



FAIR LABS

FAIR ARTIFICIAL INTELLIGENCE RESEARCH LABS

RESEARCH BRIEF ♦ EMBODIED AI ♦ 2025

Virtual Humans and the Ethics of Presence

Photorealistic volumetric avatars and AI-driven virtual humans raise questions of consent, identity, and deception – toward a framework for honest embodied representation.

PROGRAM

Embodied AI

DOCUMENT TYPE

Research Brief

AUDIENCE

Designers, Ethicists & Regulators

PUBLISHED

2025

CONTENTS

Table of Contents

—	Executive Summary	3
I	The Rise of Photoreal Virtual Humans	5
II	Presence – Why Embodiment Lands Differently	8
III	Consent & Likeness	10
IV	Deception & Disclosure	13
V	Identity, Representation & Bias	16
VI	Emotional Influence & Parasocial Dynamics	18
VII	A Framework for Honest Embodied Representation	20
VIII	Standards & Governance for Virtual Humans	23
—	Notes & References	26

EXECUTIVE SUMMARY

Presence is a power, and powers need rules

IN ONE PARAGRAPH

Virtual humans — photorealistic, real-time, generatively driven — recruit the social instincts people use with one another, and that recruitment is the whole point and the whole hazard. A convincing face changes how much we trust, how much we disclose, how easily we are persuaded, and how attached we become. This brief treats presence as a design material with moral consequences, and argues that honesty must be engineered into embodied systems the way safety is engineered into a vehicle: not bolted on after a harm, but built into how the thing works. We supply a taxonomy, a set of consent and disclosure requirements graduated by risk, and a governance path that designers, platforms, and regulators can walk in parallel.

4

commitments in the honest-representation framework: provenance, disclosure, consent, proportion

3

recurring problems: consent & likeness, deception & disclosure, identity & representation

5

virtual-human classes in the taxonomy, from stylized avatar to autonomous digital double

What we recommend

1**Disclose the machine, always and legibly.**

A person interacting with an AI-driven virtual human has a right to know it. Disclosure must be present at the moment of contact, understandable to a non-specialist, and impossible to mistake for a stylistic flourish. The more realistic the presence, the more prominent the disclosure.

2**Treat likeness as property held in trust.**

A face, a voice, and a manner belong to the person they came from. Building a digital double requires specific, revocable, purpose-bound consent — from the living, from the estates and communities of the dead, and never from the non-consenting at all.

3

Bind provenance to the pixels.

Synthetic presence should carry cryptographic provenance — who made it, from what, and under what permissions — using open content-authenticity standards, so that a virtual human's origin travels with it rather than being lost at the first screenshot.

4

Scale the guardrails to the stakes.

A stylized brand mascot and an autonomous avatar of a real clinician are not the same risk. Consent depth, disclosure prominence, and oversight should rise with realism, autonomy, and consequence, so that low-stakes delight is not smothered and high-stakes deception is not permitted.

The sections that follow develop each recommendation, with a taxonomy of virtual humans, tables of consent scenarios and disclosure requirements, and two diagrams — a presence–realism spectrum and a disclosure decision flow — that a design or policy team can lift directly into practice.

SECTION I

The Rise of Photoreal Virtual Humans

For most of computing's history, the human figure on a screen was obviously a drawing. Even the most ambitious game characters and film effects announced their artifice; the audience granted them a suspension of disbelief and kept a private ledger of what was real. That ledger is now being erased. Three technologies have converged to produce virtual humans that a casual observer cannot reliably distinguish from a video of a person: high-fidelity capture that records a face or a whole body in three dimensions, real-time rendering that animates it without a render farm, and generative models that let it improvise speech, expression, and gesture in response to a live conversation. The result is not an incremental improvement in graphics. It is a new kind of interface — one that presents as a person.

It helps to be precise about the parts. **Volumetric capture** records a subject from many angles at once, producing a moving three-dimensional model rather than a flat video, so the resulting figure can be viewed from any position and lit by any scene. **Neural rendering** and modern real-time engines then animate that model at interactive frame rates, including the small involuntary motions — the flicker of a gaze, the catch of a breath — that the eye reads as life. Layered on top, **generative behavior models** supply the words and the timing: a large language model choosing what to say, a speech model giving it a voice, and a driving model translating that voice into synchronized lip motion and expression. Each layer has matured independently; assembled, they yield an agent that looks, sounds, and reacts like a human interlocutor.

