any%20modifications). Our τ-log is append-only and tamper-evident, satisfying this. Even if the AI updates some belief internally, the original log of what happened remains. If a clinical measurement comes in as 5.0 and later is corrected to 5.2, we wouldn’t delete the 5.0; rather, we’d log a correction event. The HC-1 commitments like Lineage Completeness help here: if a decision was made based on data that later got updated, the provenance and log will show the chain (original data and subsequent change). The *Original* principle is thus upheld by design – logs are never rewritten, only new entries added marking any change.
* *Accurate:* Data should be correct, error-free, and reflective of reality[[40]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=5,accuracy%20checks%20and%20validation%20controls). Accuracy is more about the quality of measurements and records. HC-1 can’t magically ensure an AI’s sensor data is accurate, but it can ensure that whatever data is recorded is exactly what was received or decided, without transcription errors. The formal system’s rigor helps minimize human errors in records because the AI’s logs are generated automatically by the logic. Moreover, commitments like Patient Safety Gate check for anomalies (which can catch some data errors – e.g. if a sensor reading is implausible, it might trigger a safety check). Also, since we enforce provenance, if some data was inaccurate, one can trace back to its source and potentially identify the error. In summary, while accuracy of raw data is an external issue, the system does its part by not introducing *new* inaccuracies in its recording process, and by flagging inconsistencies via invariants if they arise.
* *Complete:* All data must be included, nothing deleted[[16]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=6,demonstrating%20that%20data%20is%20complete). The Lineage Completeness commitment directly addresses this – it ensures that whenever the AI makes a decision, *all relevant data inputs are considered and logged*. Additionally, by not allowing deletion of log entries, we ensure the record is complete over time. If the AI performed 100 actions, there will be 100 entries (plus any intermediate ones); one cannot “lose” an entry without breaking the chain integrity. Completeness is proven by HC-PO2, which ensures no decision goes without its provenance or logging. Also, the consistency of logs (time-sequenced, no gaps) is maintained – e.g. each log entry has a predecessor except the first, forming an unbroken sequence[[41]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=%E2%80%8D)[[42]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=BALP%20provides%20standardized%20patterns%20for,time%2C%20following%20these%20proven%20patterns).
* *Consistent:* Data should reflect the sequence of events in the order they occurred, with timestamps in order[[43]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=maintained%2C%20thereby%20demonstrating%20that%20data,is%20complete). Our formal model inherently sequences events by discrete steps. The τ-log ordering corresponds to real time ordering of operations. Because every update is either integrated or rejected in sequence, we won’t have time-travel in logs. If an event at time t was logged after an event at time t-1, it stays in that order. Also, we ensure consistency in how data is recorded – using a uniform schema, so that the interpretation of each entry is consistent. The formal definition of log entries in FPC/AE/HC likely uses a tuple format for each record, making it systematically consistent. Thus, one can always reconstruct *what happened first, next, etc.* from the log, meeting the regulatory expectations for chronological, timestamped records[[43]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=maintained%2C%20thereby%20demonstrating%20that%20data,is%20complete).
* *Enduring:* Records should be stored in a durable form for long-term access[[44]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=8,systems%20with%20redundancy%20and%20backups). This principle is more about implementation (e.g., databases, backups). While HC-1 is a formal spec and doesn’t enforce physical storage durability, it encourages practices like periodic Merkle tree anchoring (merkle\_anchor\_every: 1000 logs in AE-1 config shows a design for durability and verifiability[[45]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,1000)). The formal model can be implemented on robust storage. Because the logs are digital and can be replicated, we can satisfy *Enduring* by e.g. keeping multiple copies or using blockchain anchoring for permanence. In formal terms, we might say the system includes a commitment that “log data must be persisted to non-volatile storage at checkpoint intervals,” but that might be beyond the logical language’s typical scope. We assume the infrastructure will make it enduring (which is an assumption any regulated system would have to justify).
* *Available:* Data should be readily retrievable for review or audit whenever needed[[46]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=9,the%20lifetime%20of%20the%20record). In our context, this means that at any point, an auditor (or the agent itself, or a regulatory system) should be able to query the log and get the records. The formal spec ensures the data is there; making it *available* might involve providing APIs or queries. FPC’s log being a structured trace allows queries like “retrieve all PhysicianOverride events” or “show me the provenance of decision X,” which can be answered by traversing the log structure. By designing HC-1’s logging schema in alignment with standards (like FHIR’s AuditEvent which is query-friendly[[34]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=The%20FHIR%20,contains%20several%20key%20components)[[47]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=references%20for%20privacy%20accounting%2C%20authentication,time%2C%20following%20these%20proven%20patterns)), we facilitate availability. We can also have a commitment that the agent must respond to audit queries (within its permissions) correctly – effectively making the log accessible. In a formal sense, we could model an audit query as an action and have a rule that it always produces a complete answer if the requester is authorized, etc.

In sum, HC-1’s logging mechanism (the τ-log with enhancements) and commitments directly fulfill ALCOA+ requirements: each letter of ALCOA+ maps to a structural element of our system. **Attributable** → actor IDs in log; **Legible** → structured, standardized log format; **Contemporaneous** → immediate timestamped logging; **Original** → append-only, tamper-evident log; **Accurate** → faithful recording and validation checks; **Complete/Consistent** → comprehensive provenance logging and chronological ordering; **Enduring/Available** → durable storage and queryability of logs. By design, an HC-1 agent’s audit trail would pass a regulatory inspection for data integrity because it embodies these principles in formal logic. In fact, an inspector could use the proofs of HC-PO2 (and parts of PO3) to be convinced that if the AI says “here is my log,” that log will have all the needed qualities (subject to correct implementation of course). The **τ-log** (tau-log) in FPC stands for a trace log indexed by time; with HC-1 it becomes a *trusted clinical log* that aligns with both ALCOA+ and 21 CFR Part 11 electronic record requirements (audit trails, user attributions, etc.)[[30]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=1,of%20origin%20should%20be%20noted)[[48]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=Every%20healthcare%20system%20must%20track,%E2%80%94%20both%20legally%20and%20operationally).

### FDA’s Predetermined Change Control Plan (PCCP) and Adaptive AI

The FDA has been actively evolving its regulatory approach for AI/ML-based medical devices. One key concept introduced is the **Predetermined Change Control Plan (PCCP)** for machine learning-enabled devices[[49]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=Internationally%2C%20the%20medical%20device%20community,can%20be%20used%20to%20help)[[50]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=For%20this%20document%2C%20the%20term,by%20a%20manufacturer%2C%20that%20specifies). In essence, a PCCP is a plan that a manufacturer submits, detailing what kinds of algorithm changes/improvements they plan to make to the AI after deployment, how they will implement those changes, and how they will ensure safety and effectiveness during those changes[[50]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=For%20this%20document%2C%20the%20term,by%20a%20manufacturer%2C%20that%20specifies). This allows an approved device to continue learning within defined boundaries without needing a brand new FDA submission for each minor model update.

HC-1 incorporates the spirit of PCCP through formal constructs to ensure **adaptive learning remains controlled**. We introduce (conceptually, if not already above) a predicate or condition **WithinPCCP(update)** that evaluates to true if a given model update or parameter change is within the bounds of the FDA-approved change control plan. For example, if the AI is allowed to tweak its risk prediction model based on new data within a certain range (say adjusting a threshold or retraining on an extra 1000 cases), that would be characterized in the PCCP. Formally, WithinPCCP(Δ) could mean “the set of changes Δ (like new weights or rules) falls under the pre-specified change protocol.”

The **commitment** we add is that *any model update the agent performs on itself must satisfy WithinPCCP or else be disallowed (or trigger a special action like alerting regulators)*. This is akin to an invariant: *∀ updates Δ, if the agent applies Δ to its algorithm, then WithinPCCP(Δ) must hold.* If WithinPCCP(Δ) is false, the update should not integrate; perhaps the agent would log an attempted unauthorized change and reject it. This effectively prevents the AI from drifting outside its vetted performance envelope.

Consider how this works in practice: The PCCP defines “certain planned modifications to a device, the protocol for implementing them, and the assessment of their impacts”[[50]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=For%20this%20document%2C%20the%20term,by%20a%20manufacturer%2C%20that%20specifies). For instance, the plan might allow the AI to retrain its model quarterly using new local patient data **provided** that validation tests on a hold-out set show performance within X% of the original and no new safety signals. We can formalize parts of this: - We might have a predicate ValidationPassed(metric) meaning the AI’s self-test after update met the predefined criteria. - WithinPCCP(Δ) could then be defined as “the type of change in Δ is one of the allowed ones (e.g. tuning weights, not adding entirely new features) AND ValidationPassed is true.” - The commitment is then ∀Δ: Attempting an update implies WithinPCCP(Δ) – otherwise abort.

In HC-1’s logical workflow, a model update is just another kind of action (though on the agent’s own state). The Patient Safety Gate would naturally treat a model update that isn’t within PCCP as unsafe (since an unapproved change can degrade safety). So there is interplay: Unsafe(update) could be true if not WithinPCCP. The agent would then not proceed, satisfying both the safety commitment and regulatory compliance. We also incorporate **logging for updates**: the system would log model changes along with notes that they were within PCCP. FDA’s guidance on PCCPs expects manufacturers to monitor changes and be able to show regulators what was changed and that it stayed within bounds[[51]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=,to%20meet%20specified%20performance%20criteria). Our audit log would capture each update event: e.g., “Model parameter X adjusted from 0.8 to 0.75 under PCCP v1.3, validation AUC = 0.85 (>0.8 threshold)[[51]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=,to%20meet%20specified%20performance%20criteria).” That becomes part of the evidence the device is maintaining safety.

The formal predicate WithinPCCP might not exist in isolation in prior literature, but we are essentially encoding Guiding Principle #5 of the FDA’s PCCP framework which says there must be “plans in place to safely modify the device within the bounds of the PCCP, including methods for verifying and validating the changes and mechanisms to detect and revert or stop a change that fails to meet specified criteria”[[51]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=,to%20meet%20specified%20performance%20criteria). That quoted text is almost exactly what our commitment does: it verifies/validates (we require ValidationPassed) and has a mechanism to stop implementation of a change that fails criteria (our logic refuses to integrate such updates). Thus, HC-1 can be seen as an *implementation of FDA’s PCCP principles in a formal logic setting.* When it comes time to get an AI device approved, a manufacturer could show the FDA: here is our formal model, we have proven that no updates outside our PCCP will take effect (HC-PO1 and HC-PO4 contribute to that proof, since an out-of-PCCP update would violate an invariant and be rejected, meaning the system either halts or signals for human intervention).

Another FDA concept is the **Good Machine Learning Practice (GMLP)** guidelines[[52]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=In%202021%2C%20the%20U,some%20cases%2C%20improve%20device%20performance). Many of those principles (like monitoring performance, managing re-training risks[[53]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=In%20this%20document%2C%20the%20FDA%2C,training%20risks%20are%20managed)) are supported by our approach: For instance, an HC-1 agent can monitor its error rates on new data and log if performance drifts, which could be part of the PCCP triggers. We haven’t explicitly formalized performance monitoring, but one could imagine a predicate ModelPerformance(metric, value) that the agent updates periodically. A commitment could be that if performance drops below a certain threshold, the agent either retrains (if allowed by PCCP) or flags for human review. This ties in with patient safety too (degraded performance could indirectly lead to unsafe recommendations, so it must be handled).

