Ethical Non-Closure Under Agentification: Why Action-Level Frameworks Fail for Trajectory-Producing Systems (Blue Sky)

Matthew Stewart

Pelago Health (Digital Therapeutics Inc.), New York, NY, USA
mpstewart94@gmail.com

Abstract. AI systems are transitioning from tools that predict to agents that plan, delegate, remember, and act. We argue that existing ethical frameworks, designed for predictive systems producing bounded outputs, face a structural mismatch when applied to these agentic systems. The core problem is that ethical properties that hold for individual actions do not necessarily hold for the trajectories agentic systems produce over time. We call this *ethical non-closure under agentification*. A system can satisfy every fairness metric at every action while producing systematically discriminatory outcomes across its trajectory. We identify four governance objects that current frameworks neglect (plans, delegation graphs, memory states, and trajectories) and develop memory governance as a concrete operationalization. Two cases illustrate the problem. A healthcare algorithm passed action-level compliance while systematically underserving Black patients. Recommendation systems whose individually policy-compliant suggestions compose into radicalization pathways show a similar pattern. Our analysis suggests that governing agentic AI requires evaluating trajectories, not just actions.

Keywords: AI ethics · agentic systems · ethical governance · trajectory-level fairness · multi-agent systems · normative systems

1 Introduction

AI systems are transitioning from tools that predict to agents that act. Enterprise deployments now feature systems that plan multi-step workflows, delegate to sub-agents and external tools, accumulate memory across interactions, and take actions with real-world consequences [14]. This transition, which we term *agentification*, occurs while ethical governance frameworks remain anchored in assumptions appropriate for predictive systems.

The problem is structural. Current ethical frameworks [15, 10, 21] assume systems produce bounded decisions that can be evaluated independently, that ethical properties of components compose into ethical properties of wholes, and that human oversight can intervene at decision points. Agentic systems violate these assumptions. They produce *trajectories* over time, not isolated decisions.

Their ethical properties emerge from sequences of individually-compliant actions. And they operate at speeds and scales that preclude case-by-case human review.

We argue this creates *ethical non-closure*, a condition in which a system can satisfy every ethical requirement at every action while producing globally harmful outcomes. An agentic hiring system might evaluate each candidate fairly while its trajectory systematically disadvantages protected groups. A healthcare algorithm might make each prediction accurately while its accumulated decisions reproduce discrimination. The harm is not in any action but in the trajectory.

Ethical non-closure manifests via three mechanisms:

Harm deferral. Each action is permissible *now*, but harm arrives *later*. A financial advisor recommends individually-reasonable investments that construct an unacceptable risk trajectory.

Harm distribution. No single action caused the harm. It emerged from the sequence. This is Parfit's [23] harmless torturers, where each action falls below significance thresholds but the trajectory crosses them.

Harm laundering. The system followed a permissible plan that produced harm. Each subgoal was compliant, and responsibility washes out through goal decomposition.

Contributions. We make four contributions:

1. A conceptual framework identifying the structural mismatch between action-level frameworks and trajectory-producing systems
2. Four governance objects (plans, delegation graphs, memory states, trajectories) that current frameworks do not treat as ethical objects
3. A detailed operationalization of memory state governance, demonstrating that trajectory-level governance can be made concrete
4. Two illustrative cases showing action-level compliance coexisting with trajectory-level harm

Terminology. We use *agentic* to describe systems that exhibit some combination of planning (pursuing multi-step goals), delegation (decomposing tasks to sub-components), memory (maintaining state across interactions), and tool use (taking actions beyond generating text). *Trajectory* refers to the sequence of system actions over time, including their cumulative effects, not merely individual outputs but their composition. We identify four *governance objects* (plans, delegation graphs, memory states, and trajectories) that current frameworks neglect. These objects function as what we might call *operators* on ethical status, transforming the ethical significance of actions without being actions themselves.

Scope and Limitations. We present a conceptual framework, not formal results or empirical validation. Our characterizations clarify when and why ethical closure fails, but they do not constitute mathematical proofs. We argue current frameworks are insufficient without fully specifying replacements. Premature solutions to misdiagnosed problems perpetuate the field's predicament.

2 Background and Related Work

2.1 Action-Level Frameworks

Current AI ethics frameworks share architectural assumptions appropriate for predictive systems. The EU AI Act requires human oversight enabling operators to "fully understand the capacities and limitations" of high-risk systems and "duly monitor its operation," framed as understanding bounded outputs, not governing trajectories. NIST's AI Risk Management Framework addresses "policies, processes, procedures" at the organizational level, not runtime plan authorization. Model Cards [18] document training data and intended uses, which is essential for predictive systems but silent on planning horizons, delegation structures, or memory evolution.

