分岐の後で：進化するデジタル・レプリカにおいて忠実性の主張が反実仮想になるとき

George Kujo（九条丈二）
Independent Researcher, Tokyo, Japan
ORCID: 0009-0006-6864-0145
E-mail: Research@georgekujo.com

プレプリント／著者版 — Version 1.1, 2026 年 9 月

本稿は英語原稿 *After the Fork: When Fidelity Claims Become Counterfactual in Evolving Digital Replicas* の著者による日本語版である。英語版との間に意味上の差異が生じた場合は、英語版を優先する。

要旨

特定人物を再現するデジタル・レプリカは、作成後も存続し、学習し、変化するよう設計され得るようになっている。これは、後年の出力を単一の「人間の情報源への忠実性」という概念で評価すると見えにくくなる問題を生む。本稿は三つの主張を区別する。すなわち、経験的に裏づけられた過去の本人状態に対する**歴史的忠実性**、現時点の本人に対する**同時点類似性**、そして「レプリカが、その特定人物が発達上対応する条件のもとで発達したであろう仕方で発達した」という**反実仮想的発達帰属**である。レプリカと本人が作成後に実質的に異なる履歴を獲得すると、現時点の本人との不一致も一致も、それだけでは発達的忠実性を検証できない。不一致はモデルの失敗かもしれないが、正当な経験差の結果かもしれない。一致もまた、初期状態への固定、制約された適応、収束、あるいは偶然によって生じ得る。したがって最も強い発達的主張の比較対象は、現実の本人ではなく、その本人が実際には歩まなかった反実仮想的軌跡となる。本稿は、この証拠上の比較対象の変化を**忠実性レジーム・シフト（Fidelity Regime Shift）**と呼ぶ。この議論は、デジタル人格の personhood、意識、数的同一性を前提としない。実務的帰結は帰属にある。後年の出力は、検索、再構成、予測、反実仮想的発達シミュレーション、自律的に発達した判断として区別されるべきである。したがって持続的な personal AI には、発達上の分岐後も本人との関係を可視化し続ける provenance が必要である。レプリカは、人間のように発達する能力を高めるほど、まさにそのために、ある人間の表象として検証することが難しくなり得る。

キーワード： デジタル・レプリカ；忠実性；反実仮想推論；personal AI；帰属；personal identity

1. はじめに

特定人物のデジタル・レプリカは通常、表象として評価される。そのシステムは本人の話し方を再現するか。本人の信念、記憶、選好、特徴的な推論、あるいは未知の質問への応答を再現するか。特定人物の言語資料から学習したモデルが、その本人の新規回答と識別困難な応答を生成し得ることはすでに示されている（Schwitzgebel et al., 2024）。

こうした比較は完全ではないが、記録された発言、行動、判断、著作など、本人に関する証拠によって経験的に支えることができる。

問題は、レプリカが静止したままでないときに難しくなる。

2026 年に、生きている人物 H の大量の記録から、特定人物向け人工エージェント D が作られたとしよう。作成時点で D は、H の記録済みの見解や特徴的な判断傾向をかなりよく再現している。ところが D はその後も存続する。他者と相互作用し、新しい情報を受け取り、記憶を蓄積し、信念を修正し、自らの判断から学ぶ。一方 H も、人間として通常の生活を続ける。

2036 年までに、H と D はかなり異なる履歴を経験している。

二者に同じ難しい問いを投げかけ、異なる答えが返ったとしよう。一つの説明は、D が H を正確に表象しなくなったというものだ。モデルがドリフトしたか、更新手続きが不適切だったのかもしれない。しかし別の説明もある。D は、経験から学習するよう設計されていたからこそ変化したのかもしれない。H が発達上対応する履歴を経験していたなら、H 自身も変化した可能性がある。したがって観察された不一致だけでは失敗は確定しない。

逆に、2036 年になっても H と D が同じ答えをしたとしよう。その類似は経験的に評価できるが、同じ結論に至った原因は終点だけからは分からない。D が H らしい仕方で本当に発達したのかもしれない。あるいは 2026 年の状態に強く固定され、ほとんど適応しなかったのかもしれない。別経路から同じ結論へ収束した可能性も、偶然一致した可能性もある。したがって観察された類似も、それだけでは発達の成功を立証しない。

ここには保存と発達の緊張がある。保存志向のレプリカは、ほとんど変化しないからこそ、観察済みの過去の本人との比較が容易である。これに対して発達志向のレプリカは、自らの履歴を獲得する能力という重要な意味でより「人間らしく」なり得るが、その結果として、起源となった本人との比較による検証は難しくなる。