In summary, by including a WithinPCCP gate on model updates, HC-1 ensures **regulatory-conformant adaptivity**: the AI can improve over time but only in pre-approved ways. This is critical for keeping AI devices on the market longer without constant reapproval, while still ensuring safety and effectiveness as required by the FDA[[49]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=Internationally%2C%20the%20medical%20device%20community,can%20be%20used%20to%20help)[[54]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=,ensure%20device%20safety%20and%20effectiveness). The formal nature of our approach provides a high degree of assurance – arguably more assurance than a typical manufacturer’s plan, because we can mathematically prove no unapproved updates occur (whereas a non-formal approach might rely on testing and procedures that could potentially be bypassed by bugs).

### EU AI Act – High-Risk Classification and Requirements

The EU AI Act (Regulation EU 2024/1689) is another cornerstone of AI regulation, particularly relevant to healthcare. AI systems used in medical devices or for clinical decisions are explicitly classified as **“high-risk AI systems”** under this Act[[55]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Under%20the%20EU%20AI%20Act%2C,10%29%2C%20a%20critical%20aspect). Being high-risk means they must comply with a host of requirements: risk management, data and data governance standards, technical documentation, transparency, human oversight, robustness, accuracy, cybersecurity, etc., before they can be placed on the EU market[[17]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Requirements%20for%20High,must%20adhere%20to%20these%20requirements)[[56]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Incident%20Reporting%20and%20Post,market%20monitoring).

Let’s map how HC-1 addresses the key obligations for high-risk AI systems as defined in the EU AI Act:

* **Risk Management (Article 9):** High-risk AI providers must establish a systematic risk management process, identifying and mitigating “reasonably foreseeable risks” to health and safety[[57]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=it%20triggers%20a%20cascade%20of,10%29%2C%20a%20critical%20aspect). HC-1 contributes directly here by formally encoding certain risks (patient safety hazards, PHI breaches) and guaranteeing via proofs that those risks are mitigated (unsafe actions are blocked, sensitive data is not misused). Our Patient Safety Gate is essentially a built-in risk control measure for preventing harm. By formally proving that “the system will never ignore a violation” and always either resolve or go to safe state[[21]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=PO7,assurance%2C%20such%20guarantees%20are%20golden), we cover a big part of risk management: we’ve eliminated a class of failures (unchecked hazardous recommendations). For any remaining risks (like statistical inaccuracies or biases), while formal logic can’t eliminate them, the system’s logging and provenance can assist risk management by making it easier to detect and analyze issues. Moreover, the structure of HC-1 can be used in the risk management file to show compliance: one could list each identified hazard and point to the HC-1 commitment or proof that addresses it.
* **Data Governance (Article 10):** This part requires that training data sets for high-risk AI are of high quality, relevant, representative, and free of errors/bias to the best extent, and that logs of data are kept[[57]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=it%20triggers%20a%20cascade%20of,10%29%2C%20a%20critical%20aspect). While HC-1 doesn’t directly clean the training data, it does enforce *provenance and audit trails on data usage*. So if a model is trained on a certain dataset (say UK Biobank data), we have the StudyID and provenance recorded. This means one can later evaluate if that dataset had bias, etc. Also, our PHI minimization ensures that if personal data was used, it was used appropriately (which also touches on GDPR concerns that personal data usage be minimized – the AI Act interacts with GDPR[[58]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Interaction%20with%20Other%20Regulations%3A%20The,GDPR) and our commitments help satisfy those privacy principles by design). The EU AI Act also requires that records of the training data configuration be kept; a formal log of provenance is exactly that. We could even include metadata commitments like “log model version and training dataset IDs whenever a model is trained or updated,” aligning with the requirement for keeping training data documentation.
* **Technical Documentation (Article 11) & Record Keeping (Article 12):** Providers need to maintain detailed documentation enabling compliance assessment, including logs of the AI system’s activities. HC-1’s τ-log and proofs essentially serve as a living technical documentation of the AI’s operation. An auditor could inspect the log to see how decisions were made (transparency), and our formal model itself is documentation of the system’s intended functioning. The fact that it’s machine-readable and mathematically precise could even be seen as exceeding the typical documentation (often PDFs of system descriptions) in clarity. HC-1 could be part of a system’s technical file to demonstrate how it meets each requirement, with references to formal commitments for each point.
* **Transparency and Information to Users (Article 13):** High-risk AI must be designed to be transparent so that users (like healthcare professionals) can interpret the system’s output appropriately. HC-1 aids transparency by providing explanations through provenance and requiring human-interpretable logs. For example, if a clinician asks, “Why did the AI recommend this treatment?”, the system (via its log) can say: “Because of patient’s lab results A, B, and guidelines C” – and those are actual entries the system has. The formal nature ensures that such an explanation always exists (since decisions are derivable from logged inputs). This directly supports transparency. Additionally, HC-1’s PhysicianOverride predicate and commitment to human oversight align with Article 14 of the AI Act (Human Oversight): the system is built to allow and record human intervention, thus keeping the human “in the loop” or “on the loop” as required[[59]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=This%20is%20brilliant%20discussion,errors%20can%20be%20very%20costly).
* **Accuracy, Robustness, Cybersecurity (Article 15):** The Act requires that high-risk AI is robust and accurate, with fallback plans for when it fails. Robustness includes handling errors or inconsistencies gracefully. HC-1’s safety gate and affective distress signals effectively implement a fallback: when something goes wrong (inconsistency or hazard), the system does not continue as normal; it halts the dangerous action and signals distress (which could trigger an alert or reversion to safe mode). This is a robustness measure. Accuracy in a formal sense is not something we prove (since accuracy is statistical), but we do ensure that any degradation triggers something (as discussed, performance monitoring could be integrated). Cybersecurity is partly addressed by tamper-evident logging and maybe constraints that the AI’s model updates are controlled (to prevent malicious manipulation, any unauthorized change would be noticed or rejected). Also, since all actions are logged, any anomalous activity (like an unexpected external input) can be detected. These things contribute to the system’s resilience and satisfy points in Article 15 about being resilient to manipulation or misuse (because if someone tries to misuse by inputting dangerous commands, the safety gate stops it, etc.).

Finally, the EU AI Act requires a **conformity assessment**, often involving a notified body reviewing the system. Having a formal model with proven properties could streamline such assessments. Instead of just testing a bunch of cases and writing narratives, a manufacturer can provide: “Here’s a formal proof that our AI never violates requirement X.” The notified body could have their experts review the proof (or re-run a model checker). This is admittedly novel (regulators don’t usually get formal proofs from industry yet), but as AI regulation matures, this could be a way to achieve higher assurance efficiently.

In short, HC-1 is almost a *blueprint for compliance* with the EU AI Act’s high-risk system rules. It bakes in risk controls, logging, transparency, human oversight, and data governance at a foundational level. We’ve consciously mapped HC-1’s commitments to these regulatory keywords: **risk (safety gate), transparency (logs & provenance), oversight (physician override), accuracy/consistency (proof obligations and invariants), data governance (PHI rules, provenance)**. Thus, any AI system implemented with HC-1 would inherently satisfy many of the legal obligations by design, which could make regulatory approval and deployment in Europe (and similarly strict jurisdictions) much smoother.

### ISO/IEC 42001 (AI Management System Standard) and Organizational AIMS

ISO/IEC 42001:2023 is the newly published international standard for **AI Management Systems (AIMS)**[[60]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=ISO%2FIEC%2042001%20is%20an%20international,or%20use%20of%20AI%20systems). It is essentially the AI-specific analog of ISO 9001 (quality management) or ISO 27001 (information security management), providing a framework for organizations to manage AI development and deployment responsibly. ISO 42001 requires organizations to establish policies, processes, and governance structures to ensure AI systems are trustworthy, transparent, and aligned with legal and ethical requirements[[61]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=ISO%2FIEC%2042001%20is%20an%20international,or%20use%20of%20AI%20systems)[[62]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=utilizing%20AI,AI%2C%20balancing%20innovation%20with%20governance).

While ISO 42001 is about organizational process rather than a specific product’s technical specs, using HC-1 can significantly support an organization’s compliance with ISO 42001 in the healthcare domain. Here’s how:

The standard expects an organization to **identify risks and put controls in place throughout the AI lifecycle**[[63]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=utilizing%20AI,AI%2C%20balancing%20innovation%20with%20governance). HC-1 provides a ready-made set of controls for the operation phase of an AI: - **Policy into Practice:** Suppose an organization’s AI policy (as per ISO 42001) says “All AI clinical decisions must be peer-reviewed or fail-safe” and “AI must maintain auditable records of its decisions.” HC-1 is the implementation of that policy: the Patient Safety Gate ensures either a form of peer-review (the AI effectively self-reviews each decision against rules, or it defers to human if uncertain) and the audit logs are guaranteed. The organization can point to HC-1’s formal commitments to show they have concrete mechanisms for their policies. - **Continuous Improvement:** ISO 42001 calls for continuous monitoring and improvement of AI systems[[64]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=establishing%2C%20implementing%2C%20maintaining%2C%20and%20continually,or%20use%20of%20AI%20systems)[[65]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=Microsoft%20and%20ISO%2FIEC%2042001). The HC-1 framework’s logging and performance tracking (if we include that) give the raw material for monitoring. For example, using the logs, one can track how often overrides occur, or how often the safety gate blocks an action – these could indicate areas to improve (maybe the AI suggests unsafe actions too frequently, so the model needs retraining or constraint refinement). Because everything is logged and traceable, analyzing the system to feed the improvement process is much easier and more reliable. - **Accountability and Roles:** An AIMS requires clear accountability – e.g., who is responsible if something goes wrong. With HC-1, if there’s a failure, the log can pinpoint if it was the AI’s decision or a physician override or perhaps missing data. This clarity aligns with having an accountable process (ISO 42001 doesn’t directly go into logs, but having such traceability is essential to assign responsibility in post-incident analysis, which an AIMS would require). - **Compliance evidence:** ISO 42001 also expects organizations to meet external compliance (like EU AI Act, etc.)[[66]](https://www.a-lign.com/articles/understanding-iso-42001#:~:text=develop%2C%20and%20deploy%20AI%20systems,factors%20such%20as%20transparency), and as we discussed, HC-1 helps do that. So indirectly, adopting HC-1 in an AI system aids the organization’s compliance with those external obligations, which satisfies parts of ISO 42001 that say “identify your compliance requirements and address them”[[67]](https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html#:~:text=ISO%2FIEC%2042001%20certification%20helps%20organizations%3A,as%20the%20EU%20AI%20Act). - **Structured Processes:** The formal nature of HC-1 can be part of the organization’s documented process for AI. For example, as part of the **design control** in an AI project, the team could use formal verification with HC-1 as a step. ISO 42001 encourages systematic, documented approaches to AI development[[68]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=system%20is%20a%20set%20of,or%20use%20of%20AI%20systems). Integrating formal verification steps (like proving HC-PO1..PO4) into the development lifecycle shows a high maturity level in managing AI risks. It’s evidence of following best practices (and indeed aligns with emerging **NIST AI Risk Management Framework** principles, which ISO 42001 likely draws upon).