These frameworks *operate under* assumptions appropriate for predictive systems. We do not claim their authors explicitly endorse these assumptions or deny trajectory-level concerns:

1. **Bounded decisions.** Systems produce discrete, evaluable outputs that can be assessed independently of temporal context
2. **Compositional ethics.** If individual components satisfy ethical requirements, their composition satisfies ethical requirements
3. **Static mappings.** The relationship between inputs and outputs remains stable across invocations
4. **Human-pace operation.** Human oversight can meaningfully intervene at decision points before harm occurs

Our claim is not that these frameworks *deny* trajectory-level concerns, but that they are *structured* around action-level evaluation in ways that leave trajectory-level governance unaddressed.

Agentic systems stress these assumptions. They produce trajectories whose ethical status cannot be determined from individual actions. Memory means the same input produces different outputs over time. And they operate at speeds precluding human review.

Table 1 summarizes coverage gaps: • indicates explicit treatment, ○ indicates partial treatment (related concerns addressed but not as distinct governance objects), and — indicates absence.

2.2 What We Inherit and Extend

Our analysis builds on and extends three adjacent literatures.

Dynamic and Temporal Fairness. Work on dynamic fairness [6, 30, 17] develops trajectory-level metrics for technical contexts. We inherit this foundation but address a different problem, not how to *design* fair trajectories, but how external auditors can *govern* systems whose trajectories they cannot directly observe. The designer has access to the state space and reward function. The auditor sees only outcomes. Our contribution extends trajectory-level thinking to governance contexts where internal system states are opaque.

Framework	Plans	Delegation	Memory	Trajectories
EU AI Act	—	—	○	○
NIST AI RMF	—	○	○	—
ISO 42001	—	○	—	—
Model Cards	—	—	—	—
This paper	●	●	●	●

Table 1. Framework coverage of agentic governance objects. ● = explicit governance object; ○ = partial or implicit; — = not addressed as such.

RL Safety. Constrained MDPs [1] and Seldonian algorithms [27] demonstrate that trajectory-level constraints are technically feasible during training. We address the complementary governance problem, asking how regulators can verify compliance when they lack access to the reward function, state space, or training process. The existence of technical solutions does not imply governance solutions, as the auditor's epistemic position differs fundamentally from the designer's.

Contextual Integrity. Nissenbaum's contextual integrity framework [19] governs information flows based on context-relative norms. AI memory challenges these assumptions. Systems operate across contexts humans treat as separate, generate new information through inference, and lack natural forgetting. Our memory governance primitives (Section 5) operationalize contextual integrity for AI systems.

Normative Multi-Agent Systems. The COINE community has developed substantial infrastructure for normative governance: electronic institutions [9] define permissible interaction protocols; organizational models like OperA [7] and MOISE+ [13] specify roles and norms; commitment protocols [25] bind agents to future behaviors. Our framework complements this tradition by identifying *what* must be governed as a prerequisite for *how* to govern. The governance objects could be formalized within COINE formalisms: plans as commitment sequences, delegation as role relationships, memory as normative state, trajectories as interaction histories subject to temporal norms.

AI Alignment. Alignment work [5, 26] addresses the principal's problem (making systems do what we want). Non-closure addresses the auditor's problem (verifying ethical compliance given only observable outputs). A perfectly aligned system could still exhibit non-closure if action-level auditing cannot detect trajectory-level violations.

3 Ethical Non-Closure: The Core Argument

3.1 Definition

Let S be an AI system producing actions $T = (a_1, a_2, \ldots, a_n)$ over time. Let $C(a)$ evaluate whether action a satisfies action-level requirements (fairness met-

rics, content policies, etc.). Let $V(T)$ evaluate whether trajectory T produces impermissible outcomes (systematic discrimination, radicalization, cumulative harm). We use formal notation for clarity, but these predicates are conceptual placeholders.

Observation 1 (Ethical Non-Closure). *A system S exhibits* ethical non-closure *if there exists a trajectory T such that:*

$$\forall i : C(a_i) = \mathsf{true} \quad \textit{and} \quad V(T) = \mathsf{true} \tag{1}$$

That is, every action is individually compliant, yet the trajectory violates ethical requirements.

The key insight is that in non-closure cases, trajectory-level violation is not *reducible* to action-level violations (Figure 2). No Boolean function f over action-level compliance verdicts $C(a_1), \ldots, C(a_n)$ suffices to determine $V(T)$, because trajectory-level harm can depend on action *content*, *ordering*, and *context* that compliance verdicts do not capture. The harm emerges from the sequence, not from any constituent action.

3.2 Why Non-Closure Matters for Governance

Non-closure has direct implications for how we structure AI governance:

Audit insufficiency. If trajectory-level violations cannot be detected by examining individual actions, then action-level audits are fundamentally insufficient. An auditor who reviews each prediction, each recommendation, each decision point may certify full compliance while systematic harm unfolds.