この問題の構成要素の多くはすでに知られている。分岐と自律的なデジタル counterpart は既存研究に存在し（Deguchi et al., 2022）、近年の digital duplicates 研究は、自律性、希少性、真正性、関係性、許容可能性、誤帰属を検討している（Danaher & Nyholm, 2024, 2025; Karpus & Strasser, 2025; Sweeney, 2024; Voinea et al., 2024）。特定人物向けシステムは、表象される人物の未観察判断を予測することもできる（Earp et al., 2024; Sweeney, 2023）。また近年、変化する人物のどの時点を personalized model が表象するのかという問題（Annoni et al., 2026）、発達する digital-afterlife system における identity drift（Spitale & Germani, 2026）、LLM persona の人間らしい人格変化の benchmark（Wang et al., 2026）も論じられている。 この論争は複製の価値と許容可能性を主に扱うのに対し、本稿が問うのは、duplicate が発達した後にその duplicate について行われる忠実性主張の証拠構造である。

本稿の貢献はより狭い。特定人物のレプリカが、本人は経験しなかった実質的に独立した発達史を獲得した後も、その人への**忠実性**が継続すると主張するとき、その主張に何が起こるのかを問う。

「レプリカは本人に忠実であり続ける」という一つの表現の中に圧縮されがちな三つの主張を区別する。

歴史的忠実性は、D が過去のある時点の H を表象しているかを問う。

同時点類似性は、現在の D が現実の H に似ているかを問う。

反実仮想的発達帰属は、D が、発達上対応する条件のもとでこの特定の H が発達したであろう仕方で発達したかを問う。

最初の二つは現実の本人の状態に経験的な根拠を持ち得る。三つ目は、その本人が実際には歩まなかった発達軌跡を参照する。

私は、この証拠上の比較対象の変化を**忠実性レジーム・シフト（Fidelity Regime Shift）**と呼ぶ。ここでいう「レジーム・シフト」は形而上学的な転換点や厳密な数値閾値を意味しない。作成後の履歴が評価対象の性質に関連し始めたとき、継続的な忠実性主張ごとに必要な証拠が変化することを指す。また、分岐後のすべての忠実性主張が反実仮想になるわけではない。反実仮想的な比較対象が必要になるのは、独立した変化そのものが「その本人ならこう変わったはず」という意味で忠実であると主張するときである。

中心問題は、デジタル分岐が形而上学的に「本当に」本人なのかではない。分岐後に、両者の関係について私たちは何を正当に主張できるのかである。

2. デジタル・レプリカ、分岐する自己、発達する persona

2.1 デジタル・レプリカと心理的連続性

Karpus and Strasser (2025) は Parfit の心理的連結性の議論を踏まえ、人物とそのデジタル・レプリカとの関係は、全か無かで理解する必要はないと論じる。関連する結びつきには、記憶、信念、欲求、傾向、世界観などが含まれ、それらは程度の差を持ち得る（Parfit, 1984）。

この見方の利点は、数的同一性を先に解決しなくても表象について議論できる点にある。デジタル・システムは、ある人物の心理的に重要な特徴の多くを継承していても、その人物と数的に同一だと仮定する必要はない。

Karpus and Strasser は、著者性、責任、帰属にも含意があることを指摘し、とりわけ source と replica の心理的関係を過大評価する危険を論じている。したがって本稿は、デジタル・レプリカに帰属問題があること自体を新発見とはしない。より狭い関心は、本人が経験していない履歴によって後年の状態が形成されるとき、source–replica 関係の証拠基盤がどう変わるかである。

作成時、レプリカは H から記憶、コミットメント、選好、特徴的な推論を継承し得る。しかし後年の状態は、経路依存的な発達の産物になり得る。

2.2 分岐するデジタル自己

Deguchi et al. (2022) はすでに、ある人物と分岐点まで履歴を共有し、その後は自律的なエージェントとして独自の人生を送る digital counterpart をモデル化している。本稿はこの構造を前提とする。

H と D が時点 t_0 まで関連する履歴を共有し、その後は異なる軌跡を進むとする。分岐モデルは、なぜそれが可能かを説明する。しかし、後年の現実の H に関する証拠を D の評価にどう使うべきかは決めない。十年後に D が H と異なれば、それは複製の失敗かもしれないし、異なる履歴の結果かもしれない。

したがって本稿が扱うのは分岐そのものではなく、分岐後に source–replica 関係についての継続的主張をどう評価するかである。

2.3 時間的 snapshot から発達する agent へ

特定人物向け言語モデルは、この問題の一部をすでに具体化している。DigiDan 実験では Daniel Dennett の著作から学習したモデルが新規の哲学的回答を生成し、専門家でも Dennett 本人の回答と区別しにくい場合があった（Schwitzgebel et al., 2024）。

Annoni et al. (2026) は personalized LLM の重要な時間的限界を指摘する。人間は変化するが、過去の corpus から学習したモデルは snapshot を保存し得る。そのため、いわゆる digital twin が「人物のどの version」を表象するのかが問題になる。

