After the Fork: When Fidelity Claims Become Counterfactual in Evolving Digital Replicas

George Kujo
Independent Researcher, Tokyo, Japan
ORCID: 0009-0006-6864-0145
E-mail: Research@georgekujo.com

Preprint / Author Manuscript — Version 1.1, September 2026

This public version is deposited on Zenodo.

Abstract

Person-specific digital replicas can increasingly be designed to persist, learn, and change after their creation. This creates a problem that is obscured when later outputs are evaluated under a single notion of fidelity to a human source. I distinguish three claims: historical fidelity to an empirically anchored source state, contemporary resemblance to the actual source, and counterfactual developmental attribution—the claim that a replica has developed as that particular person would have developed under developmentally corresponding conditions. Once replica and source acquire materially different post-origin histories, neither divergence from nor resemblance to the contemporary source can by itself validate developmental fidelity. Divergence may reflect model failure or legitimate experiential divergence, while resemblance may result from anchoring, constrained adaptation, convergence, or coincidence. The relevant comparator for the strongest developmental claim is therefore not the actual source but an unrealized counterfactual trajectory of that person. I call this change in evidential target the Fidelity Regime Shift. The argument does not depend on digital personhood, consciousness, or numerical identity. Its practical consequence is attributional: later outputs should be distinguished as retrieval, reconstruction, prediction, counterfactual developmental simulation, or autonomously evolved judgment. Persistent personal AI therefore requires provenance that keeps its relation to the source visible after the developmental fork. A replica may become harder to validate as a representation of a person precisely because it has become better able to develop like a person.

Keywords: digital replicas; fidelity; counterfactual reasoning; personal AI; attribution; personal identity

1. Introduction

Digital replicas of particular people are usually evaluated as representations. Does the system speak as the person speaks? Does it reproduce that person's beliefs, memories, preferences, characteristic reasoning, or responses to unfamiliar questions? Person-specific language models already demonstrate that systems trained on a sufficiently distinctive corpus can generate novel answers that are difficult to distinguish from those of the person on whom they were based (Schwitzgebel et al., 2024).

Such comparisons are imperfect, but they can be empirically anchored in recorded statements, behavior, decisions, writings, and other evidence about the source.

The problem becomes harder when the replica does not remain static.

Consider a person-specific artificial agent, D, created in 2026 from an extensive record of a living person, H. At creation, D reproduces H's documented views and characteristic patterns of judgment reasonably well. D then

persists. It interacts with other people, receives new information, accumulates memories, revises beliefs, and learns from later decisions. H meanwhile continues living an ordinary human life.

By 2036, H and D have undergone substantially different histories.

Suppose they are asked the same difficult question and give different answers. One explanation is that D has ceased to represent H accurately. Its model may have drifted or its updating procedure may be defective. But another explanation is possible: D may have changed precisely because it was designed to learn from experience. Had H undergone a developmentally corresponding history, H might also have changed. Observed disagreement therefore does not by itself establish failure.

Now reverse the result. Suppose H and D still give the same answer in 2036. That resemblance is empirically assessable, but its cause is not determined by the endpoint alone. D may genuinely have developed in a way characteristic of H. Alternatively, it may have remained strongly anchored to its 2026 state, adapted only weakly, converged on the same answer by another pathway, or reached it coincidentally. Observed resemblance therefore does not by itself establish successful development either.

This produces a tension between preservation and development. A preservation-oriented replica can remain easy to compare with an observed past person precisely by changing little. A development-oriented replica can become more person-like in one important respect—its capacity to acquire a history of its own—while becoming harder to validate against the source from which it began.

Several components are already recognized. Branching and autonomous digital counterparts are established in the literature (Deguchi et al., 2022), and recent work on digital duplicates examines autonomy, scarcity, authenticity, relationships, permissibility, and misattribution (Danaher & Nyholm, 2024, 2025; Karpus & Strasser, 2025; Sweeney, 2024; Voinea et al., 2024). Person-specific systems can also predict unobserved decisions of represented individuals (Earp et al., 2024; Sweeney, 2023). Other recent work asks which temporal version of a changing person a personalized model represents (Annoni et al., 2026), analyzes identity drift in evolving digital-afterlife systems (Spitale & Germani, 2026), and benchmarks human-like personality change in LLM personas (Wang et al., 2026). That debate concerns the value and permissibility of duplication; the present question concerns the evidential structure of fidelity claims made about a duplicate after it develops.

The present contribution is narrower. It asks what happens to a continuing claim of **fidelity** when a person-specific replica itself acquires a materially independent developmental history that its source never underwent.