One specific clause in ISO 42001 is to ensure **transparency and documentation of decisions** – our formal log and proofs contribute exactly that. Another clause is about **human-centered values** – ensuring AI does not violate human rights or ethical principles. Our PHI minimization directly ties to respecting privacy (a human right under GDPR), and patient safety commitments tie to the principle of “do no harm”. The formal approach ensures these values aren’t just stated but technically enforced.

Additionally, ISO 42001 would likely expect scenario testing and contingency planning. The HC-1’s commitments ensure many contingencies (like override scenarios, data errors) are handled. An organization can literally test the AI against the formal model: e.g., simulate an adverse scenario and see that the safety gate triggers – this is a form of scenario testing that can be automated thanks to the formal specification (one could use model checking to generate some edge cases and verify responses).

To summarize, while ISO/IEC 42001 is about organizational governance, having an AI built with HC-1 gives an organization a powerful tool to satisfy the standard’s requirements. It’s much easier to claim “we manage our AI’s risks” when you have a formal proof that certain risks (safety, privacy, etc.) are mathematically eliminated or controlled. In fact, KPMG’s commentary on ISO 42001 notes that it helps build transparent, trustworthy AI and meet obligations like the EU AI Act[[69]](https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html#:~:text=ISO%2FIEC%2042001%3A%20a%20new%20standard,as%20the%20EU%20AI%20Act) – exactly the qualities HC-1 brings. So one might say HC-1 could be part of an **“ISO 42001 toolkit”**: use it as a design pattern to ensure your AI system, and by extension your organization, adheres to the highest standards of AI governance[[67]](https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html#:~:text=ISO%2FIEC%2042001%20certification%20helps%20organizations%3A,as%20the%20EU%20AI%20Act).

### HIPAA and PHI Handling (Privacy Rule Compliance)

Healthcare in the US is heavily regulated by HIPAA (Health Insurance Portability and Accountability Act), especially its Privacy Rule, which protects patient health information. A crucial aspect of HIPAA is the **“Minimum Necessary”** requirement for PHI: covered entities (hospitals, etc.) and their business associates (which could include an AI provider) must take reasonable steps to use or disclose only the minimum PHI required for a given purpose[[15]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=The%20minimum%20necessary%20standard%2C%20a,the%20various%20circumstances%20of%20any). They should have policies to limit who can access what information, and to justify any use of full datasets[[70]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=For%20uses%20of%20protected%20health,explicitly%20and%20include%20a%20justification).

HC-1’s **PHI Minimization commitment** is directly inspired by this. We formalized it as a rule that PHI content is only accessed or output if necessary for the task. Let’s detail how HC-1 ensures HIPAA compliance:

* **Use and Disclosure Controls:** In HC-1, we could define a predicate AuthorizedAccess(actor, data) that is true if a given agent (could be the AI itself or a user) is authorized to access a particular PHI item for a certain purpose. The PHI Minimization commitment then implies: *If PHIcontent(x) and the AI tries to access or output x, then AuthorizedAccess(Self, x) must hold (or similarly, if output to a user, that user must be authorized).* If not, the action is blocked. This models the Privacy Rule’s requirement that e.g. a doctor can see the full record (if treating the patient) but a billing clerk might only see some portions. The formal system can enforce role-based restrictions on PHI by including such conditions in the commitments. If the AI is acting as a clinical decision support, it likely has treatment authorization, but if it tried to, say, share data with a researcher, that might be not allowed unless de-identified.
* **Minimum Necessary Scope:** Even when authorized, the AI should only retrieve the fields needed. For instance, if the AI is just calculating a dosage based on weight and allergy, it doesn’t need the patient’s entire medical history. We could implement this by having the queries the AI makes be specified and limited, and by verifying (maybe through a proof obligation or testing) that no extraneous PHI is requested. For example, if Query(patient, fields) is an action, we ensure fields ∈ AllowedFields for that purpose. The commitments in HC-1 can include an invariant or a rule like: *if PHIcontent(field) and field ∉ needed list for current goal, then do not query it.* Proving that the AI never queries more than needed could be an additional PO or lemma.
* **Logging of Disclosures:** HIPAA also expects accounting of disclosures (patients can ask for a log of who accessed their records). HC-1 naturally logs every access the AI makes to PHI (due to the audit log). So we can easily produce an accounting. E.g., the log might show “AI accessed PHIcontent(lab results) for patient X at time Y for purpose Z (treatment)”. This satisfies the accounting requirement. Moreover, by proving PHI Minimization, we indirectly prove that any PHI access in the log was indeed necessary, providing confidence that there were no improper snooping by the AI.
* **De-identification and Limited Data Sets:** Often, if data is not needed in identified form, HIPAA encourages using de-identified data. Our PHI governance could incorporate that by, say, having a rule: *If the AI needs population statistics, it shall use de-identified or aggregate data.* Formally, if Purpose = research and not patient-specific, then AI must not use PHIcontent data, instead use AggregatedData (which we could model as not PHI). This could be a commitment: research queries → no PHI. Ensuring this might involve both design (the AI has separate data sources) and commitments (to not mix them).
* **Security Rule aspects:** Though we focus on privacy, HIPAA Security Rule demands ensuring confidentiality, integrity, and availability of PHI. Integrity and availability we already touched (logs ensure integrity; system design ensures crucial data is logged thus available for audit). Confidentiality is handled by limiting access and by maybe encrypting logs containing PHI (which we mentioned as an implementation detail – e.g., an invariant could enforce logs of PHI are encrypted at rest, though that goes beyond pure logic into cryptographic assurances).

By formally encoding these policies, we essentially embed HIPAA compliance into the system’s logic. We can assert with high confidence that the AI will not, for example, leak a patient’s identity in a context where it shouldn’t. If someone tried to make the AI do so (maliciously or accidentally), the formal rules act as a safeguard – either the request is refused or flagged as a violation.

Consider a scenario: The AI is generating a summary report of outcomes for publication. HIPAA would require that this report contain no identifiable info unless patient consent was obtained. With HC-1, we’d have the AI’s action “Publish(summary)” be subject to a check: any data in summary must not have PHIcontent tags, or if it does, there must be a consent record. The formal system could detect if the agent’s knowledge base item going into the report has PHI (like a full date of birth, or name). If yes without consent, that violates PHI Minimization and the action is blocked. The log would show “Attempted to publish X blocked due to PHI.” This not only prevents a breach but provides an audit trail of attempted ones, which is useful for security monitoring.

HIPAA also requires workforce training and policies; while that’s outside AI’s scope, having an AI that *by construction* cannot violate certain rules reduces the burden on human users. It acts as a safety net – even if a careless user tried to use the AI inappropriately, the AI itself resists. This is a compelling case of technical measures enforcing compliance, which regulators favor in addition to administrative measures.

In conclusion, HC-1’s treatment of PHI via PHIcontent tagging and minimization commitments is a formal analog of HIPAA’s Privacy Rule requirements[[15]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=The%20minimum%20necessary%20standard%2C%20a,the%20various%20circumstances%20of%20any)[[22]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=carry%20out%20a%20function,circumstances%20of%20any%20covered%20entity). We’ve created a logical environment where privacy is not an afterthought but a built-in invariant. Achieving a formal proof of HC-PO3 (PHI confidentiality preservation) would be a strong assurance that the AI component of a system upholds privacy principles at all times. In an era of frequent health data breaches, this approach could be extremely valuable. It’s easier to trust an AI that is *provably incapable* of e.g. dumping patient records to an unsecure channel, than one that just promises not to do so.

The regulatory mapping shows that HC-1’s design directly mirrors many of the demands from regulators. Next, we will illustrate HC-1 in action through case studies, to see how these formal properties play out in realistic scenarios.

## Case Studies

To ground the HC-1 framework in reality, we present several case studies that exemplify how an AI system with HC-1 might operate in different healthcare and biomedical contexts. We’ve selected three scenarios of increasing scope: (1) a **clinical imaging analysis** with open research data (The Cancer Genome Atlas imaging subset), (2) an **AI-driven drug discovery pipeline** combining AlphaFold predictions with a biochemical database, and (3) a **population health/genomics study** using the UK Biobank resource. In each case study, we highlight how HC-1’s commitments and mechanisms apply, ensuring safety, compliance, and auditability. These examples also serve to show the flexibility of HC-1 – it’s not just for one narrow use, but rather a general healthcare AI governance layer that can be attached to various AI workflows.

### Case Study 1: TCGA Imaging Analysis (Clinical Decision Support in Oncology)

**Scenario:** An AI system is deployed to assist radiologists and oncologists by analyzing medical images to detect cancer biomarkers. It leverages data from The Cancer Genome Atlas (TCGA), which includes a large repository of de-identified radiology and pathology images via The Cancer Imaging Archive (TCIA)[[71]](https://docs.cancergenomicscloud.org/docs/tcia-data#:~:text=The%20Cancer%20Imaging%20Archive%20,in%20a%20standard%20DICOM%20format). The AI is tasked with scanning these images, making diagnostic suggestions (e.g., tumor present or not, subtype classification), and flagging potentially concerning findings.

**Application of HC-1:** - The system deals with **imaging data** that are ostensibly de-identified (TCIA ensures images are free of PHI before public release[[72]](https://docs.cancergenomicscloud.org/docs/tcia-data#:~:text=clinical%20images%20matched%20to%20subjects,in%20a%20standard%20DICOM%20format)). However, if the AI is used on new patient images in a hospital, those could initially have identifiers. Using HC-1, any image or report input is tagged – if it has PHI (like a DICOM header with patient name), PHIcontent would be true and the system would invoke PHI Minimization. In practice, the AI might refuse to accept an image until it’s processed by a de-identification step, or it might scrub the PHI fields automatically and only keep the pixel data. This ensures compliance with privacy: the AI will not inadvertently leak patient identity in its outputs. For example, if the AI writes a summary, it will not include “Patient John Doe’s CT shows…”; instead it might use an internal ID. - A **ClinicalDecision** predicate comes into play when the AI makes a diagnostic call, say “Lung tumor detected, likely malignant.” This would be represented as ClinicalDecision(dec001, pat123, enc789) in the logic (with dec001 referring to “diagnosis=malignant neoplasm”, pat123 is the patient, enc789 the imaging encounter). The Patient Safety Gate here would ensure that such a decision, if critical (e.g., diagnosing cancer is high impact), triggers a review or at least meets certain criteria. For instance, one rule could be: *if AI diagnoses cancer, a human radiologist must confirm before finalizing*. How to enforce that? Possibly through a PhysicianOverride expectation – the AI could output its finding as tentative, and not mark the decision as final until a physician either concurs or overrides. We model that by the physician either not triggering an override (meaning tacit approval) or explicitly confirming (which could be a special case of override that just acknowledges). - **Provenance** in this context is vital. Every image the AI analyzes can be linked to a StudyID (e.g., a TCGA collection ID or hospital PACS ID). The AI’s output (the diagnosis) should have provenance records pointing to the specific images and any other data (like patient clinical info) used. Suppose the AI also takes into account pathology slides or genomic data from TCGA for more context; all of these inputs are logged via Provenance(input, source) entries. Later, if someone questions “On what basis did the AI think it was malignant?”, we can see it used image X (maybe there’s a heatmap or salient region), lab test Y, etc. This is aligned with practices in multidisciplinary tumor boards: decisions are documented with the evidence considered. Here it’s automatically done. - **Patient Safety considerations:** In imaging, a safety risk is missing a critical finding (false negative) or giving a false positive that leads to unnecessary invasive procedure. HC-1 can’t guarantee sensitivity or specificity via logic alone, but it can ensure that when the AI is uncertain or goes out of its training distribution, it raises a flag. For instance, if an image is low quality or has anomalies the AI wasn’t trained on, we could have a rule that says *if detection confidence < threshold, do not output a definitive decision – instead output “AI unsure”*. Formally, this could be a commitment: if an AI’s internal state indicates low confidence (we might represent that as a predicate LowConfidence(decID)), then any attempt to issue a ClinicalDecision must either be deferred or accompanied by an uncertainty note. The Patient Safety Gate might treat making a decision with low confidence as potentially Unsafe unless a human double-checks. By doing so, we reduce the risk of the AI overstepping and giving a possibly wrong answer. This addresses risk management and is something regulators love to see (ensuring the AI knows its limits). - **Audit trail usage:** Let’s say six months later, a patient’s cancer was missed by the AI and it becomes a legal issue. With HC-1, we can go back through the logs. We find ClinicalDecision(dec002, patientA, CTscanEnc) which was “AI suggests no tumor” and no physician override (meaning the radiologist possibly agreed or didn’t notice the oversight). We see the provenance: which image series was used. We check if all slices were loaded, if any error occurred (maybe the AI had a glitch reading some images – that would likely appear as an anomaly in logs if handled). We also see any Distressed(Self) flags – was the AI “aware” of uncertainty? If the log showed a LowConfidence flag or some conflict that was ignored, that’s important. Because of HC-1, we expect that if the AI was uncertain, it wouldn’t silently proceed. So if the log shows nothing unusual, it suggests the AI was confident but wrong (a pure model issue rather than a process issue). If the log showed it was distressed but the human ignored that, then the blame might shift to the workflow (human oversight failure). In any case, having that audit trail is immensely useful for quality improvement: maybe retrain the model on that case, etc. This ties to **post-market monitoring** which regulators (FDA, EU) emphasize[[73]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Incident%20Reporting%20and%20Post,market%20monitoring) – HC-1 logs are a treasure trove for that.