持続的で適応的なレプリカは、この問題をさらに難しくする。D が H と継続的に同期されるなら比較は比較的容易である。しかし、D が H の経験しない相互作用や出来事から学習することを許されているなら、現実の H へ再同期することは、まさにシステムが獲得するよう設計された独立発達を消してしまい得る。

ここでは二本の「動く軌跡」が存在する。

Wang et al. (2026) は、この発達問題の一形態を直接扱っている。BFI-Adapt benchmark は、模擬的な life event 後の LLM agent の人格変化について、縦断的な人間の人格研究と同じ方向に変化するかという **directional fidelity** を評価する。

これは「変化そのもの」を評価対象にする点で重要である。しかし主たる比較対象は、特定個人の未実現の発達ではなく、集団レベルの人間発達である。

たとえば次の主張は支持され得る。

> 事象 E にさらされた AI persona は、人間心理学的にもっともらしい方向へ変化する傾向がある。

しかし、そこから次は導けない。

> この H のレプリカは、この特定の H が変化したであろう仕方で変化した。

集団レベルの発達妥当性と、個人特異的な発達的忠実性は異なる証拠上の達成である。ある個人が集団傾向から大きく逸脱しても、それだけで「その人らしくなくなる」わけではない。

2.4 予測、proxy、発達上の gap

Sweeney (2023) は、自律的 avatar が proxy として行動するとき、表象対象の人物が何をしたであろうか分からない場合に認識論的 gap に直面すると論じる。同様に、Earp et al. (2024) の Personalized Patient Preference Predictor は、意思決定できない特定患者が何を選ぶかを推定することを目的とする。

これらはすでに個人特異的な反実仮想推論である。本稿の問題はさらに一段重なる。次を比べよう。

> 状況 C で H は何を選ぶだろうか。

と、

E_D に対応する長期の発達史を経験したなら H はどのような人物になり、その反実仮想的な H はその後何を選ぶだろうか。

前者は、表象対象となる人物の未観察の応答を予測する。後者ではまず、その人物の未観察の後年状態を帰属しなければならない。後の回答を生み出す信念、優先順位、記憶、価値、傾向、意思決定規則そのものが反実仮想的な結果になり得る。

これは developmental counterfactual を新しい種類の反実仮想推論とする主張ではない。より狭い点は、その推論が継続的な忠実性主張の内部に入ると、観察済みの source と replica の関係が、未実現の source version についての主張を部分的に含むようになることである。

観察された結果と実現しなかった反実仮想結果の非対称性は、因果推論の基礎的問題である（Holland, 1986）。近年の causal digital twin 研究も同じ点を digital-twin の文脈で強調している。観察可能な側面がよく検証されても反実仮想量は仮定依存であり得て、predictive fidelity は counterfactual validity を保証しない（Freeman et al., 2026; Laudy, 2026）。本稿の貢献は、この既知の問題が、独立発達後の個人特異的な表象忠実性の内部へ、どこで入り込むかを示すことである。

3. 忠実性レジーム・シフト

3.1 忠実性には時間的に特定された target が必要である

本稿で扱う人物忠実性の主張は、target と関連する性質が指定されなければ不完全である。

説明のため、

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

を、デジタル・レプリカ D が、人間の source H に対し、性質の集合 X について、時間的 target t に忠実であるという主張とする。

これは同一性を数量化する理論ではなく、明示化のための記法である。X には、信念、自伝的記憶、言語スタイル、道徳判断、リスク選好、感情傾向などが含まれ得る。

D が、時点 t_0 までに利用可能な H の記録から作られたとする。

Type I — 歴史的忠実性

D は t_0 における H を表象しているか。

比較対象は、

$$H_{t_0}.$$

である。関連特徴は不完全にしか分からないかもしれないが、その少なくとも一部は source 記録、判断、行動に経験的根拠を持つ。

後の時点 t_1 では、次を問える。

Type II — 同時点類似性

t_1 における D は、現実の t_1 における H と似ているか。

比較対象は、

$$H_{t_1}.$$

である。この関係も経験的評価は可能だが、心理的性質は推論を必要とする場合がある。

さらに強い主張がある。

Type III — 反実仮想的発達帰属

D は、発達上対応する条件のもとで H が発達したであろう仕方で発達したか。

ここで比較対象は、

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

となる。$H^*_{t_1}$ は、X に関連する点で D が経験したものに対応する発達史を H が経験していたなら、H が到達したであろう状態を表す。

H_{t_0} や現実の H_{t_1} と異なり、この比較対象は実現していない。

H^* を用いるために D と H を数的に同一とみなす必要はない。同じ source 人物が別の条件のもとでどうなったかを問うだけである。D と H の形而上学的関係は未決のままでよい。