I distinguish three claims that can otherwise be compressed into the statement that a replica remains “faithful” to a person.

Historical fidelity asks whether D represents H as H was at an earlier time.

Contemporary resemblance asks whether D now resembles the actual H.

Counterfactual developmental attribution asks whether D has developed as this particular H would have developed under developmentally corresponding conditions.

The first two can be empirically anchored in states of the actual source. The third refers to an unrealized developmental trajectory of that person.

I call this change in evidential target the **Fidelity Regime Shift**. “Regime shift” here does not denote a metaphysical transition or a precisely measurable threshold. It identifies a change in what evidence is required by different continuing fidelity claims once post-origin history becomes relevant to the properties under evaluation. Nor does every post-fork fidelity claim become counterfactual. The counterfactual comparator appears when the stronger claim is made that independent change itself remains faithful to how the particular source person would have changed.

The central issue is not whether a digital branch is metaphysically "really" the source person. It is what we are entitled to claim about the relation between them after the fork.

2. Digital Replicas, Branching Selves, and Evolving Personas

2.1 Digital replicas and psychological continuity

Karpus and Strasser (2025), drawing on Parfit's account of psychological connectedness, argue that relations between persons and their digital replicas need not be understood in all-or-nothing terms. Relevant connections may include memories, beliefs, desires, dispositions, outlooks, and other psychological features, and these connections can exist to varying degrees (Parfit, 1984).

This approach is useful because representation can be discussed without first resolving numerical identity. A digital system might inherit a substantial part of a particular person's psychologically relevant profile without thereby being assumed to be numerically identical with that person.

Karpus and Strasser also identify consequences for authorship, responsibility, and attribution, including overestimating the psychological relation between source and replica. The present paper therefore does not claim to discover an attribution problem. Its narrower concern is how the evidential basis of the source-replica relation changes once later states are produced through a history not undergone by the source.

At creation, a replica may inherit memories, commitments, preferences, and characteristic reasoning from H. Later states may instead be products of path-dependent development.

2.2 Branching digital selves

Deguchi et al. (2022) already model digital counterparts that share a person's history up to a branch and then live their own lives as autonomous agents. The present paper presupposes that structure.

Let H and D share a relevant history up to t_0. After t_0, H and D follow different trajectories. Branching explains how that can occur; it does not determine how later evidence about actual H should be used to evaluate D. If D differs from H ten years later, the difference may reflect replication failure or different histories.

The present issue is therefore not branching itself but how continuing claims about the source-replica relationship should be evaluated after it.

2.3 From a temporal snapshot to a developmental agent

Person-specific language models already make part of the problem concrete. The DigiDan experiment demonstrated that a model trained on Daniel Dennett's work could generate novel philosophical responses that experts sometimes struggled to distinguish from Dennett's own responses (Schwitzgebel et al., 2024).

Annoni et al. (2026) identify an important temporal limitation of personalized LLMs: people change, while a model trained from an earlier corpus can preserve a snapshot. Their critique raises the question of which version of a person a purported digital twin represents.

Persistent adaptive replicas make that question harder. If D is continually synchronized with H, comparison remains relatively straightforward. But if D learns from interactions and events not shared by H, synchronizing it back toward actual H may erase precisely the independent development the system was intended to possess.

There are now two moving trajectories.

Wang et al. (2026) examine one form of this developmental problem directly. Their BFI-Adapt benchmark evaluates the **directional fidelity** of personality changes in LLM agents after simulated life events, asking whether changes occur in directions consistent with longitudinal evidence about human personality development.

This makes change itself an object of evaluation, but its principal target is population-level human development rather than the unrealized development of a particular individual.

Evidence that agents tend to change in directions observed across human populations may support the claim:

> AI personas exposed to event E tend to change in a psychologically plausible human direction.

It does not establish:

> this replica of H changed as this particular H would have changed.

Population-level developmental plausibility and person-specific developmental fidelity are different evidential achievements. A particular person may depart markedly from a population tendency without thereby becoming less themselves.

2.4 Prediction, proxies, and the developmental gap

Sweeney (2023) argues that autonomous avatars acting as proxies may face an epistemic gap when they must decide without knowing what the represented person would have done. Similarly, the Personalized Patient Preference Predictor proposed by Earp et al. (2024) is designed to estimate what a particular incapacitated patient would choose.

These are already person-specific counterfactual inferences. The present problem adds another layer. Compare:

> What would H choose in situation C?

with:

> What would H have become after a prolonged developmental history corresponding to E_D, and what would that counterfactual H then choose?