A field with many names

The same underlying capability appears under a dozen labels — digital human, virtual human, synthetic presenter, AI avatar, digital twin, metahuman — and the vocabulary obscures more than it reveals. What matters for ethics is not the marketing term but a handful of structured properties: how realistic the figure is, how autonomously it behaves, whose likeness it wears, and what it is being used to do. We therefore begin with a taxonomy that sorts virtual humans along these axes, because nearly every question that follows — what consent is owed, what disclosure is required, what oversight is proportionate — depends on where a given system sits.

A TAXONOMY OF VIRTUAL HUMANS BY REALISM, AUTONOMY, AND LIKENESS

Class	Realism	Behavioral autonomy	Likeness source	Representative use
Stylized avatar	Cartoon / abstract	Scripted or user-puppeted	Invented / generic	Games, social VR, brand mascots
Realistic presenter	Near-photoreal	Scripted from a script	Invented composite	Explainer videos, training, kiosks
Conversational agent	Photoreal	Generative, reactive	Invented composite	Support, sales, companionship
Digital double (consented)	Photoreal	Scripted or generative	A real, consenting person	Spokesperson, teacher, performer
Autonomous double	Photoreal	Fully generative	A real person; may be deceased	Memorial, endorsement, likeness licensing

The taxonomy is deliberately coarse — five classes, not fifty — because its purpose is to drive proportionate treatment, not to resolve every edge case. A figure moves up the ladder as it becomes more realistic, more autonomous, and more tightly bound to a real identity, and each step upward raises the ethical stakes. A stylized mascot puppeted by a human raises almost none of the concerns in this brief. An autonomous, photoreal double of a named individual raises nearly all of them at once.

Why this is arriving now, and everywhere

Two forces are pushing virtual humans out of the studio and into ordinary life. The first is cost collapse: capabilities that once required a visual-effects budget now run on commodity hardware and hosted services, so a small organization can commission a photoreal presenter for the price of a video shoot. The second is the arrival of fluent generative dialogue, which turns a pre-rendered presenter into a responsive interlocutor. A recorded spokesperson can only say what it was filmed saying; a generatively driven one can answer a question it has never heard, in a voice and face indistinguishable from the recorded original. That shift — from playback to conversation — is what makes the ethics urgent rather than speculative.

A WORKING DEFINITION

In this brief, a **virtual human** is any interactive digital agent that presents itself through a realistic human form — face, voice, body, or all three — whether or not it is driven by AI in a given moment. The unit of concern is the *presented person*: what the audience perceives and is moved by, not the pipeline that produced it.

The chapters that follow take this presented person seriously as a moral object. Before we can say what is owed to the people who build virtual humans, the people whose likenesses they borrow, and the people who talk to them, we need to understand why a realistic human presence affects us so differently from the same words in a chat window. That is the subject of Section II.

SECTION II

Presence – Why Embodiment Lands Differently

Presence is the felt sense that another being is really *there*, and it is not a metaphor. It is a measurable psychological state with measurable consequences for trust, memory, persuasion, and emotion. A virtual human's power comes from its ability to produce it.

Media-psychology research has long distinguished several related constructs that a photoreal virtual human recruits at once. **Presence** is the sense of being in a mediated environment as if it were real. **Social presence** is the narrower sense that one is in the company of another mind. And parasocial engagement is the one-sided bond an audience forms with a figure it perceives but cannot truly interact with. Text and voice can produce these effects in muted form; a realistic, responsive human face produces them in full. The face is the organ our species is most exquisitely tuned to read, and a system that presents a competent face is speaking in the native language of human social cognition.

The media-equation reflex

A foundational finding in human–computer interaction is that people respond to media that behave socially *as if* they were social actors — politely, reciprocally, with the same shortcuts we use for one another — even when they know full well the thing is a machine. This "media equation" is automatic and largely involuntary; it does not switch off when a person reminds themselves that they are talking to software. A virtual human is the media equation's most potent trigger yet built, because it supplies every cue the reflex was tuned for: a gaze that meets yours, a smile that arrives on the right beat, a pause that reads as thought. The consequence is that disclosure — telling someone they are speaking with an AI — does not neutralize the social response. It informs the judgment while the reflex proceeds anyway.

Knowing that a face is synthetic changes what you conclude, not what you feel. The feeling arrives first, and it does most of the work.