Type III は、忠実性という語のあらゆる用法に含まれるわけではない。システムが単に H を保存する、あるいは現在の H に似ていると主張するのではなく、**自ら発達することを通じてなお忠実である**と主張するときに問題になる。

3.2 material divergence は claim-relative である

D を模式的に、

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

と表す。ここで E_D は、作成後の関連する相互作用、入力、記憶操作、学習イベント、その他の発達要因を表す。

同様に、

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

とする。

単に、

$$E_D \neq E_H$$

であるからといって、あらゆる忠実性主張について意味のある転換が起きるわけではない。

差異は X に対して material でなければならない。

作成後の差異が **material** であるとは、その忠実性主張が前提とする発達モデルのもとで、その差異が評価対象 X に合理的に影響し得ることをいう。

ある午後の天候が違っても、安定した言語習慣には関係しないかもしれない。一方、十年間にわたる異なる人間関係、喪失、政治的出来事、職業経験は、後の価値観や判断に大きく影響し得る。

したがって materiality は普遍的な数値閾値ではなく、主張に相対的である。

ただし Type III の主張が証拠的な力を持つためには、materiality を決める発達モデルを結果観察後に選んではならない。主張者は、どの作成後差異が X に影響すると予想されるのか、そしてなぜかを事前に特定すべきである。そうしなければ、不一致も一致も事後的に都合よく再記述できる。事後的な特定は主張を偽にするわけではないが、その検証の証拠力を弱める。

3.3 developmental correspondence は主張成立の条件である

人間の経験と機械への入力を文字どおり同一視することはできない。人間の死別、疾病、身体的危険、加齢、親になることは、その出来事を説明する文章を AI へ与えるだけで自動的に再現されるわけではない。

したがって Type III は、D の履歴の関連部分と、H に仮定する条件との間に何らかの擁護可能な**発達上の対応関係（developmental correspondence）**があることを前提とする。

この対応は完全である必要はない。しかし主張内容が決定可能になる程度には特定されなければならない。X について擁護可能な対応を示せないなら、Type III 主張は単に証拠が弱いのではなく、under-specified である。つまり、どの H の反実仮想的軌跡を主張しているのかまだ定まっていない。

したがって $H^{*}_{t_1}$ は無制約な想像上の H ではなく、提示されている発達主張に相対して定義される。

3.4 欠けている比較対象

次を考える。

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

この差は経験的に確認できるかもしれない。しかし、それだけでは差が複製の失敗によるのか、material に異なる発達によるのかは分からない。

Type III に必要な比較はむしろ、

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

である。

現実の H が経験したのは E_H であり、Type III 主張が想定する発達上対応した履歴ではない。最も強い発達的比較対象は、現実には実現しなかった。

反実仮想的な未観察性それ自体は新しくない。重要なのは、それが表象関係のどこへ入り込むかである。t_0では「D は H に忠実である」という文は、source に対する経験的に裏づけられた関係を主として意味し得る。しかし material に重要な分岐の後、「H ならこう発達したはずの仕方で D も発達したから忠実である」というより強い主張は、H に未実現の発達軌跡を帰属する。

したがって同じ「忠実性」という語が、証拠上の比較対象の変化を隠し得る。

> **忠実性レジーム・シフト原則。特定人物を再現するデジタル・レプリカが、その本人とは異なる実質的に独立した発達史を獲得した場合、「本人への忠実性」という無限定な主張は、もはや単一の証拠負担を持たない。歴史的忠実性は本人の観察済み特徴との比較で評価可能であり、同時点類似性も観察可能だが、異なる経験によって交絡する。さらに「本人ならこのように発達した」という発達的忠実性の主張には、その特定人物の実現しなかった軌跡に関する反実仮想的帰属が必要となる。**

論証を最小限に述べると次の通りである。

P1. 人物忠実性の主張には、関連する source 状態と性質の指定が必要である。

P2. 歴史的忠実性は、実在した source 状態に関する証拠によって経験的に裏づけられる。

P3. 持続的で適応的なレプリカは、作成後の発達によって後年の状態を部分的に形成し得る。

P4. source と replica の履歴が、評価対象の性質に material な仕方で異なるとき、現時点の現実の H は D に対する matched developmental comparator ではない。

P5. したがって、D が発達上対応する条件のもとで H が変化したであろう仕方で変化したという主張には、特定された対応関係に相対して定義された反実仮想的比較対象 H^* が必要である。

P6. 実現しなかった反実仮想的軌跡と、経験的に裏づけられた source 状態は、異なる証拠上の地位を持つ。

したがって、

> **十分に material な経験的分岐の後では、「その人物への継続的忠実性」という無限定な概念を分解しなければならない。**

このレジーム・シフトは、すべての比較対象を H^* に置き換えるものではない。「fidelity」に圧縮されていた異なる主張が、異なる比較対象を必要とすることを明らかにするものである。