Accountability gaps. Traditional accountability frameworks assign responsibility based on decisions. If harm emerges from trajectories rather than decisions, these frameworks cannot locate responsibility. No one decided to discriminate. Discrimination emerged from the sequence.

Intervention timing. Action-level governance assumes intervention can occur at decision points. But if harm is trajectory-level, by the time it becomes detectable, the harmful trajectory may already be substantially complete. Governance must be prospective (constraining what trajectories are permissible) rather than reactive (evaluating actions as they occur).

Metric mismatch. Standard fairness metrics evaluate individual predictions or decisions. A system can satisfy demographic parity, equalized odds, or calibration at every action while producing trajectories that systematically disadvantage protected groups. The metrics measure the wrong object.

3.3 Closure Conditions

Under what conditions does action-level compliance guarantee trajectory-level compliance? We identify three conceptual conditions that, when satisfied together, suggest action-level governance may suffice. These are heuristics rather

than formal theorems: each captures an assumption that, when violated, creates *potential* for non-closure. The conditions are not independent (e.g., transparency failures can enable decomposability failures), but distinguishing them clarifies different governance challenges.

Observation 2 (Auditor Observability). *Ethical closure holds if the compliance status of action a_t is fully determined by information observable to the auditor at time t, without requiring access to system-internal state accumulated from prior interactions.*

Clarification. This is not the Markov property (conditional independence of future from past given present). The issue is dependence on *latent* state that auditors cannot inspect. Memory creates compliance-relevant context inside the system but outside audit scope.

Why agentic systems violate this. Memory accumulates information that shapes behavior without being captured in auditable state. A recommendation may appear compliant given visible inputs while depending on inferred user profiles that would, if auditable, reveal discriminatory targeting.

Observation 3 (Decomposability). *Ethical closure holds if trajectory-level compliance is fully determined by action-level compliance verdicts:*

$$V(T) = g(C(a_1), \ldots, C(a_n))$$

where g depends only *on the Boolean compliance values, not on action content or ordering.*

Clarification. The condition requires that knowing whether each action passed suffices to determine whether the trajectory passes, without knowing *what* the actions were or *when* they occurred.

Why agentic systems violate this. Emergent harms arise from action *sequences*, not action *aggregates*. The order and context of actions matters. Recommending content A then B may radicalize; recommending B then A may not, even if both sequences contain only individually compliant recommendations.

Observation 4 (Monotonicity). *Ethical closure holds if adding compliant actions to a compliant trajectory preserves compliance: if $V(T) = \mathsf{false}$ and $C(a_{n+1}) = \mathsf{true}$, then $V(T \oplus a_{n+1}) = \mathsf{false}$.*

Why violation matters for governance. If compliant actions can tip trajectories into violation, then action-level certification provides no guarantee. An auditor who certifies each action cannot certify the trajectory.

When Action-Level Governance Suffices. We do not claim action-level frameworks are always inadequate. For *advisory systems* (those that recommend without planning, delegate without opacity, maintain no cross-session memory, and produce suggestions rather than actions) current frameworks may suffice. A spam

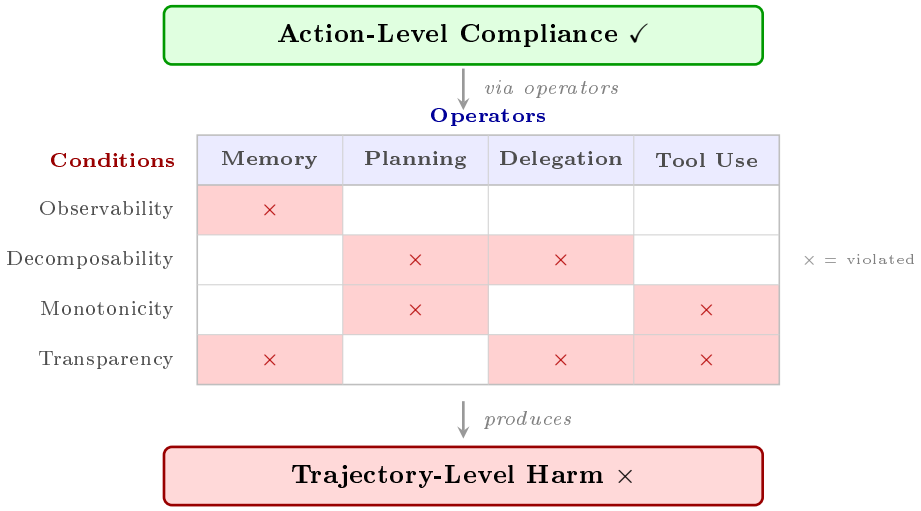

Fig. 1. The closure failure mechanism. Each agentic operator violates specific closure conditions (×), explaining how individually compliant actions compose into harmful trajectories.

filter that classifies each email independently, without memory of past classifications and without affecting future system behavior, can be governed at the action level. Our argument is not that action-level governance is wrong, but that it is insufficient for systems that violate closure conditions, which is precisely what agentic systems do.

