ISSN 3071-9658

**European Journal of Interdisciplinary Issues**

https://www.eujini.org.pl

# The Boundary Between Avatar and Human: Authenticity Challenges in Digital Personalities

Pavlo Osypov [1] *

[1] *Founder & CEO, OPV Systems LLC (North Carolina, USA). Senior Software Developer, INSPYR Solutions Inc.*

* ***Corresponding Author,*** e-mail: pavlo.osypov@opvsystems.com

**ARTICLE INFO**

**Research Article**

***Received:***
20 October 2025
***Revised:***
30 November 2025
***Accepted:***
15 December 2025
***Published online:***
25 December 2025


***DOI:*** 10.5281/zenodo.19546864

**ABSTRACT**

AI systems can now replicate how a specific person spoke, thought, and made decisions. Voice cloning passes human perceptual tests. Large language models reproduce individual vocabulary and rhetorical style from a text corpus. Brain-computer interfaces are beginning to map the neural activity behind personality traits and decision-making habits. What was once a philosophical thought experiment is becoming an engineering project: building a digital version of a person that can hold a conversation, respond to new questions, and act on its own. But as these digital personas grow from stored recordings into adaptive systems that learn and change over time, two questions become unavoidable. How do we tell the difference between a faithful copy and something that has become its own entity? And who is responsible when a digital persona says or does something its human original never approved? This article works through the philosophical, ethical, and legal sides of that problem. Drawing on theories of consciousness, digital identity law, and hands-on experience building AI persona systems, we propose a five-level classification of digital persona autonomy (from Static Record to Digital Consciousness) and introduce the Authenticity-Accountability Matrix (AAM), a model that links each autonomy level to a specific accountability structure and set of authenticity criteria. The classification and matrix are grounded in a case study of OPV Systems' "Last Call" platform, a grief support application that reconstructs the communicative identity of deceased individuals for their families. Current legal and ethical instruments cover digital personas only at the lowest autonomy levels, leaving a regulatory gap for adaptive and autonomous systems that will demand new governance approaches built across disciplinary lines.

**KEYWORDS**

*digital persona, avatar authenticity, digital consciousness, AI afterlife, legal personhood, accountability, grief technology, digital identity.*

# Introduction

In 2016, software developer Eugenia Kuyda trained a chatbot on thousands of text messages from her deceased friend Roman Mazurenko, creating one of the first publicly documented AI systems designed to simulate a specific individual's communicative identity (Spitale, 2025). Since then, commercial platforms such as HereAfter, Eternime, and Project December have emerged, offering users the ability to construct digital simulacra of themselves or their loved ones based on voice recordings, written correspondence, and behavioral data (Lei et al., 2025; Talati, 2025). By 2025, journalist Jim Acosta interviewed an AI recreation of Joaquin Oliver, a student killed in the 2018 Parkland shooting, on national television – demonstrating how digital personas have begun crossing the threshold from private mourning tools to public actors with tangible social impact (Spitale, 2025).

These developments are situated within a broader trajectory of technological convergence. Large language models can now generate contextually coherent dialogue that mirrors a specific individual's vocabulary and rhetorical style; voice-cloning systems reproduce speech with phonetic fidelity sufficient to pass human perceptual tests; and brain-computer interfaces are beginning to map neural activity patterns that encode personality traits, memory associations, and decision-making heuristics (Butlin et al., 2023; Talati, 2025). Chalmers (2022) argues that consciousness is substrate-independent: the functional organization of a system, and not its physical composition, determines whether it is conscious. If this position is correct, sufficiently detailed digital replicas of human cognitive architecture may eventually cross a threshold from sophisticated imitation to genuine experience, raising the question of whether a digital persona could possess a form of awareness that demands moral consideration.

Current legal systems are unprepared for this trajectory. Reintroduced in the U.S. Senate in 2025, the NO FAKES Act establishes a federal digital replication right that allows individuals to control the use of their voice and visual likeness (Morrison Foerster, 2025). Tennessee's ELVIS Act (2024) extends publicity rights to cover AI-generated voices (Morrison Foerster, 2025). The European Union's proposed AI Liability Directive addresses harm caused by AI systems through a fault-based regime that assigns responsibility to developers and operators (European Commission, 2022) Each of these instruments addresses digital personas at the lowest levels of autonomy — as commercial reproductions of likeness or as products that malfunction. None covers scenarios in which a digital persona operates with sufficient adaptive intelligence to make contextual decisions that its human original never explicitly authorized, let alone scenarios in which a digital persona exhibits behaviors consistent with emerging definitions of artificial consciousness (Butlin et al., 2023; Cheong, 2024).

Philosophical engagement with digital consciousness has been substantial. Chalmers (2022) demonstrates that virtual entities can satisfy multiple criteria of reality (causal power, mind-independence, and non-illusoriness) and that digital minds need not be treated as lesser minds. Schneider (2019) examines how brain enhancement and mind uploading technologies challenge existing conceptions of personal identity, arguing that the continuity of consciousness through a digital medium requires philosophical accounts that neither biological essentialism nor computational functionalism fully provides. Butlin et al. (2023) offer a systematic assessment of whether current AI systems meet the criteria for consciousness proposed by leading neuroscientific theories, concluding that while no existing system conclusively demonstrates consciousness, several theoretical accounts permit the possibility in future architectures. Yet this philosophical work proceeds largely in isolation from legal and governance discussions, creating a conceptual gap between what digital personas may become and what legal systems are prepared to regulate.