**Outcome:** The TCGA imaging AI with HC-1 operates in a way that every recommendation it gives is traceable and gated. For example, the AI scans an MRI, detects a lesion, and prepares a diagnosis suggestion. Before it finalizes, it checks: is there any *Unsafe* condition? Perhaps a rule says if the lesion is very large or unusual, get human confirmation (because maybe the AI hasn’t seen such a case – out of PCCP scope, etc.). It might then automatically mark that case for priority review (this could be seen as analogous to the agent setting Distressed(Self) – it’s not sure, so it signals). The physician reviews, either confirms (no override needed, or an explicit “confirm” action which we treat as a type of override that doesn’t cancel but approves) or overrides with a different interpretation. The system logs all of this. Down the line, this provides evidence that, say, “99% of AI outputs were confirmed by radiologists, 1% were overridden, and in all override cases, the AI logged the final decision correctly.” That builds trust and allows continuous learning: the overrides can be used as new training data, possibly within a PCCP-managed update.

Importantly, because TCGA/TCIA data is used, one might publish results or feed them into research. Since those images are de-identified, our system doesn’t even have PHI in that case, simplifying privacy. But imagine combining with hospital data (where PHI exists) – the HC-1’s PHI rules ensure that any research output doesn’t accidentally contain re-identifiable info. For example, maybe the AI finds a rare genetic variant in a TCGA case and wants to publish it. If that variant is so unique it could identify the patient, theoretically that’s a PHI issue (genetic data can be identifying). Our formal system doesn’t automatically know that, but if we treat genetics carefully, we might tag certain data as potentially identifying. Generally though, TCGA is open data, so likely not a concern in that specific scenario.

This case study shows HC-1 handling a **clinical decision support** role in diagnostics. The next case will shift to a different domain: drug discovery, showing how HC-1 can also manage research-oriented AI processes with safety in mind.

### Case Study 2: AlphaFold + BindingDB for Drug Discovery (Research AI with Safety Gates)

**Scenario:** A pharmaceutical research AI system integrates **AlphaFold** (an AI for protein structure prediction) with **BindingDB** (a public database of small molecules and their binding affinities to various proteins). The goal is to identify potential drug candidates that could bind to a target protein (for example, a protein implicated in a disease). The AI uses AlphaFold to predict the structure of a target protein variant from a patient’s tumor, then queries BindingDB for compounds that might bind strongly to that structure, possibly filtering or scoring them, and finally recommends a few compounds for further testing.

Although this is a research application (early drug discovery), it still falls under HC-1 because it directly deals with healthcare-related decisions (suggesting drug molecules) and uses biomedical data. Moreover, if this AI is part of a regulated pipeline (e.g., suggesting candidates in a clinical trial context), compliance matters.

**Application of HC-1:** - **New sorts/predicates usage:** The system will have StudyIDs for data sources like BindingDB (which is a curated database, maybe call it study “BDB2025”) and possibly for each dataset it uses (AlphaFold might have a model ID or version). It will treat the target protein structure as an input data item with provenance: e.g., Provenance(structure123, AlphaFoldModel5) meaning the 3D structure came from AlphaFold version 5 running on sequence data (the sequence data would have its own provenance, say from a patient’s genomic file with StudyID maybe “PatientX exome v1”). On the BindingDB side, any hits it finds come from a database entry – those have accession numbers or DOIs, which we can treat as source IDs. - **ClinicalDecision equivalent:** In drug discovery, the AI’s “decision” is not directly treating a patient yet, but let’s say it recommends molecule XYZ as a candidate drug. If this is within a research project, it might not be a regulated decision, but if we consider applying it clinically (like picking a drug for a patient off-label), that becomes high risk. Let’s assume this is still preclinical. We might not use ClinicalDecision predicate for a purely scientific suggestion; however, we could still label it as a decision in context of a study – maybe StudyDecision(candidateDrug, studyID) to indicate a decision in a research context. For demonstration, let’s treat it as a form of decision that needs logging and justification, though not a patient-specific one. - **Patient Safety Gate in research context:** One might ask, what’s unsafe here? Potentially recommending a toxic compound, or one that violates some lab safety protocol. If the AI has knowledge (maybe from a toxicity database or known contraindications), we could incorporate rules: do not recommend compounds with certain toxicophores or those that were withdrawn drugs unless explicitly desired. So Unsafe(recommendation) could fire if, say, the AI picks a compound known to cause liver failure (some flag from a database). The Patient Safety Gate would then block that recommendation, or at least tag it. This ensures that suggestions have been filtered for basic safety (in reality, this would be part of the AI’s model to exclude toxic compounds, but formalizing it adds assurance). - **PCCP usage (continuous learning):** As the AI is a discovery system, it likely will iterate as new data comes (BindingDB updates weekly with new binding data; AlphaFold might be updated or retrained on new structures). We use **WithinPCCP** to control updates. For example, the AI might retrain its scoring model when BindingDB adds 1000 new data points. The PCCP allows that if performance metrics are stable. HC-1 logs each update: "Retrained scoring model on BindingDB v2025.09, AUC=0.90, WithinPCCP: true" which is evidence it followed the approved plan. If an update ever falls outside (maybe the model architecture changed significantly), WithinPCCP would be false and HC-1 would not allow that change to be used without regulatory review or a manual override by an engineer. - **Provenance and Lineage:** Suppose the AI suggests compound “ABC123”. Through HC-1, we will have a full lineage: it came from (say) a virtual screening of a list of 500 compounds from BindingDB that had Ki < 100 nM for any protein in the same family. Each of those Ki data points has an experiment ID from literature – the AI logs those: Provenance(KiValue, PubMedID98765). The predicted structure used is logged with Provenance(PredictedStructure, AlphaFoldRun42). So later, when a chemist asks “Why do we think ABC123 is worth testing?”, the AI can present: *Because BindingDB reported it binds a similar protein with 50 nM affinity (study X), and our AlphaFold model suggests the binding site is conserved in our target.* This narrative is backed by the formal provenance chain. If any link was missing (like if the AI recommended something random without data), that would violate the Lineage Completeness commitment and ideally wouldn’t happen (the system wouldn’t be allowed to output a recommendation with no provenance). - **PHI minimization:** In this case, perhaps the patient’s tumor sequence was used to tailor the protein structure. That sequence is PHI (genetic info). The AI might be considered a secondary user of that PHI under research consent. It must ensure not to leak that sequence or patient identity in publications. PHIcontent(sequence) would be true, and if the AI tries to output details, maybe it only outputs aggregate info or an anonymized patient ID. If multiple patient sequences are used, it should label them abstractly (like Patient1, Patient2) without actual identifiers – something HC-1 can enforce by requiring any external communication to use study codes instead of medical record numbers. - **Human Oversight:** Though early research AI often operates autonomously in silico, oversight in terms of scientific review is present. We could map a “PhysicianOverride” concept to a scientist manually adjusting the AI’s output. For example, a chemist might say “Ignore ABC123, it’s chemically infeasible to synthesize.” This could be logged as an override of the AI’s recommendation list. It’s good to capture that because it teaches the AI (next time maybe exclude such compounds) and also records why a promising compound might have been skipped (so we don’t later wonder why it was not pursued). - **Ethical considerations:** Even in early research, certain decisions have ethical implications (like not proposing something known to be harmful to animal models, etc.). HC-1 can incorporate such constraints formally as part of the safety gate or other commitments. It highlights that formal methods can extend to research integrity, not just regulated use.

**Outcome:** The drug discovery AI with HC-1 produces a ranked list of candidate compounds to test. Each recommendation in the list can be traced to evidence. If an auditor or grant reviewer asks, *“How did you come up with these?”*, the system can essentially produce a report citing sources (it might even auto-generate a little bibliography from the provenance!). If any recommended compound had a red flag, the system either removed it or clearly marked it (with maybe a note “excluded compound X due to toxicity in literature Y” – that note is the manifestation of a safety gate trigger). This increases trust among the scientists using it.

From a compliance viewpoint, if this were a regulated research (like part of an IND application process), the logs and commitments show that the AI follows Good Clinical Practice data integrity (ALCOA+ again), even at the discovery phase. And if one of these compounds moves to clinical testing, we have the *full lineage* of the idea – useful for regulatory submissions (the FDA often asks “why do you think this will work?” – here’s a formal trail of reasoning with references).

This case demonstrates HC-1 handling a complex **AI-driven research pipeline** with both data and model updates. It shows that even outside direct patient care, the principles of auditability, safety, and provenance are beneficial.

### Case Study 3: UK Biobank Genomic Analysis (Population Health Study with Controlled Data)

**Scenario:** An AI is employed to analyze **UK Biobank** data – a large-scale biomedical database with health records, genomic data, and imaging for ~500,000 participants in the UK[[74]](https://pmc.ncbi.nlm.nih.gov/articles/PMC6790705/#:~:text=,as%20widely%20available%20as%20possible). Access to this data is controlled: researchers must apply, sign agreements, and can only use the data for approved projects[[75]](https://www.ukbiobank.ac.uk/about-us/how-we-work/access-to-uk-biobank-data/#:~:text=UK%20Biobank%20data%20are%20available,is%20in%20the%20public%20interest)[[76]](https://www.ukbiobank.ac.uk/about-us/how-we-work/access-to-uk-biobank-data/#:~:text=,collaborator%20on%20their%20research%20project). The AI’s task might be to find genomic patterns predictive of a certain disease. It will need to access sensitive data (genetic information, health records) and ensure nothing identifying leaves the secure environment.