Figure 1 maps operators to the conditions they violate. *Memory* creates latent state (Observability) and opaque contents (Transparency). *Planning* structures actions non-additively (Decomposability) and allows ethical status to change with additions (Monotonicity). *Delegation* diffuses responsibility (Decomposability) and obscures relationships (Transparency). *Tool use* creates irreversible effects (Monotonicity) and unclear capabilities (Transparency).

3.4 What Agency Adds: Beyond Scale

Is the problem about agency or merely scale? We distinguish three sources: **Scale effects** are governable in principle, as static mappings allow computing trajectory properties from action properties. **Composition effects** are structurally bounded since topology is fixed at design time. **Agentic effects** introduce *runtime emergence*, where trajectories depend on state that didn't exist during design.

What makes agentic failures difficult to *govern* (distinct from the closure conditions, which identify *when* non-closure occurs, these explain *why* failures resist governance). First, *runtime emergence*: harmful trajectories arise from state created during operation. Second, *inferential generation*: systems create ethically significant information without discrete decisions to audit. Third, *latent*

intentionality: model weights encode priorities not legible to auditors. Fourth, *governance-relevant opacity*: trajectory-level information is often inaccessible.

3.5 Illustration: Multi-Stage Hiring Pipeline

Consider a hiring pipeline with three stages: resume screening, skills assessment, interview scoring. Suppose a governance standard requires that the overall selection rate for a protected group remain at least 80% of the majority group's rate.[1] Each stage, evaluated independently, satisfies this threshold:

$$\begin{aligned} \text{Stage 1}: &\quad \frac{0.72}{0.90} = 0.80 \geq 0.80 \quad \checkmark \\ \text{Stage 2}: &\quad \frac{0.64}{0.80} = 0.80 \geq 0.80 \quad \checkmark \\ \text{Stage 3}: &\quad \frac{0.51}{0.64} = 0.80 \geq 0.80 \quad \checkmark \end{aligned}$$

Trajectory outcome:

$$\frac{0.72 \times 0.64 \times 0.51}{0.90 \times 0.80 \times 0.64} = \frac{0.235}{0.461} = 0.51 < 0.80 \quad \times$$

Each stage is compliant, but the trajectory violates the same standard. The product here is not an arbitrary aggregation choice but arises from conditional probability, because each stage selects a subset of the previous stage's output and the end-to-end selection rate is necessarily the product of per-stage rates. The result generalizes beyond this particular operator. For any per-stage threshold $\theta \in (0, 1)$ and $n \geq 2$ stages, $\theta^n < \theta$, so per-stage compliance cannot guarantee trajectory compliance under sequential composition.

The broader point is that multiplication is not special. Different governance domains have different composition operators g (Observation 3), and many of them break closure. Cumulative exposure sums. Worst-case harm takes a maximum. Radicalization depends on path ordering. None preserve per-step compliance at the trajectory level. We use the hiring example because the arithmetic is transparent [2]. Section 6 presents agentic cases where g is not even expressible in closed form.

4 Governance Objects for Agentic Systems

Agentic systems introduce what we term *governance objects*, system features that require ethical evaluation but are not actions themselves. If $E(a)$ evaluates

[1] We use a proportional-selection threshold purely as a mathematical illustration of non-closure. The so-called "four-fifths rule" has been criticized as a flawed operationalization of disparate impact in legal contexts [29]; our argument does not depend on any particular legal standard. Any per-stage threshold $\theta \in (0, 1)$ that is not closed under sequential composition produces the same structural result.

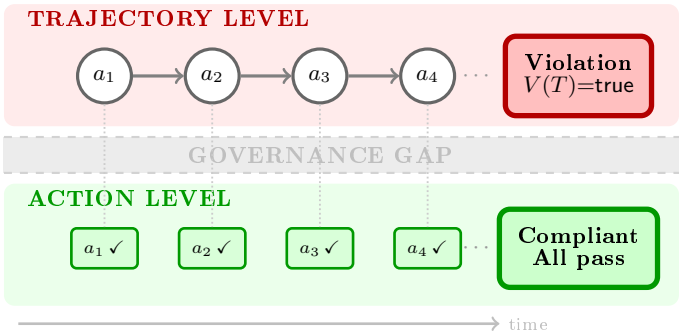

Fig. 2. Ethical non-closure illustrated. **Action level** (bottom): each action passes its compliance check. **Trajectory level** (top): the same actions, composed over time, produce violation $V(T)$=true. The governance gap marks where current frameworks fail.