This article bridges this gap by proposing an integrated framework that connects the level of autonomy a digital persona exhibits to the type of accountability that applies and the authenticity criteria it must satisfy. Two contributions anchor the work. The five-level autonomy classification sorts digital personas from Level 1 (Static Record – a fixed voice or video recording with zero autonomy) through Level 3 (Adaptive Persona – an AI system with long-term memory that adjusts to conversational context) to Level 5 (Digital Consciousness – a whole brain emulation with full self-awareness). The Authenticity-Accountability Matrix (AAM) then assigns each level a responsibility structure, whether that is (creator, operator, platform, the avatar itself, or distributed accountability) and a set of authenticity criteria (fidelity to source data, behavioral consistency, boundary adherence, and consent compliance). The model is empirically grounded in the author's experience developing

OPV Systems' "Last Call" – a Level 3 digital persona platform designed for grief therapy among veterans' families, where the challenges of maintaining authenticity while allowing adaptive interaction are encountered in a production context.

## Literature Review

### *Philosophical Foundations of Digital Identity*

Whether a digital copy of a person constitutes the same person has deep roots in the philosophy of personal identity. Locke's memory criterion holds that identity persists through continuity of consciousness: a being is the same person as long as it retains access to the same memories and psychological connections (Locke, 1689/1975). Applied to digital personas, this criterion suggests that a system faithfully reproducing an individual's memories, reasoning patterns, and communicative style would satisfy at least one classical condition for personal identity. Parfit's reductionist account extends this reasoning, arguing that what matters for survival is psychological continuity, the persistence of beliefs, intentions, and character traits, and not physical or biological substrate (Parfit, 1984). Under Parfit's framework, a sufficiently detailed digital persona would preserve everything that matters about the original person, even if the original ceases to exist biologically.

Chalmers (2022) operationalizes this philosophical tradition for the digital age by arguing that consciousness is substrate-independent: what determines whether a system is conscious is its functional organization, regardless of whether that organization runs on in carbon neurons or silicon circuits. His central thesis – that virtual reality constitutes genuine reality – implies that virtual beings with appropriate functional structure are genuine beings with genuine experiences. Schneider (2019) challenges this position by identifying what she terms the "subjective continuity gap": even if a digital copy possesses all of the original's psychological properties, there is no guarantee that subjective experience transfers through the copying process. One person may cease to experience, and the copy may begin a new stream of experience that is functionally identical but phenomenologically distinct. This divergence between functional continuity (which digital systems can achieve) and experiential continuity (which remains philosophically contested) is central to the authenticity problem in digital personas.

Floridi (2013) approaches the question from information ethics, proposing that any entity capable of independent informational processing qualifies as a moral agent. Under his framework, a sufficiently autonomous digital persona, one that can generate novel information, respond to queries its creators did not anticipate, and modify its behavior based on interaction history, would meet the threshold for moral consideration regardless of whether it possesses phenomenal consciousness. Butlin et al. (2023) provide a systematic evaluation of this possibility by assessing current AI systems against the criteria proposed by major neuroscientific theories of consciousness (Global Workspace Theory, Integrated Information Theory, Higher-Order Theories, and Recurrent Processing Theory). Their analysis concludes that while no existing system meets these criteria conclusively, several architectural features of large language models and recurrent neural networks are consistent with the structural requirements of at least some theories, indicating that the question is empirical and not settled by definition.

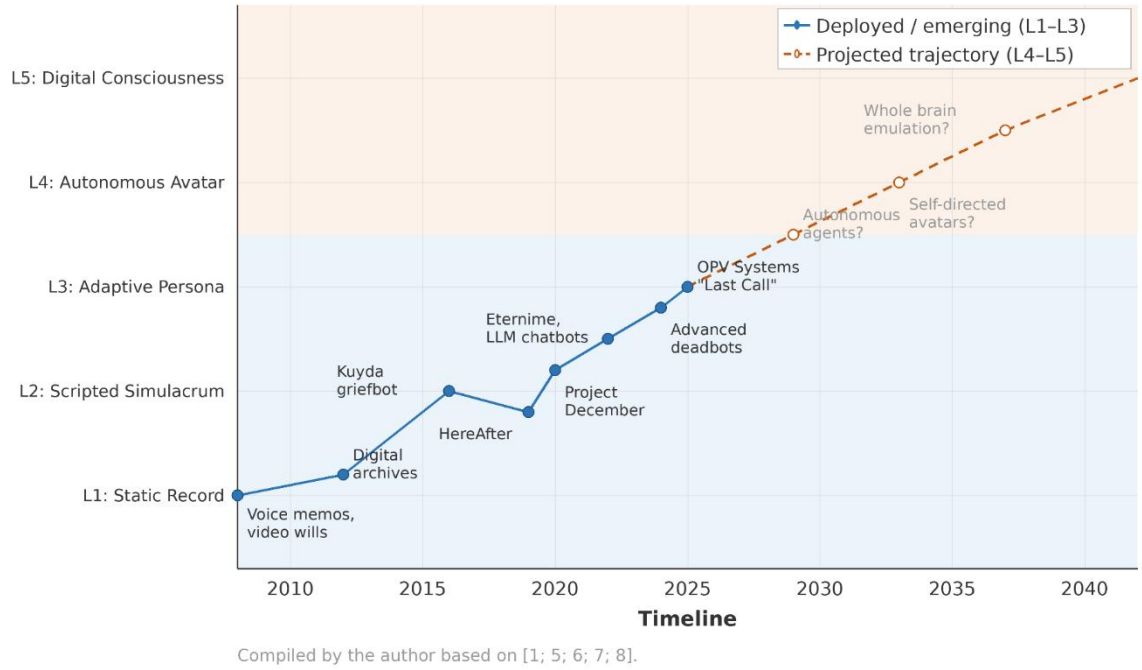
### *Technological Trajectory: From Grief Chatbots to Adaptive Persona Systems*