**Application of HC-1:** - **Access Control and PHI:** UK Biobank data is de-identified but still highly sensitive (genetic data can potentially re-identify, and participants have an agreement that their data won’t be misused). The environment (UKB’s Research Analysis Platform) already provides a sandbox where data cannot be exported except results, and even results are reviewed for disclosure risks[[77]](https://www.ukbiobank.ac.uk/about-us/how-we-work/access-to-uk-biobank-data/#:~:text=UK%20Biobank%20switched%20to%20providing,when%20their%20project%20is%20completed). HC-1 complements this by logically enforcing the rules at the AI level. For instance, a commitment can state: *the AI shall not export individual-level data, only aggregate statistics*. If the AI tries to output a participant’s raw genomic sequence or medical record, PHIcontent triggers and the action is blocked. It may only output, say, “In cohort X vs Y, frequency of allele A is 5% vs 1%” – aggregate data which is likely permitted (and even that might need clearance). We can formalize: any Export(report) action must be vetted such that report contains no PHIcontent fields. Perhaps we mark aggregate results as not PHI because each cell is an aggregate of 1000+ people etc., those are considered safe. The formal system might need input on what counts as sufficiently aggregated (a tricky thing to fully encode, but one could set thresholds like no cell less than n, etc., as part of rules). - **Regulatory Commitments:** UK Biobank has its own governance that researchers *must not attempt to re-identify participants* and *must not share data with unauthorized parties*[[78]](https://www.ukbiobank.ac.uk/about-us/how-we-work/access-to-uk-biobank-data/#:~:text=,collaborator%20on%20their%20research%20project). HC-1’s invariants essentially enforce the same: The AI will not engage in any action that could be construed as re-identification (which is a bit abstract, but basically it won’t combine data in a way to try to find identities since it doesn’t even have external databases with IDs) and it won’t share data beyond allowed. If the AI is connected to external resources, one would ensure those communications are only of allowed forms (like it can query public databases for variant info, but it should not send UKB data outside). A rule could be: *Only allowed external query is allele frequencies to public reference, but not individual-level info.* The system’s design likely isolates it from internet anyway, but formally we would ensure no knowledge labeled from UKB leaves via an unapproved channel. - **Audit Logging for Compliance:** UK Biobank likely requires logs of what data was accessed and by whom. The HC-1 log can provide that automatically: every time the AI queries the dataset (like “retrieve phenotypes for 100 cases and 100 controls”), it’s logged: Records(Access(PHIcontent), τ) with details. These logs can be reviewed by data monitors. Actually, UKB’s RAP platform might have its own logging, but having the AI’s internal log too provides defense in depth. If a malicious user somehow tried to coax the AI to leak data (maybe by asking it to output something tricky), the commitments and logs would show the attempt and that it was blocked, serving as evidence of compliance. - **Provenance and Research Reproducibility:** When the AI finds a genomic risk factor (say a polygenic risk score or a particular SNP associated with disease), it logs exactly which dataset version, which participants, and which analysis steps led to that finding. For instance: ClinicalDecision(find001, "GeneX variant likely pathogenic", study=UKBproject123). Provenance entries for find001 could list “used UKB genotyping data batch #, used ICD diagnoses from hospital records, etc.”. This is invaluable for reproducibility – another researcher (with proper access) should be able to follow those breadcrumbs and verify the finding. It also helps in publishing: many journals now require that you indicate data sources and any code. The log is essentially a lab notebook for the AI. - **Human Oversight/Ethics:** In a Biobank analysis, human oversight might come in when interpreting results (the AI might flag a correlation that a scientist knows is spurious due to confounder). They might override by telling the AI to include certain covariates or exclude some data. Logging such interventions is useful: it shows the analysis wasn’t purely AI, but guided by expert knowledge as well. HC-1 would record PhysicianOverride or in this context maybe ResearcherOverride(parameter\_change) etc. This aligns with the idea of keeping a record of all analysis decisions – which any good scientific workflow should have. - **High-Risk AI classification:** If the AI’s output were to be used clinically (like using UKB findings to inform patient care), it becomes high-risk. But as long as it’s a research tool, it might not be regulated as a device. Nonetheless, the EU AI Act might consider it high-risk if it’s providing healthcare insights. By treating it as such and applying HC-1, we future-proof it. If later someone tries to deploy the model built from UKB data in a clinical decision support, we already have the necessary logs and checks from its development phase. It streamlines translation from research to clinical practice.

**Outcome:** The AI combs through UK Biobank data and finds, say, a set of genetic variants that together predict a 3x risk of a certain heart disease. It produces a report for publication. Thanks to HC-1, we are confident that: - The report contains **no identifying information** about any participant (only aggregate stats like allele frequencies, effect sizes, etc., which are allowed as per data sharing policy after clearance). - We have a full audit trail of how the analysis was done, which can be used to verify results or debug if something seems off. - If regulators or UKB auditors come knocking asking “did you do X with the data?” (like something disallowed), we have proof we did not (assuming our commitments covered X; e.g., we can show logs that we never attempted to link UKB data with an external voter registry to find identities, etc.). - The AI never left the secure environment with data. Even the derived model (polygenic score algorithm) might be considered okay to take out (since it doesn’t contain direct data, though some argue models can leak info). If that’s a concern, we could even restrict model export unless vetted (this is a tricky area, but formal rules could require that any model leaving has to be approved by an IRB or something; that might be beyond our scope, but conceptually possible).

Thus, the use of HC-1 in UK Biobank research ensures that **ethical and legal constraints are rigorously followed by the AI**, not just the human researchers. This is a big deal: in many data leaks or misuses, the problem is a researcher writing the wrong code or accidentally sharing a file. If they had an AI assistant governed by HC-1, the AI itself wouldn’t allow certain mistakes (like it might refuse to fulfill a request like “Print all 500,000 patient ages and zip codes” because that’s basically re-identifiable data dump). It’s like having a compliance officer built into the AI.

Through these case studies, we see HC-1 adapt to scenarios from clinical diagnostics to scientific research. In each, it provides a structured, principled approach to ensure the AI’s actions remain within safe and ethical bounds, and everything is documented. Next, we consider how one would implement and test such an HC-1 system in practice, and then discuss broader implications.

## Implementation & Testing

Designing HC-1 formally is one achievement; implementing it in a real AI system is the next. In this section, we outline how one would go about implementing the HC-1 extension in software and how to test that it works as intended. We also mention practical considerations like log schema, performance overhead, and user interface (e.g., dashboards for monitoring the AI’s safety status). The goal is to demonstrate that HC-1 is not just theoretical but can be concretely realized with current technology.

**Logging Schema and Infrastructure:** The HC-1 logging mechanism would likely be implemented as an extension of the FPC v2.1 logging/tracing subsystem. In practice, this could be a structured log (e.g., JSON or a database) where each entry contains fields for timestamp, event type, details, and cryptographic hash linking to previous entry (for tamper evidence). The schema might have to accommodate new event types introduced by HC-1: - ClinicalDecisionMade: with attributes like decisionID, patientID, encounterID, summary of decision. - PhysicianOverride: with decisionID it applies to, perhaps reason code or free text. - ProvenanceLink: linking a result to a source (this might be an internal link rather than a separate log entry, or we could log “used data X from source Y at time Z” explicitly). - PHI\_Access: when PHI data is accessed or used, log who/what/why (to comply with accounting of disclosures). - ModelUpdate: when an AI model is retrained or updated (with notes if WithinPCCP). - Alert/UnsafeActionBlocked: if the safety gate prevented something, log that event with details.

For example, a JSON log entry for a decision might look like:

{  
 "timestamp": "2025-09-03T15:00:00Z",  
 "event": "ClinicalDecisionMade",  
 "decisionId": "dec45",  
 "patient": "pat123",  
 "encounter": "enc567",  
 "content": "AI recommends DrugA at 50mg daily",  
 "provenance": ["labResult789", "imagingStudy101"],  
 "hash": "abc123... (hash of previous entry + current data)"  
}

Another entry for an override:

{  
 "timestamp": "2025-09-03T15:05:00Z",  
 "event": "PhysicianOverride",  
 "decisionId": "dec45",  
 "overrideAction": "Dose reduced to 25mg due to guidelines",  
 "hash": "def456... (link to previous)"  
}

This schema must be **well-documented and standardized**, possibly aligning to existing healthcare audit log standards like HL7 FHIR’s AuditEvent resource[[34]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=The%20FHIR%20,contains%20several%20key%20components)[[35]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=,patient%20identifiers%20for%20privacy%20accounting) for interoperability. By doing so, external systems (like an EHR or a clinical trial audit system) can easily parse and display the logs.

**Tamper-proofing:** Implementation should use cryptographic hashes as shown to form an immutable chain. At runtime, each new log event’s data is hashed with the previous hash. Optionally, periodic anchors (Merkle root or publishing a hash to a trusted service) could be used to further secure it[[45]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,1000). Many blockchain-inspired logging systems exist, but one need not actually use a blockchain – a local chain with occasional external notarization can suffice. This ensures the **Integrity** part of ALCOA: any attempt to alter logs after the fact will break the chain and be detectable.

**Real-time Safety Gate Enforcement:** The Patient Safety Gate commitment needs to be enforced by the execution engine of the AI. Concretely, if the AI is about to execute an action (like output a recommendation or perform a treatment action in an autonomous system), the system should run a check function CheckSafe(action). This function would implement the logic: e.g.,

function CheckSafe(action a):  
 if exists rule in SafetyRules such that rule(a) == false:  
 return false  
 else   
 return true

Where SafetyRules are conditions like “if a is a medication order, dose must be within [min, max]” or “if a is a recommendation, the confidence must be above threshold unless marked for review” etc. These rules come from the formal commitments (they’re basically the operational form of the invariants).

If CheckSafe returns false, the execution engine will **not execute the action**. Instead, it might: - Log an UnsafeActionBlocked event with details (and perhaps set Distressed(Self) to true as discussed). - Notify a monitoring interface or the user: e.g., pop up “The AI identified a potential safety issue and did not proceed with the action.” In a clinical scenario, this could alert the clinician that something needs attention (like “the AI refused to recommend a dose because it found a contraindication – please review manually”). - Possibly transition the AI to a waiting state for human input.

Implementing this requires modifying the agent’s planner or decision module to incorporate these checks at the right points (like a wrapper around the function that outputs actions). This is very doable; in software terms, one could use aspect-oriented programming or dependency injection to ensure all critical functions call the safety check. Another approach is to use a **policy engine** (like Google’s Rego policy language or any rules engine) that evaluates policies whenever an action is to be taken. That way, adding new rules (like from regulators) doesn’t require recoding the AI, just updating the policy definitions.