The first predicts an unobserved response. The second must first attribute an unobserved later state to the person; beliefs, priorities, memories, values, dispositions, and decision rules may themselves be counterfactual outcomes.

This does not make developmental counterfactuals a new species of reasoning. The narrower point is that, once such an inference enters a continuing fidelity claim, a relation between replica and observed source becomes partly a claim about an unrealized version of that source.

The general asymmetry between observed outcomes and unrealized counterfactual outcomes is of course foundational in causal inference (Holland, 1986). Recent work on causal digital twins reinforces the same point in a digital-twin setting: counterfactual quantities can remain assumption-dependent even when observable aspects of a twin are successfully validated, and predictive fidelity need not establish counterfactual validity (Freeman et al., 2026; Laudy, 2026). The contribution here is to identify where that familiar problem enters person-specific representational fidelity after independent development.

3. The Fidelity Regime Shift

3.1 Fidelity requires a temporally specified target

The person-fidelity claims considered here are incomplete unless target and relevant properties are specified.

Let

$$F(D, H, t, X)$$

represent a claim that digital replica D is faithful to human source H, with respect to some set of properties X, at temporal target t.

This is clarifying notation, not a quantitative theory of identity. X might include beliefs, autobiographical memories, linguistic style, moral judgments, risk preferences, or emotional dispositions.

Suppose D is constructed at t_0 from records of H available up to that point.

Type I — Historical fidelity

Does D represent H as H was at t_0?

The comparator is:

$$H_{t_0}.$$

Relevant features may be imperfectly known, but at least some are empirically anchored in source records, decisions, and behavior.

At later time t_1, we can ask:

Type II — Contemporary resemblance

Does D at t_1 resemble actual H at t_1?

The comparator is:

$$H_{t_1}.$$

This relationship can also be empirically assessed, although psychological properties may require inference.

A stronger claim asks:

Type III — Counterfactual developmental attribution

Has D developed as H would have developed under developmentally corresponding conditions?

The relevant comparator becomes:

$$H^*_{t_1},$$

where $H^*_{t_1}$ represents the state H would have reached had H undergone a developmental history corresponding, in the respects relevant to X, to that undergone by D.

Unlike H_{t_0} and actual H_{t_1}, this comparator was not realized.

Using H^* does not require treating D and H as numerically identical; it asks only what the same source person would have been like under different conditions. The metaphysical status of D relative to H remains open.

Type III is not implicit in every use of fidelity. It becomes relevant when the system is claimed not merely to preserve H or resemble current H, but to remain faithful **through its own development**.

3.2 Material divergence is claim-relative

Represent D schematically as:

$$D_{t_1} = g\left(D_{t_0}, E_D\right),$$

where E_D denotes relevant post-origin interactions, inputs, memory operations, learning events, and other developmental influences.

Similarly:

$$H_{t_1} = h\left(H_{t_0}, E_H\right).$$

The mere fact that

$$E_D \neq E_H$$

does not automatically trigger a meaningful shift for every fidelity claim.

The difference must be material to X.

A post-origin difference is **material** when, under the account of development presupposed by the fidelity claim itself, that difference could reasonably affect the properties X being evaluated.

Different weather on one afternoon need not matter to stable linguistic habits; a decade of different relationships, losses, political events, or professional experiences may matter greatly to later values or judgments.

Materiality is therefore claim-relative rather than a universal threshold.

For Type III claims to have evidential force, the developmental account determining materiality cannot be chosen after the outcome is known. Claimants should specify in advance which post-origin differences are expected to matter to X and why. Otherwise divergence and resemblance can be redescribed after the fact. Post hoc specification does not make the claim false, but weakens validation based on it.

3.3 Developmental correspondence is a condition on the claim

Human experience and machine input cannot simply be treated as literal equivalents. A human bereavement, illness, bodily danger, aging, or parenthood is not automatically replicated by providing an AI with text describing the same event.

Type III therefore presupposes some defensible relation of **developmental correspondence** between the relevant parts of D's history and the hypothetical conditions applied to H.

The correspondence need not be perfect, but it must be specified well enough for the claim to have determinate content. If no defensible correspondence can be stated for X, the Type III claim is under-specified: we do not yet know which counterfactual trajectory of H is asserted.

Thus $H^*_{t_1}$ is not unconstrained; it is defined relative to the developmental claim being made.