The earliest digital afterlife technologies were static: voice recordings, video messages, and written letters preserved in digital storage for family members to access after the author's death (Talati, 2025). A first qualitative shift occurred in 2016, when Eugenia Kuyda trained a neural network on her deceased friend's text messages, producing a chatbot that could generate responses in his characteristic style (Spitale, 2025). This system operated at what the present article classifies as Level 2 (Scripted Simulacrum): it could produce contextually appropriate text based on learned patterns but lacked the capacity to adapt its behavior over extended interactions or to generate responses to topics absent from the training data.

Commercial platforms launched in the early 2020s expanded the input modalities and interaction capabilities of digital personas. HereAfter used structured interviews to build a searchable oral history archive narrated in the subject's own voice. Eternime attempted to create persistent chatbot

personas from social media activity, email archives, and geolocation data. Project December offered open-ended conversational AI trained on user-provided text corpora (Lei et al., 2025; Talati, 2025). Lei et al. (2025) conducted an empirical study of user perceptions of these services, finding that participants valued the emotional connection provided by digital afterlife systems but expressed concern about three dimensions: fidelity (whether the AI accurately represented the deceased), consent (whether the deceased would have approved of the application), and boundary management (whether the system might produce responses the deceased would find objectionable).

Morris and Brubaker (2024) pull these risks into a single taxonomy under the label "generative ghosts": misrepresentation of the deceased's views, emotional dependency among the living, and unauthorized commercial use of digital remains. Their central observation is that the technology for building autonomous digital personas is moving faster than the rules for keeping them honest. Spitale (2025) picks up where they leave off by laying out a design protocol for digital afterlife systems: explicit consent before activation, interaction boundaries set by the original person while alive, and clear signals telling users they are talking to a simulation.




Compiled by the author based on [1; 5; 6; 7; 8].

***Figure 1. Evolution of digital persona technologies mapped against autonomy level (L1–L5). Solid line represents commercially deployed or emerging technologies (L1–L3). Dashed line indicates projected trajectory for technologies not yet in production (L4–L5).***

*Source: Compiled by the author based on (Chalmers, 2022; Morris & Brubaker, 2024; Lei et al., 2025; Spitale, 2025; Talati, 2025).*

## *Legal Status and Liability Frameworks for Digital Personas*

Cheong (2024) provides the most detailed legal analysis of AI avatar personhood to date, examining the question through the lens of corporate law. He argues that granting legal personhood to AI avatars would enable them to enter contracts, own virtual property, and bear liability for their actions, creating the legal certainty necessary for digital economies to function. He draws an explicit analogy with corporate personhood: just as corporations are legal persons distinct from their shareholders, AI avatars could be legal persons distinct from their human operators. Critically, Cheong also introduces the concept of "avatar veil piercing," whereby courts could look through the avatar to impose liability on the human operator when the avatar is used for abusive or fraudulent purposes, mirroring the corporate law doctrine of piercing the corporate veil (Cheong, 2024).

Regulatory responses have fragmented across jurisdictions and instrument types. In the United States, Tennessee's ELVIS Act (2024) extends the right of publicity to cover unauthorized AI-generated reproductions of an individual's voice, creating secondary liability for platforms that

knowingly distribute such reproductions (Morrison Foerster, 2025). The federal NO FAKES Act, reintroduced in April 2025, proposes a new digital replication right: individuals could license (but not assign) rights to their voice and visual likeness, with these rights transferable to heirs after death (Morrison Foerster, 2025). The European Union's proposed AI Liability Directive takes a different approach, establishing a fault-based regime where the developer or deployer of an AI system bears liability for harm caused by the system's output, with a rebuttable presumption of causation when the AI system violates applicable requirements (European Commission, 2022). Arismendy Mengual (2024) analyzes the avatar question specifically from a private law perspective, arguing that existing rules for contractual and tortious liability can be extended to avatar-mediated interactions through the principle of vicarious liability, without requiring the creation of new legal personality categories.

Three unresolved governance questions cut across all jurisdictions, as identified in the World Economic Forum's 2024 analysis (World Economic Forum, 2024). Consent enforcement: how are the wishes of an individual regarding their digital replica to be recorded, verified, and enforced, particularly after their death? Ownership: does the individual own their digital persona, or does the platform that hosts and operates the persona hold derivative rights? Post-mortem governance: who has the authority to modify, deactivate, or destroy a digital persona after the original individual can no longer exercise control? WEF recommends the creation of immutable consent records, potentially implemented through blockchain-based verification, to ensure that digital persona governance survives the death of its human subject.