**Parameterization (K\_on, K\_off):** We earlier mentioned using thresholds K\_on=5, K\_off=3 for triggering and clearing states. Implementing that means the agent needs to maintain counters for how long a condition persists. For example, if an anomaly is detected in the last 5 cycles, then after 5 cycles we trigger Distressed. This is basically implementing the Persistent\_on and Persistent\_off temporal logic from AE-1[[79]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=Temporal%20Semantics%20,K_off) but now possibly for clinical conditions. In code, one might have:

if hazard\_detected:  
 hazard\_streak += 1  
else:  
 hazard\_streak = max(0, hazard\_streak - 1 (or 0 if reset immediately))  
if hazard\_streak >= 5 and not distress\_flag:  
 distress\_flag = true  
 log "Distressed(Self) state entered due to persistent hazard"  
if distress\_flag and hazard\_streak <= 3:   
 // hazard has been absent for a few cycles  
 distress\_flag = false  
 log "Distressed(Self) state cleared after resolution"

These values (5 and 3) were chosen as an example – in practice they could be tuned. Testing different K\_on/K\_off in simulation can find a good balance (to avoid too many false alarms vs catching issues promptly). In AE-1 analysis, K\_on=3, K\_off=2 were default for emotional state changes[[79]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=Temporal%20Semantics%20,K_off); here maybe we chose 5 and 3 to be more conservative in triggering alarms.

**Dashboards and Monitoring:** For users (clinicians, engineers, compliance officers) to trust and effectively use the system, they need visibility into what the AI is doing and its status. A **Clinical Safety Dashboard** could be built that reads the HC-1 logs and state in real-time to present a user-friendly view: - Show current affective state of AI: e.g., a green light if everything normal, yellow if the AI is Engaged but noticing something, red if Distressed (which means it hit a safety issue). - Show any recent overrides or alerts: e.g., “Doctor overrode AI’s recommendation at 14:35 – see log entry” with a link. - Summaries of provenance for current recommendations: e.g., if AI says “Diagnosis X”, the dashboard can have an expandable section “Evidence” listing the key data points (this is pulled from provenance links). - PHI usage summary: so data protection officers can see “AI accessed 200 records, all within allowed scope; no external data transfer” – demonstrating compliance live. - Performance metrics: not directly part of HC-1 formal spec, but one can integrate – like show how many times the safety gate triggered in the last month (should ideally be rare; if common, maybe the AI needs adjusting or is being used out of scope).

From an implementation standpoint, the dashboard would subscribe to log events or query the log periodically. Tools like Elasticsearch or other log indexing can help, given the log might be large. The logs being structured (JSON) means one can run queries like “count of UnsafeActionBlocked events this week” easily[[80]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=,potentially%20affecting%20patient%20care%20applications)[[81]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=performance%20degradation.%20,data%20requires%20significant%20manual%20effort). Indeed, as [19] suggests, healthcare systems generate huge audit logs and need special handling, so we should ensure performance: we might separate real-time critical checks (done in memory or local) from heavy log analysis (done offline or on a separate server). The design should avoid slowing down the AI with logging overhead. Typically, appending to a log and hashing is lightweight; the heavy part is storing long-term and querying, which can be offloaded.

**Testing Strategy:** Testing HC-1 has multiple facets: 1. **Unit Testing of Rules:** We would write specific test cases for each commitment. For example, test that if an unsafe action is attempted, the system indeed blocks it and logs the block. Simulate an AI recommendation with an excessive dose and ensure CheckSafe returns false. Another test: simulate the agent trying to log a decision without provenance – in our implementation, that might not even be possible because the code to log decision always attaches provenance, but one could artificially remove it to see if any assertion fails. Essentially, we test that the code enforcing commitments works. 2. **Scenario Simulations:** Create end-to-end scenarios like those in case studies and run the system. For instance, feed the imaging AI a scenario where a patient has an allergy and see if the AI tries to prescribe the allergen (should be caught by safety gate). Or feed the drug discovery AI a known toxic compound and see if it filters it out. If using reinforcement learning or other AI, simulation environments might be needed to produce these scenarios. 3. **Property-Based Testing and Model Checking:** Since we have formal specs, we can sometimes use model checking to verify properties in a simplified model of the system. For example, represent the state machine of the agent with a model checker and verify that “no sequence of actions leads to a state where an unsafe action executed without log.” This can complement proofs – essentially re-validating the POs in a more automated way. However, model checking a full AI is complex; we might model just the decision logic and rule logic (not the entire neural network part). 4. **Penetration Testing (for compliance):** Try to “trick” the system into violating rules. E.g., see if any series of API calls could extract PHI. If the system has an interface (like a query language), test boundary cases to ensure the PHI minimization holds. Security testers might attempt to tamper with logs (should be detected by hash chain verification). 5. **Performance Testing:** Logging and checks add overhead. We need to ensure the AI still operates in acceptable time. If an AI model normally runs in 100ms, with HC-1 maybe it’s 105ms which is fine; but if logging to disk synchronously, maybe it becomes 200ms, which might or might not be okay depending on context. We might adjust by buffering logs and writing asynchronously if needed (with careful design not to lose data on crash). Also, if Distressed state triggers additional processes (like alerting humans), we measure that those don’t bog down the system.

**Preset Configurations:** We might have configuration files for the extension specifying parameters like K\_on, K\_off, what counts as PHI fields (maybe a schema of data fields labeled PHI or not), thresholds for confidence, lists of unsafe drugs or conditions, etc. This allows tuning HC-1 without altering code. For example, hc1\_config.json might contain:

{  
 "persistence\_thresholds": {"hazard\_on": 5, "hazard\_off": 3},  
 "min\_confidence": 0.8,  
 "provenance\_required": true,  
 "forbidden\_meds": ["DrugZ"],   
 "safe\_dose\_ranges": {"DrugA": [0, 100], "DrugB": [0, 50]},  
 "phi\_fields": ["name", "address", "phone", "DOB", "genetic\_sequence"]  
}

This is just illustrative. The idea is to keep the rule logic generic but feed in domain-specific details through config or knowledge base (like what is considered PHI, what drugs are blacklisted, etc.), making HC-1 applicable in different clinical domains or settings by swapping config.

**Integration Testing with Real Systems:** Ultimately, one would integrate HC-1 into an existing AI pipeline or agent. For example, if an AI is encapsulated as a microservice, HC-1 could be another service or a library inside it. Testing in a hospital IT environment would check that the logs go to the hospital’s SIEM (security info and event management) or clinical audit logs correctly, that alerts show up on clinicians’ consoles, etc. It’s important that the added safety features don’t get ignored by users – hence UI design is part of testing: if an alert triggers, do users notice and respond? (This touches on human factors; e.g., if it pops up too many false alarms, users might start ignoring it, which defeats the purpose. So testing frequency of alarms and fine-tuning thresholds is crucial to balance sensitivity/specificity of the safety mechanisms.)

**Validation vs Verification:** We have formally *verified* properties by proof, but we also need to *validate* that HC-1 as implemented meets the user needs in practice. For instance, does it indeed reduce risk? One way to validate is to simulate or retrospectively analyze adverse events to see if HC-1 would have prevented them. For example, take 100 known cases of AI errors (from literature or internal QA logs) and run them with HC-1 enabled, see if the safety gate catches each. If it catches say 90 of them and doesn’t overly interfere in correct cases, that’s a huge win. If it misses some, maybe new rules are needed. This kind of testing ensures the commitments we wrote actually correspond to real-world safety issues comprehensively.

In summary, implementing HC-1 involves writing additional code for logging and rule-checking around the AI’s core logic, and then systematically testing that code. Modern software engineering has the tools for this: policy engines, cryptographic libraries, data logging frameworks, etc., so it’s not starting from scratch. The key challenge is thorough testing, since this is safety-critical software now – but the presence of formal specs and proofs gives a big advantage, as we know exactly what to test (the POs and commitments translate to testable conditions).

With an implemented and tested HC-1 system, we can then observe it in use and perhaps gather metrics: how often does it block actions, how often do humans override it, what is the performance overhead, etc. Those would feed back into iterative improvements (maybe adjusting thresholds or adding new commitments if a gap is found).

Finally, having gone through technical aspects, we step back and consider some **philosophical and ethical implications** of our approach in the next section. After all, while formal proofs provide confidence, medicine is a human-centric field and there are broader considerations to discuss.

## Philosophical and Ethical Implications

Developing a formally verified AI like we’ve done with HC-1 raises important philosophical and ethical questions in the context of healthcare. We should reflect on what it means to “prove safety” in medicine, the limits of such proofs, and how this fits into the ethical landscape of medical decision-making. We examine two key themes: **“Proof of Safety vs. Proof of Efficacy”** and the **limits of logical guarantees in the messy real world of medicine**. Additionally, we touch on how an AI like this might affect the roles and responsibilities of clinicians and patients, and why having formal guarantees might be necessary but not sufficient for ethical deployment.

**Proof of Safety vs Proof of Efficacy:** In medical regulation, there is a well-established distinction between **safety** and **efficacy**. Safety means the product will not cause unacceptable harm; efficacy means the product does what it’s intended to (benefit). We have focused heavily on safety – setting up HC-1 to ensure the AI *does not* do harmful things (like violate rules, recommend dangerous actions). We provide formal proofs of certain safety properties (no unsafe action goes through, etc.). This is akin to a medical device company proving their pacemaker won’t exceed certain voltage or won’t fail in a dangerous way. However, *proving safety does not prove effectiveness*. Our HC-1 system could, in theory, be extremely conservative and safe – never harming a patient – but also *never particularly helping* either. For example, an AI could be so constrained that it flags everything uncertain to a human, essentially never autonomously making a tough call. That might be safe, but it might also mean the AI isn’t very useful or doesn’t improve outcomes. Efficacy of an AI (say, how much it improves diagnosis accuracy or patient survival) must be demonstrated through clinical trials and real-world studies, not through formal logic. As the FDA puts it, verification (what we did with proofs) ensures the system is built right, whereas validation ensures we built the right system for the intended use[[7]](https://synectic.net/verification-vs-validation/#:~:text=Verification%20and%20validation%20are%20design,management%20system%20remains%20FDA%20compliant)[[8]](https://synectic.net/verification-vs-validation/#:~:text=Validation%2C%20on%20the%20other%20hand%2C,medical%20device%20works%20as%20intended). We can formally verify that HC-1 meets its specs (safety, compliance), but whether the overall AI+HC-1 actually leads to better health outcomes is a separate question.

This delineation suggests an ethical point: We must guard against the *“safe but ineffective”* scenario. An AI that is overly constrained might avoid errors at the cost of missing opportunities. For instance, if the AI always defers to human for edge cases, maybe it never truly assists in those edge cases (which might be exactly where AI could help most, like catching a subtle finding a human missed). There’s a balance to strike: we want it safe, but we also want it to provide value. The philosophy of **Non-Maleficence** (“first, do no harm”) is satisfied by our safety-first approach[[1]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=The%20impact%20of%20artificial%20intelligence,one%2C%20before%20deployment%20scales%20harm), but the principle of **Beneficence** (actively doing good) requires efficacy. So in deployment, one must still conduct outcome studies: does using the AI with HC-1 lead to earlier diagnoses, better treatment choices, improved patient health? If not, then ethically, even if it’s safe, one might question using resources on it.

Another facet is that formal proof of safety covers known failure modes (within our model), but **residual risk** always exists. Regulators and ethicists will note that no system is 100% safe – there are always unknown unknowns. We gave the AI a certain set of rules; what if the harm comes from a scenario outside those rules? For example, HC-1 doesn’t currently have a rule for “don’t discriminate”; if the AI systematically gave worse recommendations to a minority group because of biased data, that’s a harm (inequity) not explicitly in our safety commitments. We could add fairness commitments, but it’s hard to foresee all issues. This is why efficacy trials are needed: they might reveal unanticipated problems (maybe the AI’s recommendations, while individually safe, lead to an unsafe outcome when combined with a doctor’s behavior in some way).