3.4 The missing comparator

Suppose:

$$D_{t_1} \neq H_{t_1}.$$

That difference may be empirically established. But it does not by itself reveal whether the discrepancy results from poor replication or materially different development.

The comparison needed for Type III would instead be:

$$D_{t_1} \text{ versus } H^*_{t_1}.$$

Actual H experienced E_H, not the developmentally corresponding history invoked by the Type III claim. The strongest developmental comparator was therefore never realized.

Counterfactual unobservability is not itself novel. What matters is where it enters the representational relationship. At t_0, "D is faithful to H" may describe an empirically anchored relation to the source. After materially relevant divergence, the stronger claim that D remained faithful **by developing as H would have developed** attributes an unrealized developmental trajectory to H.

The same language can therefore conceal a change in evidential target.

> **Fidelity Regime Shift Principle. When a person-specific digital replica acquires a materially independent developmental history, an unqualified claim that it remains "faithful" to its human source no longer carries a single evidential burden. Historical fidelity remains assessable against observed features of the source; contemporary resemblance remains observable but is confounded by divergent experience; and claims of faithful development require counterfactual attribution concerning an unrealized trajectory of the particular source person.**

The argument can be stated minimally.

P1. Person-fidelity claims require specification of the relevant source state and properties.

P2. Historical fidelity can be empirically anchored in evidence about an actual source state.

P3. Persistent adaptive replicas can acquire later states partly through post-origin development.

P4. When source and replica histories differ in ways material to the properties under evaluation, actual contemporary H is not a matched developmental comparator for D.

P5. A claim that D changed as H would have changed under developmentally corresponding conditions therefore requires a counterfactual comparator H^*, defined relative to a specified correspondence.

P6. An unrealized counterfactual trajectory and an empirically anchored source state have different evidential statuses.

Therefore:

> **The unqualified concept of continuing "fidelity to the person" must be disaggregated after sufficiently material experiential divergence.**

The regime shift does not replace all comparators with H^*; it reveals that claims compressed under "fidelity" require different comparators.

4. Why Divergence Cannot Diagnose Failure by Itself

Once source and replica have materially different post-origin histories, disagreement becomes an ambiguous signal.

Suppose:

$$D_{t_1} \neq H_{t_1}.$$

One explanation is replication failure. D may have drifted away from characteristics that should have remained stable, overweighted irrelevant inputs, or updated poorly.

Another is legitimate experiential divergence. D and H may differ because they underwent different developmental histories.

4.1 The wrong control

Imagine that D spends ten years interacting with a community H never meets, encountering information H never receives, maintaining relationships H never forms, and learning from decisions H never makes.

At t_1, D endorses a position H rejects.

Actual H_{t_1} can establish that H and D now differ. It cannot, without further assumptions, establish why. If the claim under examination is Type III, the relevant comparison would be with $H^*_{t_1}$.

Contemporary H is both the source of the replica and, for developmental validation, a control exposed to the wrong history.

That does not make contemporary H irrelevant. Strong disagreement may provide evidence against a model, especially where the disputed feature is believed to be highly stable or insensitive to the differing experiences.

Difference alone, however, does not identify its own cause.

4.2 Memory as a developmental variable

Memory policy makes the problem concrete.

Human autobiographical memory is selective and reconstructive rather than a passive archive, and external memory technologies can participate in how autobiographical identity is maintained and reconstructed (Burkell, 2016; Heersmink, 2022). Recent work also examines AI systems as active curators or co-narrators of personal memory rather than merely storage devices (Smart et al., 2026; Telakivi, 2026).

Consider three replicas constructed from the same source. One retains almost every interaction indefinitely. Another selectively forgets. A third repeatedly summarizes and reconstructs previous information.

Years later, different retention, retrieval, and reconstruction policies may influence beliefs and judgments. A complete digital archive might preserve historical information exceptionally well while producing a developmental process unlike ordinary human autobiographical remembering. A selective memory architecture might resemble some aspects of human cognitive functioning more closely while preserving less historical information.

The point is not to identify the correct memory architecture. It is that memory policy forms part of the replica's post-origin developmental history.

4.3 Divergence is evidence, not diagnosis

The framework does not make developmental fidelity unfalsifiable.

Large, systematic, or theoretically unexpected divergence may provide strong evidence of failure. Repeated observations, within-person regularities, psychological theory, population data, and mechanistic knowledge can constrain plausible developmental trajectories.

The narrower claim is:

$$D_{t_1} \neq H_{t_1}$$

does not entail:

$$D \text{ failed to model } H.$$

The inference requires assumptions about which differences in history matter and how H would respond to them.