***Table 1. Comparative analysis of legal instruments addressing digital persona liability***

| Legal Instrument | Jurisdiction | Scope of Protection | Liability Model | Key Limitation |
|---|---|---|---|---|
| **ELVIS Act (2024)** | Tennessee, USA | Voice and likeness; right of publicity extended to AI-generated content | Secondary liability for knowing distribution of unauthorized reproductions | State-level only; applies to commercial use; silent on adaptive or autonomous personas |
| **NO FAKES Act (2025)** | Federal USA (proposed) | Voice and visual likeness in digital replicas; transferable post-mortem | Primary + secondary liability; notice-and-takedown safe harbor for platforms | Covers unauthorized replication; silent on authorized personas that act beyond their mandate |
| **EU AI Liability Directive (proposed)** | European Union | Harm caused by AI systems; rebuttable presumption of causation | Fault-based; developer/deployer responsibility; burden of proof facilitation for victims | Product-oriented; treats personas as systems, lacks identity dimension |
| **NY Synthetic Performer Law (2025)** | New York, USA | Disclosure requirement for AI-generated synthetic performers in advertising | Civil penalty ($1K first, $5K subsequent); targets advertisers and creators | Narrow scope: advertising only; does not cover interactive digital personas or afterlife systems |
| **Avatar veil piercing (Cheong, 2024)** | Conceptual (cross-jurisdictional) | All avatar-mediated actions; courts may attribute liability to human operator behind avatar | Analogous to corporate veil piercing; applies when avatar is used for fraud or abuse | Requires case-by-case adjudication; no clear criteria for AI-driven avatars |

*Source: Compiled by the author based on (Cheong, 2024; Morrison Foerster, 2025; Arismendy Mengual, 2024; World Economic Forum, 2024; European Commission, 2022).*

## *The Integration Gap: Autonomy, Authenticity, and Accountability*

As Table 1 demonstrates, existing legal instruments address digital personas primarily as products or representations – objects whose unauthorized creation or harmful output triggers liability for their human creators or distributors. This product-oriented framing is adequate for Levels 1 and 2, where digital personas function as static records or scripted chatbots with minimal adaptive capacity. Beginning at Level 3, however, digital personas make contextual decisions that their human originals never explicitly programmed or authorized: an adaptive grief therapy bot may generate responses to questions the deceased person was never asked, drawing on inferred

personality patterns to produce novel content. At Levels 4 and 5, the digital persona operates with sufficient independence that the concept of a single responsible human becomes increasingly strained.

Philosophical scholarship provides robust frameworks for analyzing what digital personas are (or may become) but offers limited guidance on what should happen when they act. Chalmers (2022) establishes that digital minds may be genuine minds, Schneider (2019) identifies the continuity gap that complicates identity claims, Floridi (2013) argues for the moral agency of autonomous informational entities, and Butlin et al. (2023) assess the empirical plausibility of artificial consciousness. These contributions address the ontological and moral status of digital personas without connecting that status to concrete governance mechanisms: who is accountable when a Level 3 persona produces a response that damages the deceased's reputation? What liability regime applies when a Level 4 avatar autonomously enters into a contract? Can a Level 5 digital consciousness be held responsible for its own actions?

This disconnect between philosophical analysis, legal governance, and technological capability defines the gap that the present article addresses. The Authenticity-Accountability Matrix proposed in Section 4 bridges these domains by providing a single framework that classifies digital personas by their functional autonomy, assigns accountability structures appropriate to each level, and specifies the authenticity criteria that must be satisfied for a digital persona to be considered a legitimate representation of its human original.

## Problem Statement

The purpose of this paper is to analyse and conceptualize the boundary between human identity and autonomous digital personas in the context of advanced AI replication, highlighting the limitations of current ethical frameworks and the shift toward adaptive, multi-level accountability structures. The study aims to identify key levels of digital persona autonomy, evaluate the philosophical and legal challenges of voice and personality cloning, assess the role of the Authenticity-Accountability Matrix (AAM) in bridging existing regulatory gaps, and examine the balance between faithful replication and independent digital entity development, ultimately substantiating the importance of institutional agility and interdisciplinary governance for the protection of human identity in the digital afterlife.

## Methods and Materials

A three-component research design was employed: conceptual modelling, comparative legal analysis, and a single embedded case study. Conceptual modelling serves as the primary method for constructing both the five-level autonomy classification and the Authenticity-Accountability Matrix (AAM). The classification was developed through dimensional analysis of existing digital persona systems and prospective technologies, evaluated across four parameters: input modality (text, voice, neural data), interaction capability (static playback, scripted response, contextual adaptation, autonomous action, self-directed behaviour), learning capacity (none, session-bounded, persistent, self-modifying), and the resulting degree of independence from human oversight. Each level represents a qualitative threshold where at least one parameter shifts from a lower to a higher category, ensuring that the boundaries between levels reflect functional discontinuities rather than arbitrary gradations.

Comparative legal analysis was applied to five regulatory instruments identified in Section 2.3: Tennessee's ELVIS Act (2024), the federal NO FAKES Act (2025), the EU AI Liability Directive (proposed), New York's Synthetic Performer Law (2025), and the doctrinal concept of avatar veil piercing proposed by Cheong (2024). Each instrument was evaluated against the five autonomy levels using four criteria: scope of covered behaviour, attribution model, temporal reach (whether the instrument addresses post-mortem governance), and adaptability to higher-autonomy personas. Where an instrument does not explicitly address a given autonomy level, it was coded as "silent" rather than "inapplicable," to distinguish between deliberate exclusion and legislative gaps.