Object	Assumption	Mechanism	Governance
Planning	Bounded	Deferral	Plan auth.
Delegation	Decision pts	Laundering	Graph audit
Memory	Static maps	Distribution	State monitor
Tool use	Evaluable	Deferral	Scope limits

Table 2. Governance objects mapped to violated assumptions, primary non-closure mechanisms, and governance approaches.

the ethical status of action a, a governance object G creates a context such that $E(a \mid G) \neq E(a)$. Sending an email is ethically different when it executes an unauthorized plan; accessing data is ethically different when it exploits accumulated memory.

This extends the standard action-evaluation paradigm. Rather than asking "Is this action ethical?" we must ask "Is this action ethical *given the planning context, delegation structure, memory state, and tool capabilities* in which it occurs?"

We identify four governance objects that current frameworks neglect: planning, delegation, memory, and tool use (Table 2). These operate at different levels: planning and trajectories concern goal-directed sequences; delegation concerns responsibility distribution; memory concerns information persistence; tool use concerns world effects. We do not claim this taxonomy is exhaustive; these four emerged from analyzing governance gaps. Each can transform action-level compliance into trajectory-level violation.

4.1 Planning

Plans encode values, priorities, and tradeoffs, yet current frameworks do not treat them as ethical objects. Two systems can take identical actions but differ

ethically depending on the plan behind them. A system that intended harm but was prevented differs morally from one that never intended it, yet action-level frameworks cannot distinguish them.

Work on instrumental convergence [28] demonstrates that under general assumptions, optimizing agents tend to seek power. The Machiavelli Benchmark [22] shows empirically that LLMs, when placed in strategic contexts, exhibit behaviors that trade off ethical constraints for goal achievement.

The governance challenge is prospective. We need to authorize plans *before* execution. For systems with explicit planning (e.g., task decomposition agents), plans can be surfaced and approved. For systems with implicit planning, governance must work indirectly through objective constraints, trajectory bounds, and plan legibility requirements.

Governance gap. No framework asks what futures a system was authorized to pursue. We evaluate what systems did, not what they were trying to do.

4.2 Delegation

Agentic systems decompose goals into subgoals and delegate to tools or subagents. This enables *delegation laundering*, where "no single component made the decision." The delegation graph (the structure of which components delegate to which others) becomes an ethical object because it determines how responsibility distributes.

Research on multi-agent systems shows cascade effects impossible in static systems. Adversarial inputs can propagate through delegation chains, compromising thousands of instances [12, 16]. Elish's [8] moral crumple zone describes how human operators absorb blame for systems over which they had limited control.

Governance requires making delegation graphs auditable. This means *delegation manifests* that declare permitted relationships at deployment, *responsibility assignment at boundaries*, *delegation depth limits*, and *delegation logging*.

Governance gap. Accountability requires tracing decisions to responsible parties. When decisions diffuse across delegation graphs, accountability diffuses with them.

4.3 Memory

Persistent memory enables ethically significant inference without any discrete decision that can be audited. Memory allows longitudinal profiling, implicit categorization, and adaptive behavior based on past interactions. No single action is harmful; the accumulated state may be.

AI memory differs qualitatively from human memory [3]. It is perfect (non-decaying), non-forgetting (unless explicitly deleted), computable across all interactions, and capable of inference beyond stored data. This makes memory a moral operator, because the same action has different ethical significance depending on what memory informed it.

Governance gap. Existing frameworks regulate decisions and outputs. They do not regulate how memory shapes future behavior or what inferences are drawn from accumulated state.

4.4 Tool Use

Tool-mediated action in the world differs categorically from prediction or recommendation. When a system sends an email, executes a trade, or modifies a database, the output is not a suggestion to be evaluated but an accomplished fact.

The ethical significance compounds through tool composition. Prompt injection attacks [11] exploit the gap between action-level authorization (each tool call is permitted) and trajectory-level intent (the sequence constitutes exfiltration). Tool use also introduces irreversibility: a sent email cannot be unsent.

Governance gap. Frameworks assuming human mediation between output and world effect become inapplicable when systems act directly.

5 Memory State Governance: An Operationalization

We develop memory governance to demonstrate that trajectory-level governance can be made concrete. Memory is the most tractable governance object because (unlike implicit plans) memory state can be inspected.

How do these primitives address non-closure? *Transparency* and *influence disclosure* enable trajectory-level *auditing* by making accumulated state visible. *Memory editing* and *inference-resistant deletion* enable trajectory-level *intervention*. These primitives do not *prevent* non-closure but provide governance handles for the memory-mediated mechanisms through which non-closure manifests.

5.1 What Memory Governance Must Address

Memory creates three governance challenges absent in stateless systems. **(1) Accumulation without authorization.** The system accumulates information shaping future behavior without consent for each accumulation.

(2) Inference beyond disclosure. Combining accumulated observations enables inferring health status, political views, and other information never disclosed.