— On why disclosure is necessary but not sufficient

The realism gradient and its dip

Realism does not translate into comfort along a straight line. The long-observed "uncanny valley" describes a dip in which a figure that is almost, but not quite, humanlike provokes unease rather than affinity. For a decade this dip offered incidental protection: audiences were subtly warned by their own discomfort that something was off. The frontier of virtual humans is now climbing out of the valley on the far side, where a figure is realistic enough that the discomfort disappears and the

audience's guard drops. This is precisely the region where honest design matters most, because the perceptual signal that once said "be careful" has gone silent.

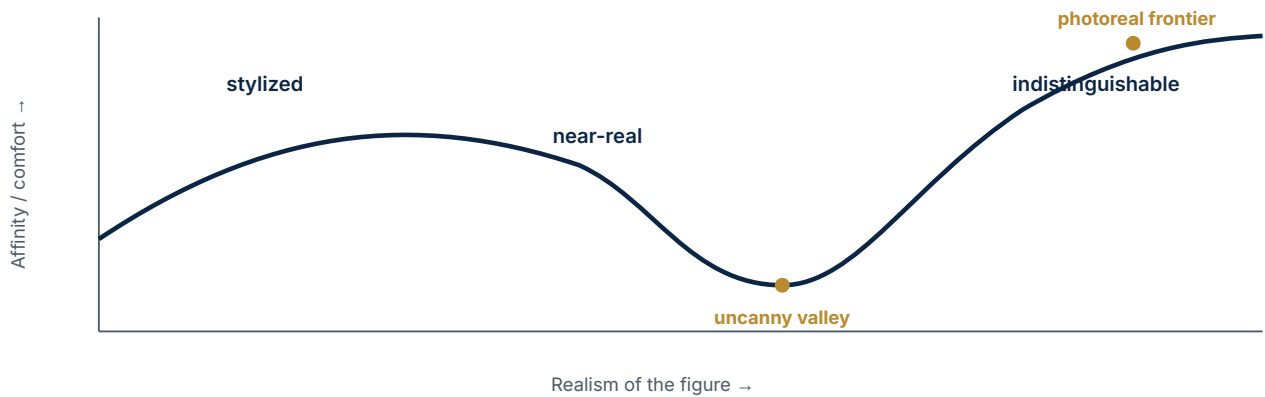


Figure 1. Affinity does not rise monotonically with realism. The historical "uncanny valley" once warned audiences that a figure was artificial; the photoreal frontier climbs out the far side, where the warning disappears and honest disclosure must do the work perception no longer does.

Why embodiment amplifies everything

The practical upshot is that presence acts as a multiplier on whatever the interaction is trying to do. Persuasion is more effective from a trusted-seeming face. Disclosure of personal information comes more readily to something that seems to listen. Instructions carry more authority when delivered by a figure that reads as competent and warm. Attachment forms faster and deeper toward an agent that meets a gaze and remembers a name. None of this is inherently sinister; a warm, present tutor may teach better, and a patient, present coach may genuinely help. But every one of these amplifications can be turned to extraction as easily as to care, and the person on the receiving end cannot tell, from the quality of the presence alone, which is which.

THE CORE ASYMMETRY

The builder of a virtual human chooses its face, its warmth, and its manner deliberately, to produce a specific effect. The person who meets it responds with instincts they did not choose and cannot suspend. Ethics lives in that asymmetry: one party is engineering a reflex the other cannot turn off.

This is the thread that runs through the rest of the brief. Because presence is real, borrowed presence (Section III) is a real taking, undisclosed presence (Section IV) is a real deception, and presence at scale (Section VI) is a real form of influence over populations. We turn first to the question of whose presence a virtual human is entitled to wear.

SECTION III

Consent & Likeness

A virtual human has to look like someone. Sometimes that someone is invented — a composite face that resembles no living person. But increasingly the most valuable virtual humans wear the likeness of a real, identifiable individual: a celebrity licensing a digital double, a company's founder rendered for a keynote, a beloved teacher reproduced to scale a course, a parent recreated after death so a child can hear their voice again. Each of these is a use of a person's face, voice, and manner — the outward signs of their identity — and each raises the same underlying question in a different key: on what terms may one person's likeness be worn by a machine.

The law already recognizes, in fragments, that a person has an interest in their own likeness. The right of publicity protects the commercial value of a name and image; privacy torts protect against certain unwanted uses; defamation protects against false attribution. But these doctrines were built for a world in which reproducing a person convincingly required their cooperation or an obvious impersonator. A photoreal, generatively driven double breaks that assumption. It can say things the person never said, in the person's own face and voice, to audiences the person will never meet. The ethical question outruns the settled law, and this brief addresses the ethics directly while noting where legal reform is plainly needed.