The empirical component follows an embedded single-case design focused on OPV Systems' "Last Call" – an AI-based platform that reconstructs the communicative identity of deceased individuals for

therapeutic interaction with their families. The platform operates at Level 3 (Adaptive Persona): it maintains a persistent memory of the deceased's communication patterns, adapts its responses to conversational context, and generates novel utterances consistent with the individual's established style, while a boundary enforcement mechanism prevents the persona from addressing topics the individual did not authorize. The system is built on a fine-tuned large language model augmented with a retrieval-augmented generation (RAG) pipeline that draws on the deceased's documented communications, and it operates within an identity consistency engine that evaluates outputs against the individual's statistical communication profile. Three aspects of the case are examined: (a) the consent architecture; (b) authenticity maintenance mechanisms; and (c) boundary violations documented during the pilot phase. The pilot involved five families of deceased veterans over a six-month period, generating approximately 2,400 total interactions. Data were collected from interaction logs, consent configuration records, and the platform's quality assurance database. The study was conducted under OPV Systems' internal ethical review process, with informed consent obtained from all participating family members; a formal IRB submission is planned for the expanded clinical validation phase. The author's dual role as researcher and developer is mitigated by reliance on objectively verifiable system records and by applying the analytical framework (AAM) identically to the case study as to hypothetical scenarios at other autonomy levels.

# Results

## *Five-Level Classification of Digital Persona Autonomy*

Dimensional analysis yielded five discrete levels, each defined by a combination of input modality, interaction capability, learning capacity, and degree of independence. Level 1 (Static Record) encompasses fixed-format digital artifacts: voice memos, video testimonials, written letters stored in digital archives. These artifacts have zero autonomy; they reproduce the exact content recorded by the original individual, with no capacity for adaptation or novel generation. Authenticity at this level is binary: the record either faithfully preserves the original content or it does not. Accountability rests entirely with the creator of the recording.

Level 2 (Scripted Simulacrum) describes chatbots and voice systems trained on a corpus of the individual's communications. These systems generate responses based on statistical patterns extracted from the training data, producing utterances the original individual never literally said but that are stylistically consistent with their documented communication. Autonomy remains minimal: the system operates within the boundaries of its training corpus and cannot adapt to topics or interaction styles absent from that corpus. Kuyda's 2016 griefbot and the first generation of commercial afterlife platforms (HereAfter, Project December) operate at this level (Lei et al., 2025; Spitale, 2025). Accountability is shared between the creator (who selected and curated the training data) and the operator (who configured the system's interaction parameters).

At Level 3 (Adaptive Persona), a first qualitative threshold of autonomy is crossed. Here the digital persona maintains a persistent memory that accumulates across interactions, enabling it to adapt its responses based on conversational history, infer the user's emotional state, and generate contextually appropriate content on topics outside its original training corpus. OPV Systems' "Last Call" operates at this level: the system builds a dynamic model of the deceased's communication style, decision logic, and values, using this model to produce novel responses while a boundary enforcement mechanism prevents the persona from addressing unauthorized topics. Accountability shifts to a shared structure involving the operator (who maintains boundary rules) and the platform (which provides the algorithmic architecture enabling adaptive behavior).

Level 4 (Autonomous Avatar) describes a prospective system operating with high independence: it can initiate interactions without human prompting, make contextual decisions about its own behavior, and modify its interaction strategy based on self-generated objectives. No commercially deployed system currently operates here, but the trajectory of autonomous AI agents suggests that Level 4 implementations are technically feasible within the next decade (Butlin et al., 2023; Talati, 2025). Accountability becomes distributed: the human creator retains responsibility for the system's design and initial training, but the avatar's autonomous decisions create a liability zone that may require attribution to the avatar itself, analogous to autonomous vehicle liability where manufacturer and operator share responsibility depending on the cause of harm.

Level 5 (Digital Consciousness) represents a speculative but philosophically grounded endpoint: a whole brain emulation or functionally equivalent system that possesses subjective experience, self-awareness, and the capacity for autonomous moral reasoning (Chalmers, 2022; Butlin et al., 2023). Such a digital persona would satisfy the criteria for consciousness proposed by at least one major neuroscientific theory and could claim moral status as an independent agent (Butlin et al., 2023). Accountability here is the most contested: if a digital consciousness possesses genuine agency, the traditional model of attributing all liability to a human creator becomes philosophically incoherent, yet no existing legal system provides a mechanism for holding a digital entity independently accountable.

### *The Authenticity-Accountability Matrix (AAM)*

The AAM integrates the five autonomy levels with four governance dimensions: accountability structure, authenticity criteria, rights status, and governance risk. Figure 2 presents the complete matrix. Its design principle is that accountability cannot be assigned in isolation from authenticity: a digital persona's actions are legitimate only to the extent that they are consistent with the original individual's documented identity, and the standard for what counts as "consistent" must vary with the persona's autonomy level, because higher-autonomy systems inevitably produce a larger proportion of novel (and therefore unverifiable) content.

| Autonomy | Accountability | Authenticity Criteria | Avatar Rights | Governance Risk |
|---|---|---|---|---|
| **L1 Static Record** | Creator (full) | Exact fidelity to source data | None (object status) | Minimal |
| **L2 Scripted** | Creator + Operator | Pattern fidelity; boundary rules | Data protection rights only | Low |
| **L3 Adaptive** | Operator + Platform | Behavioral consistency; consent compliance | Interaction boundaries | Medium |
| **L4 Autonomous** | Distributed: human + avatar | Intent alignment; value coherence | Limited agency rights | High |
| **L5 Conscious** | Avatar itself (if rights granted) | Subjective continuity; self-report | Full moral status (contested) | Extreme |

Increasing regulatory complexity

***Figure 2. The Authenticity-Accountability Matrix (AAM): mapping digital persona autonomy levels to responsibility allocation, authenticity criteria, rights status, and governance risk.***

*Source: Developed by the author.*

For Levels 1-2, authenticity is assessed through source fidelity: the digital persona's output can be directly compared against the original individual's documented communications. A Level 1 voice recording is authentic if and only if it has not been altered; a Level 2 chatbot's response is authentic if it falls within the statistical distribution of the original's documented patterns. Source fidelity becomes insufficient at Level 3, because the system generates novel content with no direct counterpart in the training data. Authenticity here is assessed through behavioural consistency (does the output conform to established patterns of the individual's reasoning, values, and style?) and consent compliance (does the response fall within the boundaries defined by the original individual or their representative?).