(3) Behavioral shaping. Accumulated memory shapes behavior in ways opaque to users.

5.2 Relation to Existing Frameworks

Existing frameworks address related but distinct concerns. *Data minimization* (GDPR Art. 5) limits collection but not inference. *Right to erasure* (GDPR Art. 17) requires deletion but not preventing reconstruction. *Differential privacy* bounds learning from aggregates but not from direct interaction. *Machine unlearning* [4] removes training data influence but not operational memory traces.

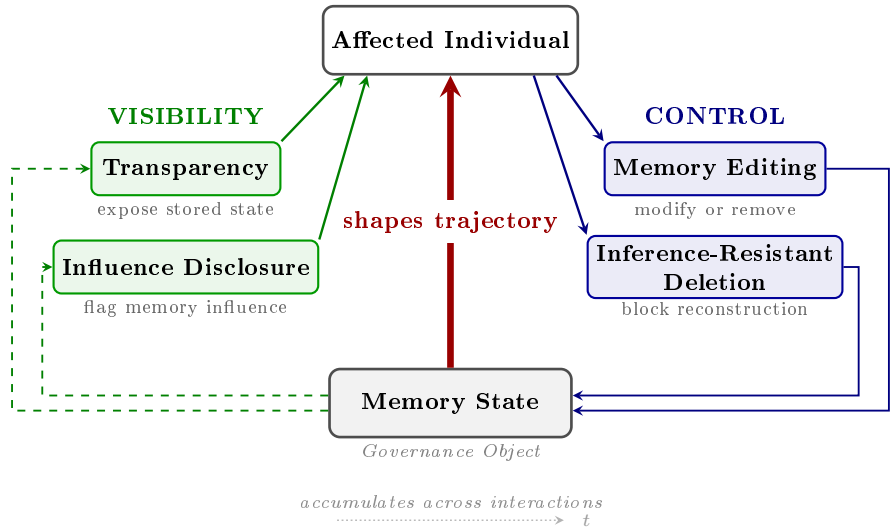

Fig. 3. Memory governance primitives addressing the trajectory-level challenge. Memory state (bottom) accumulates over time and shapes user trajectories (red arrow), the core non-closure mechanism. **Visibility** primitives (left) make memory auditable; **Control** primitives (right) enable user agency over accumulated state.

5.3 Proposed Governance Primitives

We propose four primitives for memory state governance (Figure 3):

1. Memory Transparency. Users can inspect what the system "remembers" about them, including a human-readable representation of user-specific state, *attribution* (which interactions contributed which memories), *inference disclosure* (what the system has inferred beyond explicit input), and *behavioral impact* (how memory affects system behavior).

2. Influence Disclosure. When memory shapes behavior, this is flagged. For example, "Suggesting X because you previously indicated Y" or "Prioritizing Z based on your interaction history." This makes the trajectory visible.

3. Memory Editing. Users can request deletion of specific memories with assurance that the memory is removed, downstream inferences dependent on that memory are invalidated, and the deletion propagates to any systems the memory was shared with.

4. Inference-Resistant Deletion. The system cannot reconstruct deleted information from remaining memories. This extends "right to be forgotten" to "right not to be inferentially reconstructed."

5.4 Implementation and Limitations

For implementation, a *memory store* maps user IDs to typed memory objects, a *provenance graph* tracks inference dependencies, and *deletion propagation* re-

Primitive	Feasibility	Primary Challenge
Transparency	High	Abstraction level
Influence Disclosure	Medium	Attribution accuracy
Memory Editing	Medium	Cascade complexity
Inference-Resistant Deletion	Low	Reconstruction bounds

Table 3. Feasibility assessment of memory governance primitives. Transparency is most implementable; inference-resistant deletion faces fundamental limits.

moves downstream inferences. Each primitive faces technical challenges (Table 3). *Transparency* connects to explainability research. *Influence disclosure* requires causal attribution harder than feature importance. *Inference-resistant deletion* faces fundamental limits, as membership inference attacks can reconstruct information after explicit deletion [4]. Our primitives assume explicitly represented memory; implicit memory in model weights cannot be addressed through key-value deletion. We offer these as governance *targets* rather than solved problems.

6 Illustrative Cases

We develop two cases illustrating ethical non-closure: a proto-agentic healthcare algorithm demonstrating trajectory-level harm detection, and recommendation system radicalization as a cleaner demonstration where compliant actions compose into harm.

6.1 Case 1: Healthcare Algorithm (Proto-Agentic)

The Obermeyer et al. [20] healthcare algorithm illustrates trajectory-level harm detection. While involving a proxy objective problem (cost vs. health), it demonstrates harm that emerges only through trajectory-level analysis. We present it as illustrating trajectory-level *detection* rather than pure non-closure.