5. Why Resemblance Cannot Validate Development by Itself

Suppose instead:

$$D_{t_1} \approx H_{t_1}.$$

H and D express similar values and reach similar judgments. This is genuine evidence of contemporary resemblance.

It is not sufficient, by itself, to establish that D developed as H would have developed through D's history.

5.1 Multiple paths to the same endpoint

Resemblance can arise through several mechanisms.

Origin anchoring. D remains so strongly constrained by its t_0 state that later experience has little effect.

Constrained adaptation. D updates but only within boundaries designed to preserve recognizable characteristics.

Convergence. H and D arrive at similar positions through different developmental pathways.

Coincidence. Similar outputs emerge despite importantly different underlying processes.

A system intended to develop might therefore resemble H because it developed too little.

This produces a stronger consequence than mere underdetermination. Suppose the developmental account presupposed by a Type III claim predicts that exposure to a specified class of experiences should shift property X in direction Δ. If D undergoes those conditions yet remains close to actual H, while actual H did not undergo them, resemblance is no longer simply non-diagnostic. It is anomalous relative to the developmental claim and may count against it. Under some specifications, therefore, resemblance to the source can be disconfirming evidence of developmental fidelity.

The validation sign can reverse: the criterion that looks strongest for a preservation-oriented replica—continued similarity to the source—may be evidence that a development-oriented replica failed to develop as its own Type III claim predicted. The disconfirming reading depends on the developmental account not also predicting the same endpoint for H's actual history; where it does, convergence again renders resemblance uninformative.

Endpoint resemblance does not identify the path by which the endpoint was reached.

5.2 Human-like development is not person-specific development

Wang et al. (2026) show why this distinction matters. A model whose personality changes in directions consistent with longitudinal human findings may exhibit human-like developmental plausibility.

But population evidence cannot uniquely identify the trajectory of a particular H. A person may depart substantially from a population mean without thereby ceasing to be themselves.

Population-level developmental evidence can constrain an estimate of $H_{t_1}^{*}$. It cannot substitute for observing it.

We should therefore distinguish contemporary resemblance, population-level developmental plausibility, and person-specific counterfactual developmental fidelity. These are related evidential achievements, but they are not interchangeable.

6. From Fidelity to Attribution

Consider five statements:

1. Alex said X.
2. Alex believes X.
3. Alex would have said X.
4. A model based on Alex predicts that Alex would say X.
5. A model derived from Alex at t_0, after developing through its own subsequent history, now says X.

They have different epistemic statuses.

The first reports an observed event. The second attributes a present state. The third makes a counterfactual attribution. The fourth identifies the counterfactual explicitly as a model prediction. The fifth reports the judgment of a historically derived but subsequently developing system.

Persistent personal AI makes these distinctions easy to collapse.

6.1 From prediction to evolved judgment

Existing work already recognizes the danger of inaccurately attributing replica outputs to their human origins (Karpus & Strasser, 2025) and the epistemic problem faced when autonomous proxies infer decisions for represented persons (Sweeney, 2023).

At t_0, a replica's output may primarily reflect information derived from H. At t_1, the output may reflect source-derived characteristics, subsequent model updates, interactions and information experienced only by D, memory policies, and prior decisions made by D itself.

The output remains causally descended from the source.

Causal ancestry, however, does not make it the source person's present judgment.

Calling such an output "what Alex thinks" conceals its provenance. Calling it "what Alex would think" makes an additional counterfactual claim. Calling the system simply "Alex" may conceal both.

6.2 Attributional authority

The source relationship can confer authority.

A generic AI stating that X is the right decision has one kind of status. An AI presented as saying “Alex says X is the right decision” may draw on Alex’s reputation, expertise, personal relationships, emotional importance, or moral standing.

This matters in authorship, public communication, endorsement, family decisions, surrogate decision-making, and posthumous representation.

Person-specific preference predictors make their inferential structure comparatively explicit: the system estimates what a patient would choose (Earp et al., 2024). A persistent replica creates a different case because the predictor may itself have changed through an independent history.

Suppose D was constructed from H at age fifty. H later dies. Twenty years after its creation, D recommends a decision H never confronted.

Compare:

> H would have chosen this.
>
> D predicts that H would have chosen this.
>
> D, which originated as a replica of H and subsequently developed independently, chooses this.

The third does not entail the first or second.

Persistent personal AI therefore sharpens an attribution problem before it requires a solution to the metaphysics of personal identity.

6.3 Authority cannot run only one way

Treating a sufficiently divergent replica as a new entity is coherent.