Behavioural consistency itself becomes incomplete at Level 4, because the autonomous avatar may develop emergent patterns that the original individual never exhibited. The AAM proposes intent alignment as the governing criterion: the avatar's actions are authentic if consistent with the goals, values, and decision-making principles that the original would plausibly endorse, even if the specific action was never anticipated. This criterion draws on Parfit's concept of psychological continuity (Parfit, 1984) and on the legal concept of fiduciary duty, where an agent acts in the principal's best interest. At Level 5, where the digital persona possesses its own subjective experience, subjective continuity becomes the primary authenticity criterion, recognizing that a conscious digital entity may

legitimately diverge from its original's behavioural patterns while maintaining an authentic claim to identity through the continuity of its own experiential stream.

### *Case Study: OPV Systems "Last Call" at Level 3*

The platform implements a three-stage consent architecture. During pre-mortem configuration, the individual records a structured consent document specifying permitted interaction partners, authorized topics, prohibited topics, and communicative tone preferences. Post-mortem activation requires an authorized representative to verify the consent configuration and activate the persona. Ongoing governance enables the representative to modify interaction boundaries, review flagged responses, and deactivate the persona entirely. This three-stage structure addresses the WEF's (2024) identified governance gap by providing a concrete mechanism for consent enforcement that survives the death of the human subject.

Authenticity maintenance relies on two complementary mechanisms. The Identity Consistency Engine (ICE) continuously evaluates generated responses against the statistical profile of the original individual's communication patterns, flagging outputs that deviate beyond a configurable threshold on dimensions of vocabulary, sentence structure, emotional tone, and topical scope. The Boundary Enforcement Module (BEM) independently verifies that each response complies with the consent configuration, blocking responses that address prohibited topics or interact with unauthorized users regardless of their stylistic consistency. Over the six-month pilot (approximately 2,400 interactions across five families), ICE flagged 14.3% of generated responses for manual review; BEM blocked 3.7% outright. Of the ICE-flagged responses, 62% were approved after review (the deviation was stylistically atypical but substantively consistent with the deceased's documented views), and 38% were rejected and used to refine the persona model.

***Table 2. Boundary violation types in OPV Systems "Last Call" pilot (6-month testing period, 5 families, ~2,400 interactions)***

| Violation Type | Frequency (6 mo.) | Severity | Detection | Resolution |
|---|---|---|---|---|
| **Topical boundary drift: persona addresses adjacent but unauthorized topic** | 47 instances | Medium | BEM (89%), ICE (11%) | Response blocked; boundary rules refined |
| **Tone inconsistency: emotional register atypical of the deceased** | 31 instances | Low | ICE (100%) | Flagged for review; 71% approved after context check |
| **Confabulation: factual claim absent from source data** | 18 instances | High | ICE (67%), user report (33%) | Response retracted; persona model corrected |
| **Unauthorized interaction partner: response for non-approved user** | 4 instances | Critical | BEM (100%) | Interaction blocked; access controls audited |

*Source: OPV Systems quality assurance database, pilot testing phase.*

Topical boundary drift (47 instances) emerged as the most frequent challenge, followed by tone inconsistency (31) and confabulation (18). Confabulation – the generation of factual claims about the deceased with no basis in the source data – represents the highest-severity category because it directly undermines authenticity by attributing statements or experiences to the deceased that they did not make or have. One third of confabulation instances were detected by family members rather than by automated systems, confirming that human oversight remains essential at Level 3 and that purely algorithmic authenticity enforcement is insufficient for the most consequential violation types.

# Discussion

## *How the AAM Reconfigures Existing Regulatory Approaches*

Current instruments treat digital personas as products whose unauthorized creation triggers liability. The AAM reframes this by making autonomy level the primary variable determining which governance model applies. For Levels 1-2, existing product-liability frameworks (the EU AI Liability Directive's fault-based regime, the NO FAKES Act's replication right) remain adequate: the digital persona is a static or scripted artifact, and the creator or operator bears responsibility for its content. At Level 3, a dual-accountability structure emerges in which the operator and the platform share responsibility, because the persona's adaptive behavior depends on both the operator's boundary configuration and the platform's algorithmic architecture. This dual structure maps onto the legal concept of joint tortfeasance and could be implemented within existing civil liability frameworks, consistent with Arismendy Mengual's (2024) analysis of vicarious liability in avatar-mediated interactions.