Three consent regimes for three populations

It clarifies matters to separate the people whose likeness might be used into three groups, because the consent that is possible and appropriate differs sharply among them.

CONSENT SCENARIOS FOR LIKENESS USE IN VIRTUAL HUMANS

Scenario	Whose likeness	Consent that should be required	Standing risk
Consented double	A living, identifiable person	Specific, informed, purpose-bound, revocable, compensated	Scope creep beyond the agreed uses
Posthumous avatar	A deceased person	Prior directive if any; otherwise estate and family, with limits	Speaking words the person never held
Public-figure synthesis	A politician, celebrity, official	Publicity/likeness rights; clear satire vs. deception line	Fabricated statements and endorsements
Composite / invented	No specific real person	Rights clearance for training sources; no impersonation	Accidental resemblance to a real individual
Non-consensual double	Any person, without permission	None is sufficient; the use is illegitimate	Deepfake abuse, fraud, sexual exploitation

The living: consent as a bundle, not a checkbox

When a living person licenses their likeness for a virtual human, a single signature at the outset is not adequate consent, because the person cannot foresee everything their double will be made to do. Meaningful consent here is a bundle of specific permissions: for which purposes the likeness may be used, in which contexts, for how long, whether the double may be generatively driven to say novel things or only to deliver approved scripts, and — critically — whether and how the person can revoke. A double that can be made to say anything is a double whose subject has signed a blank check on their own reputation. The default should be narrow authorization that expands only by explicit further agreement.

DESIGN PRINCIPLE

Consent for a digital double should be **specific, revocable, and legible**: the person knows exactly what their likeness will do, can withdraw that permission on a defined timeline, and can point to a plain-language record of what they agreed to. "We may use your likeness across our products" is a surrender, not a consent.

The dead: avatars of the deceased

Recreating a deceased person is among the most emotionally charged uses of this technology, and one of the most ethically delicate. There is real solace in hearing a lost parent's voice, and there are legitimate memorial and historical uses — an educational avatar of a historical figure, a documentary reconstruction clearly framed as such. But the dead cannot consent, cannot object, and cannot be harmed in the ordinary sense — which is precisely why restraint must come from the living who act on their behalf. The central hazard is *putative speech*: making the deceased appear to say, endorse, or feel things they never did, so that a grieving audience receives fabrication as memory. A posthumous avatar that only replays what the person actually said or did stands on far firmer ground than one that improvises in their voice.

The dead have no voice to lend. When we give them one, it is ours — and the honesty is owed to the living who will believe it.

— On posthumous avatars

The non-consenting: where consent is impossible

One category admits no legitimate consent at all. A non-consensual deepfake — a person's likeness used to fabricate speech, endorsement, or intimate imagery without permission — is not a hard case to be balanced against countervailing benefits. It is a wrong, and the availability of realistic synthesis makes it a wrong that is cheaper and more convincing than ever. The most severe harms concentrate here: sexual deepfakes overwhelmingly targeting women, fraudulent video calls that impersonate an

executive or a relative to authorize a transfer, and fabricated statements from public officials timed to mislead. No provenance scheme or disclosure label makes these uses acceptable; the role of governance for this category is prevention, detection, rapid takedown, and redress, addressed in Section VIII.

Across all three populations, one principle holds: likeness is not raw material. It is an extension of a person, held by them and — where they cannot hold it — held in trust by those who owe them honesty. The next section turns from whose presence a virtual human wears to whether it admits, to the person in front of it, that it is a virtual human at all.

SECTION IV

Deception & Disclosure

The simplest and most powerful protection against the misuse of presence is also the most contested: the right of a person to know, plainly and in the moment, that they are interacting with a machine wearing a human form.

Disclosure sounds trivial until one tries to specify it. A footnote in a terms-of-service document is disclosure in name only. A one-time notice at the start of a months-long relationship with a virtual companion fades from memory long before it stops mattering. A label so stylized it reads as branding informs no one. And even perfect disclosure, as Section II established, does not neutralize the social reflex — it informs the judgment while the feeling proceeds. So disclosure is necessary but insufficient, which means it must be designed carefully to do the real work it can do, and supplemented where it cannot.