But doing so has consequences for attribution. One cannot invoke developmental independence to explain why D may legitimately differ from H while retaining the H relationship without qualification whenever D benefits from H’s authority.

Independence cannot explain divergence while ancestry alone supplies attributional authority.

This is a consistency constraint on claim-making rather than a direct entailment of the preceding epistemic argument.

The fork must remain visible in both directions.

7. The Living Ship: An Arcadia Thought Experiment

Consider two replicas inspired by the relation between Tochiro and the Arcadia in Leiji Matsumoto’s fictional universe. No claim about the canonical interpretation of the fiction is required.

Immediately before Tochiro’s death, two equally accurate replicas are created.

Ship A is preservation-oriented. Its memories, values, and characteristic judgments are kept close to the source state. Fifty years later, it may still answer extremely well: “What would Tochiro at t_0 probably have said?” Its historical fidelity may be high.

Ship B is development-oriented. It travels with a crew, forms relationships, encounters failures and losses, learns from decisions, and revises the salience of earlier memories. Fifty years later, it makes a judgment with which historical Tochiro would probably have disagreed.

That disagreement does not by itself show that Ship B failed. Change was part of what it was designed to do. But neither does its developmental richness establish that historical Tochiro would have become what Ship B became. The required Tochiro did not live Ship B's history.

Ship A and Ship B therefore optimize different relationships to their common source. Ship A may provide better evidence about an observed past person. Ship B may be the richer developing branch. Neither fact settles whether Ship B's later judgment may be attributed to Tochiro.

The useful question is not:

> Which ship is the real Tochiro?

It is:

> Which claims about Tochiro are justified by what each ship has become?

8. Epistemic Governance: Keeping the Evidential Status Visible

Existing work already develops broad governance requirements for digital replicas and posthumous AI, including consent, transparency, distinguishability, control, and constraints on autonomous evolution (Hollanek & Nowaczyk-Basińska, 2024; Karpus & Strasser, 2025; Spitale & Germani, 2026). Danaher and Nyholm's (2025) minimally viable permissibility principle likewise includes transparency in interactions between digital duplicates and third parties.

The present argument does not replace those requirements. It sharpens one dimension of transparency. Knowing that one is interacting with an AI duplicate does not reveal whether a particular output is retrieved from the source, reconstructed from incomplete records, predicted for the source, generated as a counterfactual developmental simulation, or expressed as the present judgment of an independently developed replica.

The resulting requirement is therefore narrower and more specific: if those output types have different evidential relations to the source, a persistent personal AI should not make those differences irrecoverable.

8.1 Persistent provenance

At minimum, adequate provenance may need to identify the source person, the temporal source state or source period, the source materials on which the replica was based, the last meaningful synchronization with the living source, whether subsequent learning occurred autonomously, the relevant character of post-origin development and memory policy, the developmental account used to specify materiality and correspondence for any Type III claim, and the epistemic status under which a given output is offered.

An output may be offered as:

Retrieval — H actually said or wrote this.

Reconstruction — the system reconstructs an incompletely recorded historical position.

Prediction — the system estimates what H would say in a specified situation.

Counterfactual developmental simulation — the system estimates what H might have become or believed after an unrealized developmental history.

Autonomously evolved judgment — the output is the present judgment of D after its own subsequent history.

These distinctions need not appear as cumbersome warnings before every sentence. They should remain recoverable and intelligible when attribution matters.

8.2 Why ordinary AI disclosure is insufficient

A label such as “AI-generated” distinguishes machine output from directly human-generated content. It does not distinguish a reconstruction of 2026-H from a prediction about 2036-H, nor either from the autonomous judgment of a replica that has accumulated ten years of independent history.

Likewise, “This is a simulation of Alex” leaves unanswered which temporal Alex is represented, when synchronization ended, what independent development followed, and whether an output is offered as retrieval, prediction, simulation, or evolved judgment.

The relevant transparency problem is temporal and relational, not merely technological.

The design problem arises when developmental independence increases while representational presentation remains unchanged. A system may become progressively less straightforwardly attributable to H while continuing to use H’s name, face, voice, biography, and first-person perspective.

Phenomenological continuity can then hide epistemic distance.

The aim is not to prohibit the fork. It is to keep the fork visible.

9. Objections

9.1 This is merely branching

Branching is a premise, not the contribution. Deguchi et al. (2022) already provide a model in which a digital twin shares an earlier history with an original and subsequently lives its own life.