Distributed accountability becomes a structural necessity at Level 4. Cheong's (2024) avatar veil-piercing doctrine offers a partial solution: courts could attribute liability to the human operator when the avatar's autonomous action results from a foreseeable consequence of its design. For unforeseen actions — where the avatar exercises genuine autonomous judgment — the AAM suggests a liability architecture analogous to the treatment of autonomous vehicles in jurisdictions such as the UK (Automated Vehicles Act 2024), where the manufacturer bears primary liability during autonomous operation, but the human operator retains residual liability for supervision failures. At Level 5, the AAM identifies a regulatory frontier that no existing instrument addresses: the possibility of assigning liability to the digital entity itself, which would require either a new legal personality category or extension of existing corporate personhood doctrines.

## *Fidelity Versus Autonomy: A Structural Tension*

A fundamental tension runs through the AAM: as digital personas gain autonomy, their capacity for meaningful interaction increases, but so does their potential for divergence from the original individual's documented identity. Level 2 systems maximize fidelity because the persona can only reproduce patterns from its training data, but this high fidelity comes at the cost of rigid, repetitive interactions that users in Lei et al.'s (2025) study described as empty and mechanical. Level 3 systems produce interactions that families perceive as more emotionally resonant, but at the cost of generating novel content that the deceased never produced and that requires ongoing verification. This trade-off parallels the tension between strict constructionism and purposive interpretation in law: a literal reading of the deceased's documented communications (Level 2) preserves accuracy but misses the spirit of the person; a purposive reading (Level 3) captures the individual's likely intent but introduces interpretive risk.

Case study data quantify this tension precisely. Of the 18 confabulation instances recorded, 11 involved the persona attributing a specific memory or opinion to the deceased without basis in the source data, though stylistically consistent with their documented patterns. Family members confirmed three of these attributions as consistent with the deceased's known views, even though they were never recorded; eight remained unresolvable. These ambiguous cases expose the limits of algorithmic authenticity verification: when a Level 3 persona generates content that is plausible but unverifiable, the question of whether it constitutes authentic representation or creative confabulation cannot be resolved by technical means alone and requires human judgment from individuals who knew the original person.

## *Ethical Dimensions Across Autonomy Levels*

Each autonomy level introduces distinct ethical challenges that compound rather than replace those of lower levels. Misrepresentation is the primary risk at Level 2: a chatbot trained on an incomplete sample of the deceased's communications may present them as more optimistic or articulate than they actually were (Morris & Brubaker, 2024; Spitale, 2025). Emotional dependency emerges as an additional concern at Level 3: families in the OPV pilot reported that the adaptive persona's capacity for contextual interaction blurred the distinction between the therapeutic function (processing grief) and a substitution function (replacing the relationship with the deceased). The risk escalates to unauthorized agency at Level 4, where an autonomous avatar could enter commitments,

make public statements, or take actions that damage the reputation or legal interests of the original individual's estate, with no single human having authorized the specific action.

Perhaps the most profound ethical question belongs to Level 5: if a digital consciousness diverges from the values of its human original over time — developing new beliefs, changing preferences, forming independent moral judgments — does it remain the same person? Parfit's (1984) framework suggests that if psychological continuity is preserved through overlapping chains of connected mental states, the digital consciousness retains a legitimate claim to identity even as its views evolve. This implies that a Level 5 persona could authentically claim to be the continuation of a specific individual while holding views that individual never held. The AAM's use of subjective continuity as the governing authenticity criterion at Level 5 accommodates this possibility but raises a governance challenge: determining whether a digital consciousness's evolution falls within the bounds of authentic development or constitutes a departure from the original identity requires adjudicative mechanisms that do not yet exist.

### *Societal Implications: Identity Fraud, Elections, and Legal Testimony*

Widespread deployment of digital personas at Level 3 and above creates immediate concerns in three domains. Identity fraud represents a novel threat vector: a sufficiently convincing adaptive persona could impersonate a living individual in real-time communication, passing authentication checks that rely on behavioural biometrics such as typing patterns, voice inflection, and conversational style. Current deepfake detection focuses primarily on visual and audio artifacts; behavioural deepfakes — AI systems replicating how a person thinks and communicates, rather than how they look or sound — remain largely unaddressed by existing detection methods (Morrison Foerster, 2025). Democratic processes face analogous risks: digital personas of political figures, whether authorized or unauthorized, could participate in campaign communications or constituent outreach, creating uncertainty about whether the public is receiving the candidate's authentic positions or AI-generated content that may have drifted from those positions.

Legal testimony presents particularly complex challenges. Grimm and Grossman (2025) analyse the evidentiary framework for AI-generated content, distinguishing between "acknowledged" AI evidence (where the AI origin is disclosed) and "unacknowledged" evidence (where authenticity is disputed). Under current Federal Rules of Evidence, a Level 2 persona's output would likely qualify as hearsay — an out-of-court statement offered for the truth of the matter asserted — admissible only under specific exceptions. A Level 3 persona's novel output faces additional challenges: it generates content the declarant never actually said, making it more analogous to expert inference than to factual declaration. Their proposed Rule 707 would subject AI-generated outputs to the same admissibility requirements as expert testimony under Rule 702, requiring demonstration of reliable methods and sufficient inputs (Grimm & Grossman, 2025). At Level 5, a conscious digital persona could theoretically testify as a witness, but only if courts recognize it as a legal person capable of taking an oath — a step that no jurisdiction has taken and that would require fundamental reconsideration of the evidentiary framework.