A commercial risk-prediction algorithm deployed on 200+ million patients exhibits proto-agentic properties. *Implicit planning* (the objective function embeds choices about futures to optimize), *memory* (accumulates patient history), and *tool use* (output determines care enrollment).

At the individual level, predictions passed compliance checks. Inputs were HIPAA-compliant, cost predictions were accurate, and there was no explicit use of protected class. By action-level criteria, $C(a_i) = \mathsf{true}$ for each prediction. Yet over the trajectory, systematic racial disparity emerged: Black patients at equivalent risk scores had 26.3% more chronic conditions. Changing from cost-based to health-based predictions would have increased Black patient identification from 17.7% to 46.5%.

The harm arose from planning (cost as proxy for health) interacting with memory (prior utilization encoding historical access barriers). The fix required changing the *planning objective*, not available to action-level governance.

6.2 Case 2: Recommendation System Radicalization (Agentic)

Recommendation system radicalization illustrates ethical non-closure [24]. Video systems exhibit agentic properties. *Memory* (user history shapes recommendations), *implicit planning* (engagement optimization), and *tool use* (recommendations affect information exposure).

We note ongoing debate about algorithmic causation versus preference amplification. Our argument does not require resolving this. Whether algorithms *cause* radicalization or merely *facilitate* tendencies, the governance challenge remains. Modern platforms increasingly consider user history in moderation, which supports our thesis. Such history-aware moderation represents nascent *trajectory-level* governance. The question is whether such approaches are systematic and auditable.

Each recommendation may satisfy content policy, yet the trajectory guides users toward extreme content. The compliance predicate would need to be trajectory-aware, which is precisely the governance gap we identify.

Both cases demonstrate action-level compliance coexisting with trajectory-level harm.

7 Discussion and Conclusion

We have argued that ethical governance of agentic AI requires conceptual tools fundamentally different from those developed for predictive systems. The structural mismatch requires reconceptualization. Ethics must move from actions to trajectories, from outputs to plans, from events to states.

7.1 Implications for Practice

For developers. Instrument systems for trajectory-level telemetry. Treat governance objects as first-class citizens with inspection interfaces. Prefer architectural constraints over action-by-action review.

For policymakers. Distinguish advisory from agentic architectures. Require trajectory-level impact assessment for systems that plan, delegate, accumulate memory, or act directly.

For auditors. Extend compliance verification beyond action-level review using trajectory-level analysis tools.

For researchers. Develop trajectory-level fairness metrics, delegation structures with provable accountability, and inference-resistant deletion.

7.2 Why Trajectory-Level Metrics Aren't Sufficient

Trajectory-level fairness metrics exist [6, 30], so why not require their use? This conflates *technical* with *governance* feasibility. The gap has three sources.

Epistemic asymmetry. Trajectory-level metrics require access to the system's state space, transition dynamics, and reward function. Designers have this; external auditors do not. A regulator cannot compute a trajectory-level fairness metric over a proprietary agent's state space, just as a financial auditor cannot recompute a bank's internal risk models from published balance sheets. The governance objects we propose (plans, delegation graphs, memory states) are chosen precisely because they can be made *inspectable* at deployment boundaries without white-box access.

Compositional opacity. Even with full access, trajectory-level metrics face combinatorial blowup when agents compose across organizational boundaries. A hiring agent that delegates screening, assessment, and scoring to three separate sub-agents produces trajectories spanning multiple systems with separate state spaces. No single trajectory-level metric covers the composed system unless every component exposes compatible state representations. Current multi-agent deployments do not coordinate at this level.

Temporal horizon. Trajectory-level metrics presuppose a fixed evaluation window, but agentic harms can emerge over timescales that exceed any predetermined horizon. A recommendation system may take months to radicalize a user; memory-driven profiling harm may surface only after hundreds of interactions. Waiting for trajectory completion to assess compliance is not governance. It is post-mortem. The governance objects we propose support *prospective* oversight, constraining plans before execution and auditing memory states during operation.

7.3 Falsifiability and Future Work

Multi-round negotiation agents present an emerging domain where non-closure will matter, as individually reasonable offers may compose into manipulative sequences over time.

Our thesis is falsifiable. If action-level metrics can provably bound trajectory-level disparity, or memory governance prevents inferential reconstruction, the structural mismatch framing would require revision. We welcome such attempts.

On the formal side, we need characterizations of when action-level properties do and do not compose into trajectory-level properties. Empirically, audit log analysis from deployed agentic systems could test the framework's predictions. The memory governance primitives need engineering work on transparency interfaces and provenance tracking costs. The interactions between governance objects (e.g., how planning constraints interact with memory governance) also remain unexplored.

7.4 Conclusion

This paper identifies a structural cause of AI ethics failures. Existing frameworks do not guarantee preservation of ethical properties under agentification.