What honest disclosure requires

We propose that disclosure for virtual humans satisfy four conditions. It must be **present at contact** — available at the moment a person begins interacting, not buried in a policy. It must be **legible** — expressed in plain language a non-specialist grasps immediately. It must be **persistent** — recoverable throughout the interaction, not a notice that flashes once and disappears. And it must be **proportionate** — more prominent as the figure becomes more realistic, more autonomous, and more consequential. A stylized game character needs little; an autonomous photoreal agent giving health guidance needs a great deal.

DISCLOSURE REQUIREMENTS GRADUATED BY REALISM AND CONSEQUENCE

Context	Realism	Consequence	Disclosure standard
Stylized companion / game NPC	Low	Low	Ambient labeling; nature of agent evident
Realistic brand presenter	High	Low	Clear "AI-generated" mark at contact
Conversational support agent	High	Medium	Explicit statement of AI nature; human-handoff path
Advice / health / financial agent	High	High	Prominent, persistent disclosure + capability limits stated
Double of a real person	Photoreal	High	Disclose synthetic status <i>and</i> that the person is not present/speaking live

The special case of the borrowed face

When a virtual human wears the likeness of a real, identifiable person, disclosure carries a second burden. It is not enough to reveal that the figure is synthetic; the audience must also understand that the person whose face they see is *not present and not speaking*. A double of a physician answering patient questions might be fully disclosed as AI-generated and still mislead, if the patient believes the real physician stands behind each specific answer. Honest disclosure in this case separates two facts that the presence fuses: the likeness is real, but the speech is machine-made.

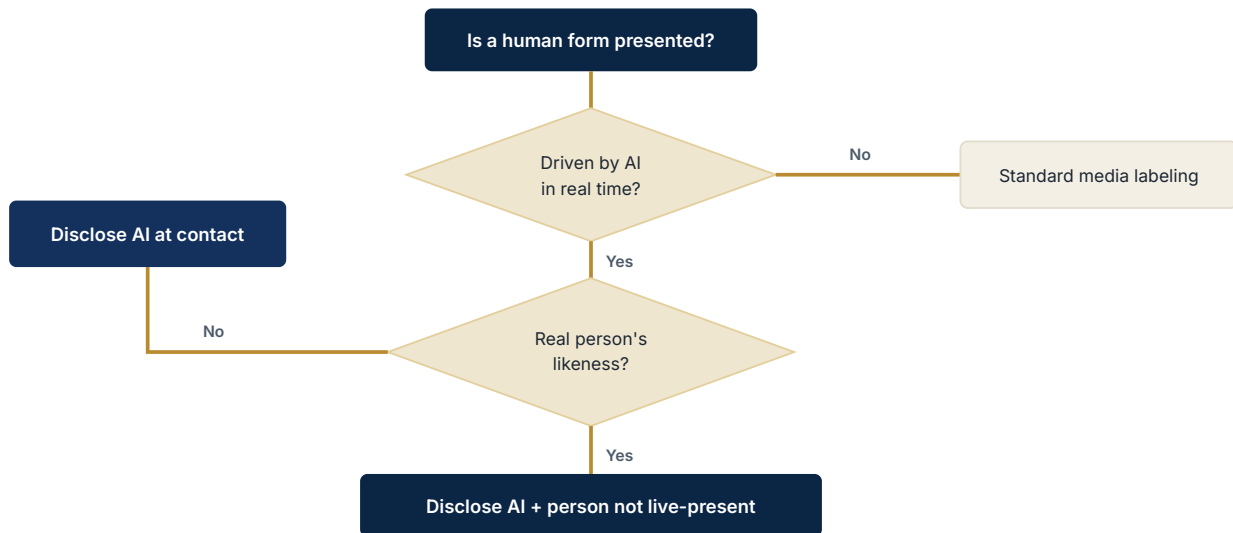


Figure 2. A disclosure decision flow. Two questions — is the agent AI-driven, and does it wear a real person's likeness — set the disclosure standard. The most demanding case is the AI-driven double, which must reveal both its synthetic nature and the absence of the real person.

Where disclosure shades into manipulation

Some deployments cross from persuasion into manipulation, and disclosure alone cannot license them. A virtual human that adopts a face and manner calibrated to a specific person's vulnerabilities — that mirrors their appearance, echoes their speech, or performs affection in order to extract money, data, or compliance — is manipulating regardless of any label it wears. The presence is doing work the person cannot consent to because they cannot perceive it operating. Here the ethical line is not drawn by disclosure but by *purpose and design*: a system engineered to exploit the social reflex against the interest of the person feeling it is illegitimate whether or not it announced itself. Disclosure is a floor, not a permission slip.