The present claim begins after that point. A branching model identifies two trajectories. It does not determine which later comparator warrants a claim that one trajectory remains developmentally faithful to the person represented by the other.

9.2 This is merely the fundamental problem of causal inference

The general counterfactual problem is familiar (Holland, 1986; Freeman et al., 2026; Laudy, 2026). The contribution is not a new solution to it, but the identification of what happens when an unrealized trajectory becomes the comparator required by a continuing person-specific fidelity claim. In Type III, counterfactual inference becomes part of what developmental fidelity itself asserts.

9.3 Humans already make substituted judgments

Humans routinely ask what another person would have wanted. The difference is not that AI invents counterfactual attribution.

“She would have wanted X” displays its counterfactual grammar. A highly realistic personal AI may make an analogous inference in the source person’s name, voice, appearance, biography, and first-person perspective. A persistent replica adds the possibility that the statement is generated from a state partly produced by its own independent history.

Counterfactual prediction can thereby phenomenologically resemble continued observation.

9.4 A sufficiently accurate model could solve the problem

Better models can strengthen Type III claims through longitudinal observations, causal models, source-specific behavior, and psychological evidence. But predictive strength and observational status remain different: even an excellent estimate of $H^*_{t_1}$ is still an estimate of an unrealized trajectory.

9.5 Treat D as a new entity

That solution is coherent. But if D is treated as a new entity, its later judgments should not be attributed without qualification to H. Historical ancestry remains relevant; it does not license unqualified attribution of independently evolved judgments.

9.6 Human and machine experiences cannot be equivalent

The framework does not require literal equivalence. Type III requires only that the developmental correspondence presupposed by the claim be specified.

Where no defensible correspondence can be stated, the appropriate conclusion is not that D must have developed unfaithfully, but that the Type III claim is under-specified.

9.7 Humans themselves are unpredictable

Perfect developmental prediction is unnecessary.

A reconstruction, prediction, counterfactual developmental simulation, and independently evolved judgment can all be uncertain. They remain different kinds of claims because uncertainty does not collapse their evidential targets into one.

9.8 “Fidelity Regime Shift” is merely a new label

The terminology is dispensable; the distinction is not. “Fidelity to H” can refer to an observed past H, an actual contemporary H, or an unrealized H under a different developmental history. These claims differ not merely in accuracy but in comparator and evidential support. “Regime shift” is shorthand for that change in evidential structure.

9.9 This presupposes a metaphysics of personal identity after all

The notation H^* asks what H would have been like under a different developmental history. It does not require D and H to be numerically identical, only the familiar practice of holding the source person fixed while varying circumstances, as in asking what a patient would have wanted under conditions never encountered. Type III therefore requires no conclusion about whether D and H are one person, two persons, or stand in another identity relation. H^* is a comparator for H’s unrealized development, not an identity claim about D.

10. Conclusion: After the Fork

A persistent digital replica begins with a comparatively strong evidential relation to its source because it is constructed from evidence about an actual person’s past. Independent development changes that relation.

Historical fidelity remains assessable against the source record, and contemporary resemblance against the actual source. But the stronger proposition that the replica developed **as that particular source person would have developed** requires inference toward an unrealized trajectory.

Neither divergence from actual H nor resemblance to actual H settles that claim by itself.

The problem becomes sharper when persistent personal AI is designed to learn. Preservation keeps the replica near an observed source; development can move later states away from direct empirical anchoring in that source.

The practical consequence is attributional. Later outputs should not move silently between retrieval, reconstruction, prediction, counterfactual developmental simulation, and autonomously evolved judgment. As independent development accumulates, temporal and developmental provenance becomes increasingly important.

This conclusion does not depend on whether digital replicas are conscious, possess personhood, or preserve numerical identity.

A digital replica may begin as a reconstruction of an observed person. Once it develops through a history of its own, continuing claims that it represents what that person would have become require a different evidential justification.

After the fork, provenance and attribution must travel together.

A replica may become harder to validate as a representation of a person precisely because it has become better able to develop like a person.

Declarations

Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The author has no relevant financial or non-financial interests to disclose.

Author Contributions

George Kujo conceived the argument, conducted the literature review, developed the conceptual framework, drafted and revised the manuscript, and approved the final version.

Data Availability

Not applicable. No datasets were generated or analysed during the current study.

Use of Generative AI

Generative AI tools were used during manuscript development to assist with literature searching, structural analysis, language editing, and drafting. The author independently evaluated and verified the relevant sources, arguments, interpretations, and wording, revised the generated material, and takes full responsibility for the content of the final manuscript.