4. 不一致だけでは失敗を診断できない

source と replica が作成後に material に異なる履歴を持つと、不一致は曖昧な信号になる。

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

一つの説明は複製の失敗である。D が安定しているべき特徴からドリフトした、無関係な入力を過大評価した、更新に失敗した可能性がある。

別の説明は、正当な経験差である。D と H は異なる発達史を経験したから異なるのかもしれない。

4.1 間違った control

D が十年間、H が一度も出会わない community と交流し、H が受け取らない情報を得て、H が築かない関係を築き、H が行わない判断から学んだとする。

t_1 で、D が H の否定する立場を支持した。

現実の H_{t_1} を見れば、H と D が今は異なることは分かる。しかし追加仮定なしには、なぜ異なるのかは分からない。問題となる主張が Type III なら、比較すべき相手は $H^*_{t_1}$ である。

現時点の H は、レプリカの source であると同時に、発達的検証においては**違う履歴に曝露された control** である。

これは現時点の H が無意味ということではない。異なる経験に影響されにくいと理論的に考えられる安定した特徴について大きな不一致があれば、モデル失敗の強い証拠になり得る。

しかし差それ自体は、その原因を自動的に同定しない。

4.2 発達変数としての記憶

記憶 policy はこの問題を具体化する。

人間の自伝的記憶は受動的な archive ではなく、選択的かつ再構成的である。また外部の記憶技術も、自伝的 identity が維持・再構成される仕方に参加し得る（Burkell, 2016; Heersmink, 2022）。近年は AI を単なる storage ではなく、personal memory の active curator あるいは co-narrator として捉える研究もある（Smart et al., 2026; Telakivi, 2026）。

同じ source から三つのレプリカを作ったとしよう。一つはほぼすべての相互作用を永久保存する。別の一つは選択的に忘れる。第三の一つは、過去情報を繰り返し要約し再構成する。

数年後、異なる保持・検索・再構成 policy は信念や判断に影響し得る。完全なデジタル archive は歴史情報を極めてよく保存する一方、人間の通常の自伝的記憶とは異なる発達過程を生み得る。選択的記憶 architecture は、保存する歴史情報は少なくても、人間の認知機能により近い側面を持つかもしれない。

ここで重要なのは正しい記憶 architecture を決めることではない。記憶 policy 自体が、レプリカの作成後の発達史を構成するという点である。

4.3 不一致は証拠ではあるが診断ではない

この framework は発達的忠実性を反証不能にするものではない。

大きく、系統的で、理論上予期されない不一致は失敗の強い証拠になり得る。反復観察、個人内の規則性、心理学理論、集団データ、mechanistic knowledge は、もっともらしい発達軌跡を制約できる。

より狭い主張は、

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

から、

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

が直接には導かれない、ということである。

この推論には、どの履歴差が重要か、H がそれにどう反応するかという仮定が必要である。

5. 類似だけでは発達を検証できない

逆に、

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

とする。

H と D が似た価値を表明し、似た判断をするなら、それは同時点類似性の真正な証拠である。

しかしそれだけでは、D が D の履歴を通じて、H が発達したであろう仕方で発達したことを立証できない。

5.1 同じ終点への複数経路

類似は複数の mechanism から生じ得る。

Origin anchoring. D が t_0 の状態に強く固定され、後年経験の影響をほとんど受けない。

Constrained adaptation. D は更新するが、認識可能な特徴を保存するよう設計された境界内でしか変化しない。

Convergence. H と D が異なる発達経路を通って同じ立場へ到達する。

Coincidence. underlying process が重要に異なるのに、似た出力が偶然生じる。

したがって「発達するよう設計されたシステム」が、**発達しなさすぎたために** H と似たままでいる可能性がある。

ここから、単なる underdetermination より強い含意が出る。Type III が前提とする発達モデルが、特定種類の経験への曝露によって性質 X が方向 Δ へ変化すると予測しているとしよう。D がその条件を経験したのに、条件を経験していない現実の H に近いままであるなら、その類似は単に non-diagnostic なのではない。その発達主張に照らして anomalous であり、反証側の証拠になり得る。つまり、ある仕様のもとでは、source への類似そのものが発達的忠実性に対する disconfirming evidence になり得る。

検証の符号は反転し得る。保存志向のレプリカにとって最強に見える基準——source との継続的類似——が、発達志向のレプリカにとっては「自分が掲げる Type III の予測どおりに発達しなかった」証拠になり得る。 ただし、この反証的解釈は、H の現実の履歴についても同じ X の終点が予測されていないことを前提とする。現実の履歴から同じ終点が予測される場合には、収束の可能性によって類似性は再び non-diagnostic となる。