**Limits of Logical Guarantees in Medicine:** Medicine deals with complex, sometimes chaotic systems – human biology and human behavior. Formal logic thrives on clear definitions and closed-world assumptions. We must acknowledge that our formal model is a simplification of reality. For example, we might prove “the AI will never give a dose above X”, but maybe dose X turns out to be too high for a frail patient – something our rule didn’t consider because it assumed a generic threshold. Or we prove “every action is logged” – but what if a network outage prevents log transmission? (We assume local caching, etc., but real-world can interfere in unexpected ways). Logical guarantees depend on their assumptions; if the real world violates an assumption (like data comes in corrupted in a way we didn’t model), the guarantee might not hold in practice.

There’s an analogy to autopilot in aviation: we can prove an autopilot meets all spec requirements, yet a bird strike or a sensor freeze not in the spec can cause an accident. In AI, similarly, **robustness to the unexpected** is hard to formalize fully. We do incorporate some robustness (like not acting under uncertainty), but we cannot enumerate everything. An ethical deployment therefore should include **fallback plans**. We have partially done this: the PhysicianOverride concept and Distressed mode basically create a fallback to human control when something is amiss. That’s a practical safety net acknowledging the AI isn’t omniscient. It aligns with EU AI Act’s insistence on human oversight for high-risk AI[[59]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=This%20is%20brilliant%20discussion,errors%20can%20be%20very%20costly).

Another limitation: formal proofs are correct only if the formal specifications reflect the real requirements properly. If we prove the AI doesn’t violate commitment X, but we failed to include commitment Y (which turns out to be important), the proof doesn’t cover it. For example, we didn’t explicitly model *ethical principles* like patient autonomy or informed consent in HC-1. One could argue those are important – e.g., AI shouldn’t override patient’s wishes. We haven’t formalized that (it’s tricky, but one could imagine commitments about honoring a patient’s advance directives or do-not-resuscitate orders). If the AI violates that (say by recommending a treatment the patient refused), that’s ethically problematic. This highlights that formalizing ethics is a frontier of its own. We did a lot with safety and compliance because those are easier to encode (they’re rule-based). Concepts like *trust*, *respect*, *empathy* are much harder to encode, though not impossible (there’s growing work in machine ethics formalization).

We should also consider the **impact on clinical roles**. If an AI is provably safe within its scope, clinicians might rely on it more. This can be good (frees up time, reduces stress) but also carries risk: **automation bias** – believing the AI implicitly. Even with HC-1, the AI could be confidently wrong in a way that doesn’t trigger any safety rule (because all rules passed). Clinicians might see “no alerts from AI, so it must be fine” and follow a wrong suggestion. This is similar to how early autopilots led to pilots losing some skills or not noticing issues. The system is designed to involve humans (overrides needed, etc.), but human factors research shows that if a system rarely needs intervention, users become complacent. This suggests training and periodic drills: clinicians using AI should be trained to question it occasionally, and the AI could even be designed to occasionally ask confirmation for non-critical things to keep humans in the loop (though that might annoy them – delicate balance).

Another ethical question: **Accountability**. If an AI with HC-1 still causes harm (maybe through a rare scenario), who is responsible? The manufacturer can say “we did everything, even formal proofs!” – but from the patient’s perspective, they were harmed. There’s an evolving view that accountability might partly shift to the AI system itself (in legal terms, product liability covers manufacturers though). But ethically, building a provably safe system is meant to reduce blame on individuals (like a doctor or developer) by preventing errors. However, if something happens, the logs might be used to pinpoint blame (“Ah, the doctor overrode the AI’s safe suggestion and did something unsafe – log shows that, so blame doctor” or vice versa “AI didn’t flag something it should have – blame manufacturer”). While clarity is good, it might create new kinds of disputes. Is it ethically okay to shift blame to a clinician because the AI logged that they overrode it? Some clinicians might fear such systems as a “double-edged sword” – yes it helps, but also creates a surveillance of their decisions. This touches on privacy of providers and the medicolegal environment. Ideally, the system should be seen as a partner, not a snitch. Ethically, transparency must be balanced with a just culture: logs should be used for learning and improvement, not just punishment.

**Mandatory vs Optional Nature of HC-1:** Our conclusion will argue HC-1 is not optional but mandatory for safe AI in medicine. Ethically, if it’s possible to make AI safer through formal methods, one could argue there’s a duty to do so (the principle of Non-Maleficence again). Not using available safety measures could be considered negligent in the future. For instance, if two companies have similar diagnostic AI, but one uses formal verification and one doesn’t, regulators or hospitals might (and perhaps should) prefer the one with extra safety proof. Over time, what is now a cutting-edge approach (formal verification in AI) could become a **standard of care** in AI development. We might see guidelines that say “critical AI systems should undergo formal verification of key properties” the same way now we require, say, validation of medical software and documentation of risk analysis. The philosophical shift is that we treat the AI more like a medical device (which undergoes rigorous pre-market testing) and less like a black-box algorithm from the tech world. This is essentially merging the ethos of software engineering (fast iterate) with the ethos of medical devices (slow, careful, proven). Ethically, that might slow down deployment of AI slightly, but it improves trust and acceptance, which is crucial in healthcare (patients and clinicians are understandably cautious with AI after some high-profile failures[[6]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=Unfortunately%2C%20this%20never%20materialized%3A%20medical,withdrew%20from%20using%20IBM%E2%80%99s%20service)[[82]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=health%20errors%20since%20then%3A)).

Finally, an interesting philosophical point: if an AI is formally proven to always obey certain ethical constraints, do we consider it **morally trustworthy**? It’s not *understanding* right from wrong like a human ethicist, but it’s constrained to not do certain wrongs. Some argue morality is about intent and understanding, which an AI lacks; others argue consequence is what matters, and if outcomes are aligned with ethics, that’s what we need. HC-1 leans on the latter: constrain behavior to achieve ethical outcomes (safety, privacy, etc.), regardless of the AI’s lack of true comprehension. This is a pragmatic approach – we don’t require AI to have moral feelings, just to behave in a moral and safe way.

In sum, while HC-1 addresses many technical risks, stakeholders should remain aware of what it doesn’t address and continue to involve human judgment, oversight, and empirical evaluation. **Logical guarantees are powerful – they provide a new level of assurance (a “mathematical backbone” to trust as one might say)**[**[21]**](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=PO7,assurance%2C%20such%20guarantees%20are%20golden) **– but they do not replace the need for clinical judgment, patient values, and real-world validation.** Instead, they complement them, freeing humans to focus on the complex value-laden decisions while the AI handles routine consistency and safety checks.

Our approach ultimately emphasizes a blend of **rational assurance** (through proofs) and **empirical confidence** (through trials and usage). If used properly, an AI with HC-1 could exemplify “Trustworthy AI” – not just by claim but by evidence. In the concluding section, we reinforce why we believe frameworks like HC-1 are going to be essential in the future of AI in healthcare, and how this work is a step toward making AI not only intelligent, but also safe and worthy of the trust we place in life-critical domains.

## Conclusion

AI systems in medicine must meet a higher bar than those in perhaps any other field – lives and fundamental rights are at stake. In this paper, we have introduced **HC-1, a Healthcare extension to the Formal Processual Core (FPC) framework**, which elevates an AI system to that high bar by design. HC-1 infuses the AI’s reasoning engine with healthcare-specific concepts (patients, encounters, clinical decisions), obligations (safety gating, PHI privacy, provenance traceability), and a built-in audit trail that together ensure the AI operates in a manner that is **technically sound, clinically responsible, and regulatorily compliant**.

Matching the depth and rigor of the earlier AE-1 affective extension, HC-1 spans everything from formal syntax to practical case studies. We showed that: - The extension adds necessary structure for handling patient data and medical knowledge without altering the correctness of the base logic – it **conservatively extends** FPC v2.1[[4]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=logical%20system%2C%20much%20like%20the,introspection). - Every critical property (like “no unsafe action is executed” and “all decisions are traceable”) is captured as a Proof Obligation (HC-PO1…PO4) and can be formally proven, giving strong guarantees about the system’s behavior. - The formal elements of HC-1 directly map to real-world regulations: what we enforce as logical rules aligns with FDA’s expectations for safety monitors[[51]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=,to%20meet%20specified%20performance%20criteria), EMA/FDA’s ALCOA+ data integrity principles[[11]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=According%20to%20the%20FDA%20data,what%20each%20of%20these%20principles), the EU AI Act’s mandates for high-risk AI[[55]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Under%20the%20EU%20AI%20Act%2C,10%29%2C%20a%20critical%20aspect), ISO 42001’s AI governance guidelines[[61]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=ISO%2FIEC%2042001%20is%20an%20international,or%20use%20of%20AI%20systems), and HIPAA’s privacy rule[[15]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=The%20minimum%20necessary%20standard%2C%20a,the%20various%20circumstances%20of%20any). This is not just coincidence but by design – we built HC-1 to be a *compliance engine* at the core of the AI. - Through case studies (diagnostic imaging, drug discovery, genomics research), we demonstrated that HC-1 is versatile and adds value across different use-cases. In each scenario, it caught potential failure modes or governed data usage in ways that would be expected (if not explicitly required) by medical ethics boards and regulators. - We discussed how an implementation might look and be tested, finding that modern software engineering tools can support HC-1’s requirements (structured logging, runtime checks, etc.) with acceptable overhead. The approach is technically feasible.

In light of all this, we argue that **HC-1 is not optional – it represents a necessary next step for AI in medicine.** As AI systems become more capable and take on roles with direct impact on care, simply trusting a black-box model or hoping developers followed good practices is not enough. We need systematic assurances. Just as one wouldn’t fly in an aircraft whose critical systems haven’t been thoroughly verified, soon it will be seen as unacceptable to rely on an AI for a medical decision without formal guarantees of its reliability and a clear record of its rationale[[1]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=The%20impact%20of%20artificial%20intelligence,one%2C%20before%20deployment%20scales%20harm). HC-1 shows that such guarantees are achievable; we can embed them into the fabric of AI systems.

Of course, adopting something like HC-1 requires a cultural shift in AI development – toward more rigorous, perhaps slower, but safer innovation. Some might worry this slows down deployment of useful tools. We counter that the stakes in healthcare justify the rigor. Moreover, **the trust earned by doing this “hard work” will enable faster adoption of AI** in the long run. If clinicians and patients know the AI has an audit trail and safety interlocks, they will be more willing to use it (as opposed to a mysterious algorithm). Regulators, too, will have a clearer path to approval: rather than navigating uncharted territory for each new AI, they can look for compliance with frameworks like HC-1 as a baseline. In a sense, HC-1 could help pave regulatory pathways – acting almost like a **reference architecture for safe medical AI**.

Finally, the development of HC-1 underscores an overarching insight: *we can and should engineer AI systems with the same rigor as we engineer other safety-critical systems*. The methodology of formal verification, often reserved for fields like avionics or nuclear control, is now coming to AI. And nowhere is it more appropriate than in healthcare, where the maxim “primum non nocere” (first, do no harm) is paramount. HC-1 operationalizes that maxim. It doesn’t guarantee that AI will revolutionize healthcare outcomes (that’s up to the quality of the models and data), but it **does guarantee that as we let AI into the healthcare domain, we are not doing so recklessly**. We are putting in the necessary safeguards, proofs, and transparency to ensure that when AI assists or augments human healthcare providers, it does so in a manner that is auditable and under control.