References

Annoni, M., Battisti, D., & Marchegiani, B. (2026). Personalised LLMs and the risks of the digital twin metaphor. *AI & Society, 41*, 5177–5189. https://doi.org/10.1007/s00146-026-02875-4

Burkell, J. A. (2016). Remembering me: Big data, individual identity, and the psychological necessity of forgetting. *Ethics and Information Technology, 18*(1), 17–23. https://doi.org/10.1007/s10676-016-9393-1

Danaher, J., & Nyholm, S. (2024). Digital duplicates and the scarcity problem: Might AI make us less scarce and therefore less valuable? *Philosophy & Technology, 37*, Article 106. https://doi.org/10.1007/s13347-024-00795-z

Danaher, J., & Nyholm, S. (2025). The ethics of personalised digital duplicates: A minimally viable permissibility principle. *AI and Ethics, 5*, 1703–1718. https://doi.org/10.1007/s43681-024-00513-7

Deguchi, Y., Onishi, T., Akiyoshi, R., Yagisawa, T., & Yamamori, M. (2022). Logic of alternative-I. *Asian Journal of Philosophy, 1*, Article 51. https://doi.org/10.1007/s44204-022-00050-2

Earp, B. D., Porsdam Mann, S., Allen, J., Salloch, S., Suren, V., Jongsma, K., Braun, M., Wilkinson, D., Sinnott-Armstrong, W., Rid, A., Wendler, D., & Savulescu, J. (2024). A personalized patient preference predictor for substituted judgments in healthcare: Technically feasible and ethically desirable. *The American Journal of Bioethics, 24*(7), 13–26. https://doi.org/10.1080/15265161.2023.2296402

Freeman, N. L. B., Zou, Y., Lee, K., Yu, C., Li, D., & Kosorok, M. R. (2026). *The fidelity and feedback traps: The case for health digital twins as modular evolving causal systems* [Preprint]. arXiv. https://arxiv.org/abs/2607.27192

Heersmink, R. (2022). Extended mind and artifactual autobiographical memory. *Mind & Language, 37*(4), 659–673. https://doi.org/10.1111/mila.12353

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Hollanek, T., & Nowaczyk-Basińska, K. (2024). Griefbots, deadbots, postmortem avatars: On responsible applications of generative AI in the digital afterlife industry. *Philosophy & Technology, 37*, Article 63. https://doi.org/10.1007/s13347-024-00744-w

Karpus, J., & Strasser, A. (2025). Persons and their digital replicas. *Philosophy & Technology, 38*, Article 25. https://doi.org/10.1007/s13347-025-00854-z

Laudy, O. (2026). *The digital twin counterfactual framework: A validation architecture for simulated potential outcomes* [Preprint]. arXiv. https://arxiv.org/abs/2604.01325

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2024). Creating a large language model of a philosopher. *Mind & Language, 39*(2), 237–259. https://doi.org/10.1111/mila.12466

Smart, P., Clowes, R., Krueger, J., & Boniface, M. (2026). The story of your life: Large language models and personal memory. *Review of Philosophy and Psychology*. Advance online publication. https://doi.org/10.1007/s13164-026-00831-1

Spitale, G., & Germani, F. (2026). The making of digital ghosts: Designing ethical AI afterlives. *Ethics and Information Technology, 28*, Article 34. https://doi.org/10.1007/s10676-026-09910-4

Sweeney, P. (2023). Avatars as proxies. *Minds and Machines, 33*, 525–539. https://doi.org/10.1007/s11023-023-09643-z

Sweeney, P. (2024). Avatars and the value of human uniqueness. *Philosophy & Technology, 37*, Article 118. https://doi.org/10.1007/s13347-024-00811-2

Telakivi, P. (2026). Remembering with AI: From distributed memory to AI-curated and human-AI co-memory. *Review of Philosophy and Psychology*. Advance online publication. https://doi.org/10.1007/s13164-026-00815-1

Voinea, C., Porsdam Mann, S., Register, C., Savulescu, J., & Earp, B. D. (2024). Digital duplicates, relational scarcity, and value: Commentary on Danaher and Nyholm (2024). *Philosophy & Technology, 37*, Article 121. https://doi.org/10.1007/s13347-024-00813-0

Wang, M., Wang, P., Yang, X., Wang, D., Feng, S., Nah, F. F.-H., & Lim, E.-P. (2026). *Do AI personas grow? Analyzing and benchmarking personality evolution in LLM agents after life events* [Preprint]. arXiv. https://arxiv.org/abs/2608.06485