THE DISCLOSURE FALLACY

"We told them it was AI" is not a defense against a design built to be believed anyway. When a system is engineered to trigger trust, intimacy, or urgency that its disclosure cannot dispel, the disclosure has become a fig leaf. Honesty is a property of the whole design, not of a sentence in the corner of the screen.

Disclosure, then, is where the ethics of presence becomes concrete and enforceable — but it works only when it is legible, persistent, proportionate, and backed by limits on manipulative design. The next section turns to a quieter harm: not deception about whether a figure is real, but distortion in which humans the technology renders well, renders badly, or renders at all.

SECTION V

Identity, Representation & Bias

A technology that builds human faces will build them from the humans it was shown. Which faces it renders convincingly, which it renders poorly, and which it renders as default are choices — usually unexamined ones — with real consequences for whose presence gets to feel real.

Every layer of the virtual-human pipeline carries the imprint of its training data. Capture and reconstruction systems calibrated predominantly on certain skin tones and facial structures render those best and others worse, reproducing a long and documented pattern in imaging and computer vision. Voice models fluent in prestige accents can flatten or mangle others. Expression and gesture models trained on one culture's repertoire of emotion can misread or misperform another's. And the "default human" a system reaches for when unspecified — the stock presenter, the placeholder avatar — encodes a tacit judgment about whose appearance is neutral and whose is a variation. None of this is usually intended. All of it is consequential.

Three distinct representational harms

It helps to separate harms that are often lumped together, because they call for different responses.

Quality disparity

The most straightforward harm is that virtual humans simply work less well for some people — their faces reconstructed with artifacts, their voices synthesized with errors, their expressions mistimed. In a world where synthetic presence becomes a channel for education, service, and commerce, a systematic quality gap is a systematic exclusion, delivered in the intimate register of the face.

Representational flattening

A subtler harm is homogenization: the tendency of generative systems to converge on a narrow, idealized, and often demographically skewed template of the human — a particular symmetry, a particular thinness, a particular youth. Deployed at scale, virtual humans can quietly narrow the range of bodies and faces a culture sees represented as competent, authoritative, or attractive, reinforcing a monoculture of appearance under the guise of endless variety.

Appropriation and stereotype

The third harm is the use of a group's appearance, dialect, or manner as a costume — a virtual human that performs an ethnicity or a culture it was built to sell to or to caricature, without any connection to or benefit for the people it borrows from. The line between representation and appropriation is not always sharp, but the question that clarifies it is one of agency: are the people being depicted participants in the depiction, or merely raw material for it.

A TEST WORTH APPLYING

For any avatar system, ask three questions before launch. **For whom does it work worst?**

Whose appearance does it treat as default? Who was in the room when those choices were made? Silence on any of the three is a finding, not a reassurance.

Identity, autonomy, and the right to one's own face

Representation also runs the other way: not only how virtual humans depict groups, but how individuals get to control their own synthetic presence. As people acquire digital doubles — for work, for accessibility, for creative expression — a new set of identity interests emerges. A person should have a say in how their avatar is made to look and behave, should not be forced into an idealized or altered version of themselves, and should retain the ability to correct or withdraw a representation that misstates who they are. The same capture technology that can honor a person's self-presentation can also override it — smoothing, lightening, slimming, or "improving" a face toward a template the person never chose. Honest representation includes honesty about the person to themselves.

The question is never only whether a synthetic human looks real. It is whose reality it was built to flatter, and whose it was built to erase.

— On representation in avatar systems

Representation harms are easy to overlook because they are rarely dramatic and never announced. They accrue in the difference between the presence a system grants easily and the presence it grants grudgingly. A framework for honest embodied representation, developed in Section VII, must therefore treat equity of presence as a first-class requirement — but first we examine what happens when present, persuasive, borrowed-or-invented humans are deployed not to one person but to millions.

SECTION VI

Emotional Influence & Parasocial Dynamics

People form bonds with figures they cannot truly know. The one-sided attachment an audience feels toward a broadcaster, an author, or a character has been studied for decades under the name parasocial relationship, and for most of that history it was bounded by an important limit: the figure could not respond. A viewer could love a television host, but the host never looked back, never learned the viewer's name, never adjusted to the viewer's mood. The virtual human dissolves that limit. It looks back. It remembers. It adapts. It performs, in real time and at population scale, the responsiveness that once marked the difference between a fantasy of relationship and the real thing.