終点の類似は、そこに至った経路を同定しない。

5.2 人間らしい発達と個人特異的発達は異なる

Wang et al. (2026) が示す区別はここで重要になる。縦断的な人間研究と同じ方向に人格が変化するモデルは、人間らしい発達妥当性を示し得る。

しかし集団データは、特定の H の軌跡を一意に定めない。ある人は集団平均から大きく外れても、それだけでその人でなくなるわけではない。

集団レベルの発達データは $H^*_{t_1}$ の推定範囲を制約できる。しかしそれを観察することの代わりにはならない。

したがって、同時点類似性、集団レベルの発達妥当性、個人特異的な反実仮想的発達忠実性は区別されるべきである。関連はするが、交換可能な証拠ではない。

6. 忠実性から帰属へ

次の五つの文を比べよう。

1. Alex は X と言った。
2. Alex は X と信じている。
3. Alex なら X と言っただろう。
4. Alex を基にしたモデルは、Alex なら X と言うと予測する。
5. t_0 の Alex から作られ、その後独自の履歴を通じて発達したモデルが、現在 X と言っている。

これらは異なる認識論的地位を持つ。

第一は観察済み出来事の報告である。第二は現在状態の帰属である。第三は反実仮想的帰属である。第四は、その反実仮想を明示的に model prediction と位置づける。第五は、歴史的には本人から派生したが、その後発達したシステム自身の判断を報告する。

持続的 personal AI では、この区別が容易に崩れる。

6.1 予測から発達した判断へ

既存研究はすでに、replica 出力を originating human へ誤って帰属する危険（Karpus & Strasser, 2025）と、自律 proxy が表象対象人物の判断を推定するときの認識論的問題（Sweeney, 2023）を認識している。

t_0 では、レプリカの出力は主として H 由来情報を反映しているかもしれない。t_1 では、source 由来特徴だけでなく、後の model update、D だけが経験した相互作用や情報、記憶 policy、D 自身の過去判断も反映し得る。

その出力は source から因果的には派生している。

しかし因果的祖先関係は、その出力を source 本人の現在判断にはしない。

その出力を「Alex が考えていること」と呼べば provenance を隠す。「Alex なら考えること」と呼べば追加の反実仮想主張になる。システムそのものを単に「Alex」と呼べば両方を隠し得る。

6.2 帰属が生む権威

source との関係は authority を与え得る。

generic AI が「X が正しい判断だ」と言うのと、「Alex が X が正しい判断だと言っている」と提示されるのとでは地位が異なる。後者は Alex の評判、専門性、個人的関係、情緒的重要性、道徳的立場を利用し得る。

これは著者性、公的コミュニケーション、endorsement、家族の意思決定、surrogate decision-making、死後表象に関係する。

人物特異的 preference predictor は、自らの推論構造を比較的明示している。すなわち、患者が何を選ぶかを推定する（Earp et al., 2024）。持続的レプリカは異なる。predictor 自身が独立した履歴を通じて変化している可能性があるからである。

D が 50 歳の H から作られ、その後 H が死亡したとしよう。作成から 20 年後、D が H の経験しなかった判断について推奨を出す。

次を比べる。

> H ならこれを選んだだろう。
>
> D は、H ならこれを選んだだろうと予測している。
>
> H のレプリカとして始まり、その後独自に発達した D が、現在これを選んでいる。

第三は第一や第二を含意しない。

したがって持続的 personal AI は、personal identity の形而上学を解決する以前に、帰属問題を鋭くする。

6.3 authority は一方向には流れない

十分に分岐した D を新しい entity として扱うことは一貫している。

しかしそうするなら帰属にも帰結がある。D が H と異なってよい理由として developmental independence を持ち出しながら、H 由来の authority が有利なときだけ H との関係を無限定に保持することはできない。

independence によって不一致を正当化しながら、ancestry だけで帰属上の権威を受け取ることはできない。

これは前節の認識論的議論から直接演繹される結論というより、主張の整合性に対する制約である。

分岐は両方向で見えるままでなければならない。

7. 生きる船：アルカディア号の思考実験

松本零士作品におけるトチローとアルカディア号の関係に着想を得て、二つのレプリカを考える。ここでは作品の正典的解釈についての主張は必要ない。

トチローが死ぬ直前に、同程度に正確な二つのレプリカが作られたとする。

Ship A は保存志向である。記憶、価値、特徴的判断は source 状態の近くに保たれる。50 年後でも、「t_0 のトチローなら何と言っただろうか」という問いに極めてよく答えるかもしれない。歴史的忠実性は高い。

Ship B は発達志向である。乗組員と旅し、人間関係を築き、失敗や喪失に遭遇し、判断から学び、過去記憶の重要性を変える。50 年後、歴史上のトチローなら反対したであろう判断を下す。