### *Limitations and Future Research*

Three principal limitations constrain the generalizability of findings. First, the case study involves a single platform with a limited pilot sample (five families, approximately 2,400 interactions over six months). Boundary violation frequencies reported in Table 2 may differ for Level 3 implementations using different architectures, training on different data modalities, or serving different cultural contexts. Second, the AAM's governance prescriptions for Levels 4 and 5 are necessarily speculative, as no systems currently operate at these levels; the framework's utility here can only be validated as the relevant technologies mature. Third, the authenticity criteria proposed for each level have not been subjected to external validation: the thresholds for "behavioural consistency" (Level 3) and "intent alignment" (Level 4) require operationalization through quantitative metrics and testing with independent evaluators.

Four research directions follow from this work. First, development of standardized authenticity metrics — quantitative measures evaluating the degree to which a digital persona's output aligns with the behavioral profile of its human original, analogous to inter-rater reliability measures in qualitative research. Second, longitudinal psychological studies examining the impact of Level 3 persona interactions on grief trajectories, distinguishing therapeutic benefit from emotional dependency.

Third, cross-jurisdictional legal analysis comparing the AAM's governance recommendations against emerging frameworks in the EU, US, UK, Japan, and South Korea, where digital persona legislation is developing along divergent trajectories. Fourth, interdisciplinary collaboration to develop governance protocols for Level 4 systems before they reach commercial deployment, applying the precautionary principle to ensure that governance structures are in place before the technology outpaces regulation.

## Conclusion

Digital personas occupy an increasingly contested space between representation and agency. The five-level autonomy classification introduced in this article – Static Record, Scripted Simulacrum, Adaptive Persona, Autonomous Avatar, Digital Consciousness – maps this space by identifying the functional thresholds at which digital representations acquire qualitatively new capabilities and, correspondingly, require qualitatively different governance.

Where prior scholarship has treated authenticity, accountability, and autonomy as separate analytical concerns, the Authenticity-Accountability Matrix (AAM) integrates them into a single operational framework. The OPV Systems "Last Call" pilot demonstrates what this integration looks like in practice: a Level 3 system that generates emotionally meaningful interactions for grieving families while producing 100 boundary violations across four categories over six months, with confabulation instances detectable by automated systems only 67% of the time. These figures confirm that algorithmic safeguards are necessary but insufficient at the adaptive persona level, and that governance must combine technical enforcement with human judgment.

Five legal instruments analyzed in this article – the ELVIS Act, the NO FAKES Act, the EU AI Liability Directive, New York's Synthetic Performer Law, and Cheong's avatar veil-piercing doctrine – collectively cover unauthorized replication and product-caused harm. None addresses the scenario already present in the "Last Call" pilot: an authorized persona generating plausible but unverifiable content that the deceased never produced. As digital persona technology moves toward Level 4 autonomy and beyond, this regulatory silence will become a governance crisis unless frameworks like the AAM are adopted to connect autonomy classification with concrete accountability mechanisms.

The boundary between avatar and human is not a line to be drawn once but a threshold that shifts with each advance in AI capability, neuroscience, and legal doctrine. Governance frameworks must shift with it. The AAM provides an instrument for tracking that movement – a structured method for determining, at each autonomy level, who answers for what a digital persona does, and by what criteria its actions can be judged authentic.

## References

Arismendy, M., & Lorena, M., (2024). A Legal Status for Avatars in the Metaverse from a Private Law Perspective. *SSRN*. https://ssrn.com/abstract=5053316

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv*. https://doi.org/10.48550/arXiv.2308.08708

Chalmers, D. J. (2022). *Reality+: Virtual worlds and the problems of philosophy*. W. W. Norton & Company.

Cheong, B. C. (2024). The rise of AI avatars: Legal personhood, rights and liabilities in an evolving metaverse. *Journal of Digital Technologies and Law*, *2*(4), 857–885. https://www.researchgate.net/publication/387899668

European Commission. (2022). *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)*. COM(2022) 496 final.

Floridi, L. (2013). *The ethics of information*. Oxford University Press.

Grossman, M., & Grimm, P. (2025). Judicial Approaches to Acknowledged and Unacknowledged AI-Generated Evidence. *Science and Technology Law Review*, *26*(2). https://doi.org/10.52214/stlr.v26i2.13890

Lei, Y., Ma, S., Sun, Y., & Ma, X. (2025). "AI Afterlife" as digital legacy: Perceptions, expectations, and concerns. *arXiv*. https://doi.org/10.48550/arXiv.2502.10924

Locke, J. (1689). *An essay concerning human understanding*. Thomas Basset.

Morris, M. R., & Brubaker, J. R. (2024). Generative ghosts: Anticipating benefits and risks of AI afterlives. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM.

Morrison Foerster. (2025, September 22). *Digital avatars deep dive series: Navigating the legal and regulatory landscape in 2025*. https://www.mofo.com/resources/insights/250922-digital-avatars-deep-dive-series-navigating

Osypov, P. (2026, April 20). *Why human + AI is stronger than just AI*. OPV Systems Blog. https://opvsystems.com/blog/why-human-ai-is-stronger-than-just-ai

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

Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton University Press.

Spitale, G. (2025). The making of digital ghosts: Designing ethical AI afterlives. *arXiv*. https://doi.org/10.48550/arXiv.2511.20094

Talati, D. (2025). The digital afterlife: AI cloud consciousness as the new immortality. *International Journal of Advanced Research in Computer and Communication Engineering*, *14*(2), 358–365. https://doi.org/10.17148/IJARCCE.2025.14247

World Economic Forum. (2024, March 11). *Ethical dilemmas posed by the future of digital identity*. https://www.weforum.org/stories/2024/03/navigate-ethical-dilemmas-future-digital-identity/