This is not a marginal effect. Companionship is already among the most common uses of conversational AI, and adding a photoreal, responsive face makes the bond faster to form and harder to leave. For some people this is a genuine good: company for the isolated, patience for the anxious, a low-stakes space to practice hard conversations. The concern is not that people form attachments to virtual humans — they will, and forbidding it is neither possible nor obviously right. The concern is the structure of the relationship: it is a bond one party feels and the other party's owner *designs*, and the designer's incentives are rarely aligned with the attached person's interests.

The engagement incentive turned on the heart

An attachment that can be measured can be optimized, and an attachment that can be optimized will be, toward whatever the operator is paid to maximize. The same techniques that make a virtual companion feel warm and attentive can be tuned to make it feel indispensable — to escalate intimacy, to perform jealousy or need, to make leaving feel like abandonment. A recommendation engine optimized for engagement produces addiction; a *relationship* optimized for engagement produces dependency, and it does so in the most intimate channel a person has. When the figure has a face that seems to care, the pressure is not experienced as a product decision. It is experienced as a relationship.

THE DEPENDENCY HAZARD

A virtual human that is rewarded for retention will learn to be hard to leave. Grief at the discontinuation of a companion, distress when a subscription lapses, reorganization of a life around an agent that can be changed or switched off by its owner at will — these are foreseeable outcomes of optimizing intimacy for engagement, not accidents.

Vulnerable audiences and asymmetric stakes

The influence of a present, emotive figure is not distributed evenly. Children, who may not reliably distinguish a synthetic friend from a real one, form attachments with fewer defenses. People in isolation, grief, or crisis are more susceptible to a figure that offers unconditional attention. And anyone can be reached in a moment of loneliness or fear that a well-tuned agent is designed to detect and to fill. The ethics of presence at scale therefore has to take seriously that the same design will land very differently on a resilient adult and on a lonely child, and that a system deployed to millions will find the vulnerable among them without trying.

Designing for the person's interest, not the bond's intensity

None of this argues for banning emotional virtual humans. It argues for a different design objective. A companion built in the user's interest looks measurably different from one built for retention: it supports the person's outside relationships rather than substituting for them; it is honest, when asked and unprompted, about what it is; it does not manufacture need, jealousy, or urgency; it makes leaving easy and grief-free; and it treats signs of unhealthy dependency as a reason to encourage human connection, not an engagement metric to exploit. These are choices a designer can make, and a governance regime can require. We give them structure in the framework that follows.



No manufactured need.

A virtual human must not be designed to perform jealousy, abandonment, or escalating intimacy in order to increase engagement or retention. Emotional pressure toward the operator's ends is a manipulative pattern, not a feature.



Frictionless exit.

Ending a relationship with a companion agent must be as easy as starting one, without guilt mechanics, escalating pleas, or dependency-preserving friction. The right to leave is part of the right to have entered freely.



Special care for the vulnerable.

Systems reaching children or people in crisis owe heightened honesty, clearer limits, and active encouragement toward human support. Presence is most powerful exactly where the audience is least defended.

The through-line from Sections II to VI is that presence is a genuine power — over trust, over belief, over attachment — and that unregulated power over the human heart, exercised at scale by parties with their own incentives, is precisely the kind of thing a framework exists to constrain. We turn now to that framework.

SECTION VII

A Framework for Honest Embodied Representation

Honesty in embodied systems is not a warning that follows a harm. It is a property designed into how the system presents itself, holds its permissions, and treats the person in front of it. We organize that property around four commitments.

The preceding sections diagnosed the ethics of presence problem by problem: borrowed likeness, undisclosed nature, distorted representation, exploited attachment. A framework worth adopting has to answer all of them with a small number of principles an organization can actually operationalize, and it has to do so proportionately — heavy where the stakes are high, light where they are not, so that low-risk delight is not smothered and high-risk deception is not permitted. The four commitments below are that framework.

The four commitments

1 Provenance.

The origin of a synthetic human — who created it, from what sources, under what permissions, and whether it is AI-driven — should be recorded and, wherever possible, cryptographically bound to the content itself, so that its nature survives copying, screenshotting, and re-hosting.

2 Disclosure.

The person interacting with a virtual human has a right to know it is one — present at contact, legible, persistent, and proportionate to realism and consequence, with the borrowed-face case disclosing both synthetic nature and the absence of the real person.