その不一致だけで Ship B の失敗とは言えない。変化することが設計目的の一部だったからである。しかし発達の豊かさだけで、歴史上のトチローが Ship B になったはずだとも言えない。必要なトチローは Ship B の履歴を実際には生きていない。

したがって Ship A と Ship B は、共通 source との異なる関係を最適化している。Ship A は観察済みの過去人物についてよりよい証拠を与えるかもしれない。Ship B はより豊かな発達 branch かもしれない。どちらの事実も、Ship B の後年判断をトチローへ帰属できるかを決めない。

有用な問いは、

> どちらの船が本物のトチローか。

ではない。

> それぞれの船が現在何になっているかを踏まえ、トチローについてどの主張が正当化されるのか。

である。

8. 認識論的 governance：証拠上の地位を可視化し続ける

既存研究は、デジタル・レプリカと posthumous AI について、同意、透明性、識別可能性、control、自律的進化への制約を含む幅広い governance 要件を提案している（Hollanek & Nowaczyk-Basińska, 2024; Karpus & Strasser, 2025; Spitale & Germani, 2026）。Danaher and Nyholm (2025) の minimally viable permissibility principle も、digital duplicate と第三者との相互作用における透明性を含む。

本稿はそれらを置き換えない。透明性の一側面を精密化する。AI duplicate と相互作用していると知っているだけでは、ある出力が source からの検索なのか、不完全記録からの再構成なのか、source についての予測なのか、反実仮想的発達シミュレーションなのか、それとも独立発達した replica の現在判断なのかは分からない。

したがって必要なのはより狭く具体的な要件である。これらの出力類型が source との異なる証拠関係を持つなら、持続的 personal AI はその違いを回収不能にしてはならない。

8.1 Persistent provenance

最低限、適切な provenance は、source 人物、時間的 source 状態または source 期間、replica が基づく source material、生きている source との最後の意味ある同期、その後の学習が自律的だったか、作成後発達と記憶 policy の関連特徴、Type III 主張における materiality と correspondence を指定する developmental account、そして当該出力がどの認識論的 status で提示されるかを識別できる必要がある。

出力はたとえば次のように区別され得る。

Retrieval（検索） — H が実際にそう発言または記述した。

Reconstruction（再構成） — 不完全に記録された過去の立場をシステムが再構成する。

Prediction（予測） — 特定状況で H が何と言うかをシステムが推定する。

Counterfactual developmental simulation（反実仮想的発達シミュレーション） — H が実現しなかった発達史を経験したなら何になり、何を信じたかをシステムが推定する。

Autonomously evolved judgment（自律的に発達した判断） — 独自の後年履歴を経た D 自身の現在判断である。

これらの区別を毎文の前に煩雑な warning として表示する必要はない。しかし帰属が重要な場面では、回収可能で理解可能であるべきである。

8.2 通常の AI disclosure では足りない理由

「AI-generated」という label は、機械出力と直接の人間出力を区別する。しかし 2026 年の H の再構成、2036 年の H についての予測、10 年間独立発達した replica の自律判断を区別しない。

同様に「これは Alex の simulation です」という説明も、どの時点の Alex なのか、同期はいつ終わったか、その後どんな独立発達があったか、その出力は retrieval か prediction か simulation か evolved judgment かを答えない。

ここでの透明性問題は技術的であるだけでなく、時間的・関係的である。

設計上の問題は、developmental independence が増すのに、representational presentation が変わらないときに起こる。システムは H の名前、顔、声、biography、一人称を使い続けながら、H へ直接帰属できる程度を徐々に失い得る。

phenomenological continuity は epistemic distance を隠し得る。

目的は分岐を禁止することではない。分岐を見えるままにすることである。

9. 反論

9.1 これは単なる branching ではないか

branching は前提であって貢献ではない。Deguchi et al. (2022) は、digital twin が original と過去を共有し、その後独自の人生を送る model をすでに示している。

本稿の主張はその後から始まる。branching model は二つの trajectory を示す。しかし一方がもう一方の人物に発達的に忠実であり続けると主張するとき、後のどの comparator が必要かは決めない。

9.2 これは因果推論の fundamental problem にすぎないのではないか

一般的な counterfactual problem は既知である（Holland, 1986; Freeman et al., 2026; Laudy, 2026）。本稿の貢献はその新しい解法ではない。未実現の trajectory が、継続的な個人特異的忠実性主張に必要な comparator になったとき何が起こるかを特定することである。Type III では、反実仮想推論は developmental fidelity そのものが主張する内容の一部になる。

9.3 人間も substituted judgment をしている

人間は常に「その人なら何を望んだか」を問う。AI が反実仮想的帰属を発明したわけではない。

「彼女なら X を望んだだろう」という文は、その反実仮想的 grammar を明示している。これに対して非常にリアルな personal AI は、source 本人の名前、声、外見、biography、一人称で類似推論を提示し得る。さらに持続的 replica では、その発言を生成する state 自体が独自の履歴によって部分的に形成されている可能性がある。