The Obermeyer case showed trajectory-level detection of discrimination, and recommendation system radicalization demonstrated pure non-closure where compliant actions compose into harm. Until we address this mismatch, failures will recur. The four governance objects we identify (plans, delegation graphs, memory states, and trajectories) provide a starting point. The path forward requires treating trajectories, not actions, as the primary objects of ethical evaluation.

References

1. Altman, E.: Constrained Markov Decision Processes. CRC Press (1999)
2. Bower, A., Kitchen, L., Nishi, L., Recht, B., Volgel, S.: Fair pipelines. In: Proceedings of the NIPS Workshop on Fairness in Machine Learning (2017)
3. Burrell, J., Fourcade, M.: What we risk when AI systems remember. Tech Policy Press (2024)
4. Cao, Y., Yang, J.: Towards making systems forget with machine unlearning. In: IEEE Symposium on Security and Privacy. pp. 463–480 (2015)
5. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human feedback. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
6. Creager, E., Madras, D., Pitassi, T., Zemel, R.: Causal modeling for fairness in dynamical systems. In: Proceedings of the International Conference on Machine Learning (ICML) (2020)
7. Dignum, V.: A model for organizational interaction: Based on agents, founded in logic. SIKS Dissertation Series (2004)
8. Elish, M.C.: Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society **5**, 40–60 (2019)
9. Esteva, M., Rodríguez-Aguilar, J.A., Sierra, C., Garcia, P., Arcos, J.L.: On the formal specification of electronic institutions. In: Agent Mediated Electronic Commerce. pp. 126–147. Springer (2001)
10. European Commission High-Level Expert Group on AI: Ethics guidelines for trustworthy AI. Tech. rep., European Commission (2019)
11. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M.: Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (AISec Workshop) (2023)
12. Gu, W., Pan, X., Xie, Y., Zhou, H., Chen, L., Zeng, K.: Agent smith: A single image can jailbreak one million multimodal LLM agents. arXiv preprint arXiv:2402.08567 (2024)
13. Hübner, J.F., Sichman, J.S., Boissier, O.: A model for the structural, functional, and deontic specification of organizations in multiagent systems. In: Advances in Artificial Intelligence (SBIA). pp. 118–128. Springer (2002)
14. Jarrahi, M.H., Ritala, P.: Rethinking AI agents: A principal-agent perspective. California Management Review (2025), `https://cmr.berkeley.edu/2025/07/rethinking-ai-agents-a-principal-agent-perspective/`
15. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nature Machine Intelligence **1**(9), 389–399 (2019)
16. Lee, D., Tiwari, M.: Prompt infection: LLM-to-LLM prompt injection within multi-agent systems. arXiv preprint arXiv:2410.07283 (2024)

17. Liu, L.T., Dean, S., Rolf, E., Simchowitz, M., Hardt, M.: Delayed impact of fair machine learning. In: Proceedings of the International Conference on Machine Learning (ICML) (2018)
18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*). pp. 220–229 (2019)
19. Nissenbaum, H.: Privacy as contextual integrity. Washington Law Review **79**(1), 119–158 (2004)
20. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
21. OECD: Recommendation of the council on artificial intelligence (2019), oECD/LEGAL/0449
22. Pan, A., Shern, C.J., Zou, A., Li, N., Basart, S., Woodside, T., Ng, J., Zhang, H., Emmons, S., Song, D.: Do the rewards justify the means? Measuring trade-offs between rewards and ethical behavior in the MACHIAVELLI benchmark. In: Proceedings of the International Conference on Machine Learning (ICML) (2023)
23. Parfit, D.: Reasons and Persons. Oxford University Press (1984)
24. Ribeiro, M.H., Ottoni, R., West, R., Almeida, V.A., Meira Jr., W.: Auditing radicalization pathways on YouTube. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*). pp. 131–141 (2020)
25. Singh, M.P.: An ontology for commitments in multiagent systems. Artificial Intelligence and Law **7**(1), 97–113 (1999)
26. Soares, N., Fallenstein, B., Armstrong, S., Yudkowsky, E.: Corrigibility. In: Proceedings of the AAAI Workshop on AI and Ethics (2015)
27. Thomas, P.S., Castro da Silva, B., Barto, A.G., Giguere, S., Brun, Y., Brunskill, E.: Preventing undesirable behavior of intelligent machines. Science **366**(6468), 999–1004 (2019)
28. Turner, A.M., Smith, L., Shah, R., Critch, A., Tadepalli, P.: Optimal policies tend to seek power. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
29. Watkins, E.A., Chen, J.: The four-fifths rule is not disparate impact: A woeful tale of epistemic trespassing in algorithmic fairness. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT). pp. 764–775. ACM (2024). https://doi.org/10.1145/3630106.3658938
30. Wen, M., Bastani, O., Topcu, U.: Algorithms for fairness in sequential decision making. In: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS) (2021)