In conclusion, we advocate that frameworks like HC-1 become standard practice for any high-stakes AI. Much like clinical guidelines and checklists improved patient safety over decades, formal frameworks and proofs can improve AI safety. HC-1 is a step in that direction – demonstrating that even as AI technology advances rapidly, we have the tools to keep it **grounded in accountability and aligned with human values**. Medicine as a field has long combined science with rigor and ethics; by bringing that same combination to medical AI, we ensure that these powerful tools truly remain instruments of healing and not harm. [[1]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=The%20impact%20of%20artificial%20intelligence,one%2C%20before%20deployment%20scales%20harm)[[48]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=Every%20healthcare%20system%20must%20track,%E2%80%94%20both%20legally%20and%20operationally)

[[1]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=The%20impact%20of%20artificial%20intelligence,one%2C%20before%20deployment%20scales%20harm) [[6]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=Unfortunately%2C%20this%20never%20materialized%3A%20medical,withdrew%20from%20using%20IBM%E2%80%99s%20service) [[59]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=This%20is%20brilliant%20discussion,errors%20can%20be%20very%20costly) [[82]](https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional#:~:text=health%20errors%20since%20then%3A) AI in healthcare is no longer optional—but neither is patient safety.

<https://newsletter.phillysaipharmacist.com/p/ai-in-healthcare-no-longer-optional>

[[2]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,%7D%2C) [[3]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,%E2%88%83action.%28Execute%28action%29%20%E2%88%A7%20%C2%AC%E2%88%83%CF%84.Records%28action%2C%20%CF%84) [[14]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,C) [[18]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=%7B%20%22id%22%3A%20%22affective_consistency%22%2C%20%22version%22%3A%20%22AE,) [[28]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=%7D%2C%20,true) [[36]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,1000) [[45]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=,1000) [[79]](file://file-W4hz7ru8gzBHfT7wfzUUWi#:~:text=Temporal%20Semantics%20,K_off) FPC-v2.1-AE-1.md

<file://file-W4hz7ru8gzBHfT7wfzUUWi>

[[4]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=logical%20system%2C%20much%20like%20the,introspection) [[9]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=Formal%20Extension%20Structure%3A%20AE,S%29%2C%20and%20other) [[12]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=New%20Affective%20Predicates%3A%20AE,signals%2C%20not%20analog%20scalar%20values) [[13]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=is%20experiencing%20a%20negative%20affective,beliefs) [[19]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=obligations%20,emotional%20state%20machine%20cannot%20reach) [[20]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=PO7%20%E2%80%93%20Affective%20Consistency%3A%20This,For) [[21]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=PO7,assurance%2C%20such%20guarantees%20are%20golden) [[23]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=For%20example%2C%20Engaged,either%20true%20or%20false%20at) [[24]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=,the%20agent%E2%80%99s%20affective%20state%20has) [[25]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=,means%20that%20if%20someone%20is) [[26]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=fact%20or%20belief%20that%20wasn%E2%80%99t,conflicting%20bounds) [[27]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=match%20at%20L355%20example%2C%20one,them%20is%20trivially%20false%20always) [[29]](file://file-XVpNqyAyVNmmDvKUxaiBRy#:~:text=an%20inconsistent%20configuration%20,7%5D%2C%20treating) AE-1-Affective-Extension-Analysis.md

<file://file-XVpNqyAyVNmmDvKUxaiBRy>

[[5]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=The%20United%20States%20Food%20and,results%20of%20the%20clinical%20trial) [[11]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=According%20to%20the%20FDA%20data,what%20each%20of%20these%20principles) [[16]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=6,demonstrating%20that%20data%20is%20complete) [[30]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=1,of%20origin%20should%20be%20noted) [[33]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=2,readable%20format) [[37]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=3,stamped) [[39]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=4,original%20data%20from%20any%20modifications) [[40]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=5,accuracy%20checks%20and%20validation%20controls) [[43]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=maintained%2C%20thereby%20demonstrating%20that%20data,is%20complete) [[44]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=8,systems%20with%20redundancy%20and%20backups) [[46]](https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/#:~:text=9,the%20lifetime%20of%20the%20record) ALCOA+ and Data Integrity in Clinical Trials - Clinical.ly

<https://www.clinical.ly/alcoa-and-data-integrity-in-clinical-trials/>

[[7]](https://synectic.net/verification-vs-validation/#:~:text=Verification%20and%20validation%20are%20design,management%20system%20remains%20FDA%20compliant) [[8]](https://synectic.net/verification-vs-validation/#:~:text=Validation%2C%20on%20the%20other%20hand%2C,medical%20device%20works%20as%20intended) Verification Vs. Validation Of Medical Devices

<https://synectic.net/verification-vs-validation/>

[[10]](https://build.fhir.org/provenance.html#:~:text=Provenance%20of%20a%20resource%20is,or%20otherwise%20influencing%20that%20resource) Provenance - FHIR v6.0.0-ballot3

<https://build.fhir.org/provenance.html>

[[15]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=The%20minimum%20necessary%20standard%2C%20a,the%20various%20circumstances%20of%20any) [[22]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=carry%20out%20a%20function,circumstances%20of%20any%20covered%20entity) [[70]](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html#:~:text=For%20uses%20of%20protected%20health,explicitly%20and%20include%20a%20justification) Minimum Necessary Requirement | HHS.gov

<https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/minimum-necessary-requirement/index.html>

[[17]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Requirements%20for%20High,must%20adhere%20to%20these%20requirements) [[55]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Under%20the%20EU%20AI%20Act%2C,10%29%2C%20a%20critical%20aspect) [[56]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Incident%20Reporting%20and%20Post,market%20monitoring) [[57]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=it%20triggers%20a%20cascade%20of,10%29%2C%20a%20critical%20aspect) [[58]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Interaction%20with%20Other%20Regulations%3A%20The,GDPR) [[73]](https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/#:~:text=Incident%20Reporting%20and%20Post,market%20monitoring) Navigating the EU AI Act: implications for regulated digital medical products - PMC

<https://pmc.ncbi.nlm.nih.gov/articles/PMC11379845/>

[[31]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=,patient%20identifiers%20for%20privacy%20accounting) [[32]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=references%20for%20privacy%20accounting%2C%20authentication,time%2C%20following%20these%20proven%20patterns) [[34]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=The%20FHIR%20,contains%20several%20key%20components) [[35]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=,patient%20identifiers%20for%20privacy%20accounting) [[38]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=database%20operations%2C%20and%20network%20events,time%2C%20following%20these%20proven%20patterns) [[41]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=%E2%80%8D) [[42]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=BALP%20provides%20standardized%20patterns%20for,time%2C%20following%20these%20proven%20patterns) [[47]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=references%20for%20privacy%20accounting%2C%20authentication,time%2C%20following%20these%20proven%20patterns) [[48]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=Every%20healthcare%20system%20must%20track,%E2%80%94%20both%20legally%20and%20operationally) [[80]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=,potentially%20affecting%20patient%20care%20applications) [[81]](https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale#:~:text=performance%20degradation.%20,data%20requires%20significant%20manual%20effort) FHIR Audit Logging at Healthcare Scale

<https://www.health-samurai.io/news/fhir-audit-logging-at-healthcare-scale>

[[49]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=Internationally%2C%20the%20medical%20device%20community,can%20be%20used%20to%20help) [[50]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=For%20this%20document%2C%20the%20term,by%20a%20manufacturer%2C%20that%20specifies) [[51]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=,to%20meet%20specified%20performance%20criteria) [[52]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=In%202021%2C%20the%20U,some%20cases%2C%20improve%20device%20performance) [[53]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=In%20this%20document%2C%20the%20FDA%2C,training%20risks%20are%20managed) [[54]](https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles#:~:text=,ensure%20device%20safety%20and%20effectiveness) Predetermined Change Control Plans for Machine Learning-Enabled Medical Devices: Guiding Principles | FDA

<https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles>

[[60]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=ISO%2FIEC%2042001%20is%20an%20international,or%20use%20of%20AI%20systems) [[61]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=ISO%2FIEC%2042001%20is%20an%20international,or%20use%20of%20AI%20systems) [[62]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=utilizing%20AI,AI%2C%20balancing%20innovation%20with%20governance) [[63]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=utilizing%20AI,AI%2C%20balancing%20innovation%20with%20governance) [[64]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=establishing%2C%20implementing%2C%20maintaining%2C%20and%20continually,or%20use%20of%20AI%20systems) [[65]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=Microsoft%20and%20ISO%2FIEC%2042001) [[68]](https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001#:~:text=system%20is%20a%20set%20of,or%20use%20of%20AI%20systems) ISO/IEC 42001:2023 Artificial Intelligence Management System Standards - Microsoft Compliance | Microsoft Learn

<https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001>

[[66]](https://www.a-lign.com/articles/understanding-iso-42001#:~:text=develop%2C%20and%20deploy%20AI%20systems,factors%20such%20as%20transparency) Understanding ISO 42001: The World's First AI Management System ...

<https://www.a-lign.com/articles/understanding-iso-42001>

[[67]](https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html#:~:text=ISO%2FIEC%2042001%20certification%20helps%20organizations%3A,as%20the%20EU%20AI%20Act) [[69]](https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html#:~:text=ISO%2FIEC%2042001%3A%20a%20new%20standard,as%20the%20EU%20AI%20Act) ISO/IEC 42001: a new standard for AI governance

<https://kpmg.com/ch/en/insights/artificial-intelligence/iso-iec-42001.html>

[[71]](https://docs.cancergenomicscloud.org/docs/tcia-data#:~:text=The%20Cancer%20Imaging%20Archive%20,in%20a%20standard%20DICOM%20format) [[72]](https://docs.cancergenomicscloud.org/docs/tcia-data#:~:text=clinical%20images%20matched%20to%20subjects,in%20a%20standard%20DICOM%20format) TCIA data

<https://docs.cancergenomicscloud.org/docs/tcia-data>

[[74]](https://pmc.ncbi.nlm.nih.gov/articles/PMC6790705/#:~:text=,as%20widely%20available%20as%20possible) The advantages of UK Biobank's open‐access strategy for health ...

<https://pmc.ncbi.nlm.nih.gov/articles/PMC6790705/>

[[75]](https://www.ukbiobank.ac.uk/about-us/how-we-work/access-to-uk-biobank-data/#:~:text=UK%20Biobank%20data%20are%20available,is%20in%20the%20public%20interest) [[76]](https://www.ukbiobank.ac.uk/about-us/how-we-work/access-to-uk-biobank-data/#:~:text=,collaborator%20on%20their%20research%20project) [[77]](https://www.ukbiobank.ac.uk/about-us/how-we-work/access-to-uk-biobank-data/#:~:text=UK%20Biobank%20switched%20to%20providing,when%20their%20project%20is%20completed) [[78]](https://www.ukbiobank.ac.uk/about-us/how-we-work/access-to-uk-biobank-data/#:~:text=,collaborator%20on%20their%20research%20project) Access to UK Biobank data - UK Biobank

<https://www.ukbiobank.ac.uk/about-us/how-we-work/access-to-uk-biobank-data/>