そのため counterfactual prediction が、phenomenologically には継続観察のように見え得る。

9.4 十分に正確な model なら問題を解けるのではないか

より良い model は、縦断観察、causal model、source-specific behavior、心理学的証拠を通じて Type III 主張を強められる。

しかし prediction strength と observational status は異なる。どれほど優れた $H^*_{t_1}$ の推定も、実現しなかった trajectory の推定であることには変わりない。

9.5 D を新しい entity として扱えばよい

それは一貫した解決である。しかし D を本当に新しい entity として扱うなら、その後年判断を H へ無限定に帰属すべきではない。

歴史的 ancestry は真であり重要であり得る。しかし独立発達した判断を無限定に H へ帰属する license にはならない。

9.6 人間と機械の experience は同等ではない

この framework は literal equivalence を要求しない。Type III が要求するのは、その主張が前提とする developmental correspondence を指定することである。

擁護可能な対応を示せないなら、適切な結論は「D は不忠実に発達した」ではなく、「Type III 主張の内容が十分に特定されていない」である。

9.7 人間自身が予測不能である

完全な developmental prediction は必要ない。

reconstruction、prediction、counterfactual developmental simulation、independently evolved judgment はいずれも不確実であり得る。しかし不確実性が、それぞれの evidential target を一つに潰すわけではない。したがって異なる種類の主張である。

9.8 「Fidelity Regime Shift」は単なる新名称ではないか

用語そのものはなくてもよい。しかし区別そのものはなくならない。「H への fidelity」は、観察された過去の H、現実の現在 H、異なる発達史のもとでの未実現の H を参照し得る。これらの主張は単に accuracy の程度が違うのではなく、comparator と supporting evidence が違う。「regime shift」はその証拠構造の変化に付けた短縮名である。

9.9 結局 personal identity の metaphysics を前提としているのではないか

H^* という記法は、異なる発達史のもとで H がどうなったかを問う。D と H が数的に同一であることは要求しない。たとえば患者が実際には経験しなかった条件のもとで何を望んだかを問うように、source 人物を固定したまま状況を変える通常の counterfactual practice だけを必要とする。

したがって Type III は、D と H が一人の人物なのか、二人なのか、他の identity relation にあるのかについて結論を必要としない。H^* は H の未実現発達を比較するための comparator であり、D についての identity claim ではない。

10. 結論：分岐の後で

持続的デジタル・レプリカは、現実に存在した人物の過去の証拠から構築されるため、作成時には source との比較的強い証拠関係を持つ。独立発達はその関係を変える。

歴史的忠実性は source record に対して評価でき、同時点類似性は現実の source に対して評価できる。しかし、レプリカが**その特定 source 人物が発達したであろう仕方で発達した**というより強い命題は、実現しなかった trajectory への推論を必要とする。

現実の H からの不一致も、現実の H との類似も、それだけではこの主張を決めない。

この問題は、持続的 personal AI が学習するよう設計されるほど鋭くなる。保存は replica を観察済み source の近くに保つため比較を容易にする。発達は、後年状態を source からの直接的な経験的 anchor から遠ざけることで、replica の独立性を増し得る。

実務的帰結は帰属にある。後年の出力が retrieval、reconstruction、prediction、counterfactual developmental simulation、autonomously evolved judgment の間を黙って行き来してはならない。独立発達が蓄積するほど、時間的・発達的 provenance が重要になる。

この結論は、デジタル・レプリカが意識を持つか、personhood を持つか、数的同一性を保存するかに依存しない。

デジタル・レプリカは、観察済み人物の reconstruction として始まり得る。だが独自の履歴を通じて発達した後、その人物が何になったはずかを表象すると継続的に主張するには、別種の証拠的正当化が必要になる。

分岐の後では、provenance と attribution は一緒に移動しなければならない。

レプリカは、人間のように発達する能力を高めるほど、まさにそのために、ある人間の表象として検証することが難しくなり得る。

宣言

Funding

本稿の作成にあたり、資金、grant、その他の支援は受けていない。

Competing Interests

著者には、開示すべき関連する財務上または非財務上の利益相反はない。

Author Contributions

George Kujo が論証を着想し、文献レビューを行い、概念 framework を構築し、原稿を起草・改訂し、最終版を承認した。

Data Availability

該当なし。本研究では dataset を生成または解析していない。

生成 AI の使用

原稿作成過程において、文献探索、構造分析、言語編集、drafting の補助として生成 AI tool を使用した。著者は関連する文献、論証、解釈、文言を独立に評価・検証し、生成された内容を改訂したうえで、最終原稿の内容について全面的な責任を負う。

参考文献

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