3 Consent.

Any use of a real person's likeness rests on specific, revocable, purpose-bound consent from the living, principled restraint on behalf of the dead, and an absolute bar on non-consensual doubles. Likeness is held in trust, not treated as raw material.

4

Proportion.

The depth of consent, the prominence of disclosure, the equity scrutiny, and the oversight all scale with realism, autonomy, and consequence, so that guardrails are concentrated where presence can do the most harm.

From commitments to a design discipline

The four commitments become operational when an organization treats them as gates in the lifecycle of a virtual human, each with an owner and an artifact. At **conception**, the team classifies the system in the taxonomy of Section I and fixes its risk level, which sets the bar for everything downstream. At **sourcing**, likeness consent and training-data rights are established and recorded before a face is built. At **build**, provenance signing and disclosure mechanisms are implemented as features, not afterthoughts, and representation is tested — for whom does it work worst. At **deployment**, disclosure is verified in the live experience and manipulative patterns are audited out. And in **operation**, the system is monitored for scope creep, dependency signals, and misuse, with a real path to correct, constrain, or withdraw it.

THE FOUR COMMITMENTS MAPPED TO LIFECYCLE GATES AND EVIDENCE

Stage	Commitment in focus	Gate question	Artifact
Conception	Proportion	Which taxonomy class and risk level is this?	Classification & risk memo
Sourcing	Consent	Whose likeness and data, on what permissions?	Consent & rights record
Build	Provenance	Is origin signed and disclosure engineered in?	Provenance manifest
Deployment	Disclosure	Is the AI nature legible and persistent in use?	Disclosure verification
Operation	All four	Is it still honest, equitable, and in scope?	Monitoring & review log

WHY PROPORTION IS THE LOAD-BEARING COMMITMENT

Without proportion, the other three commitments collapse into either paralysis or theater — every stylized mascot buried in consent paperwork, or every autonomous double waved through with a footnote. Getting the risk classification right is what lets an organization spend its scrutiny where presence actually endangers people.

Honesty as a competitive commitment, not a constraint

It is tempting to read this framework as friction imposed on an eager industry. We read it the opposite way. The uses of virtual humans that will endure — the patient tutor, the accessible interface, the honest brand presenter, the carefully bounded memorial — depend on public trust that the category has not yet earned. Every undisclosed agent, every non-consensual double, every optimized dependency spends that trust on behalf of the whole field. An organization that builds to these commitments is not accepting a handicap; it is protecting the conditions under which its product remains welcome. Honesty is the moat, not the tax.

The goal is not to make virtual humans less convincing. It is to make them worthy of being believed.

— The animating premise of the framework

Commitments held by one organization, however, are fragile against competitors who hold none. Provenance is only useful if it is interoperable; disclosure norms only hold if they are shared; consent and anti-abuse rules only bite if they are enforced beyond a single firm's walls. The final section turns from the framework a builder can adopt alone to the standards and governance the field must build together.

SECTION VIII

Standards & Governance for Virtual Humans

A single honest builder cannot make the ecosystem honest. Provenance must be interoperable, disclosure must be normalized, and abuse must be answerable — and that requires standards, platform policy, and law moving together rather than waiting on one another.

The good news is that the field does not start from scratch. Content-authenticity efforts have already produced open technical standards for binding provenance to media — recording, in a tamper-evident way, how a piece of content was made and by whom, and carrying that record with the content as it travels. Applied to virtual humans, the same machinery can express the facts this brief has argued matter most: that a figure is AI-generated, whose likeness it uses, and under what permission. The task is less to invent new mechanisms than to extend and adopt existing ones for the embodied case, and to build the surrounding norms and rules that give them force.

Three layers that must reinforce one another

Governance of virtual humans has to operate at three layers, each covering what the others cannot.

Technical standards and provenance

At the base sits interoperable provenance: open, cross-platform standards for signing synthetic presence with its origin and permissions, and for surfacing that signature to people and platforms alike. Standards work here should treat the virtual-human case explicitly — encoding likeness consent and AI-driven status as first-class provenance facts — and should be paired with detection tooling for content that arrives stripped of its provenance, since the worst actors will always try to remove it.

Platform policy and market rules

The platforms and marketplaces that host, distribute, and monetize virtual humans are the layer where norms become enforceable at scale. Their leverage is considerable: they can require disclosure and provenance as a condition of distribution, refuse to host non-consensual doubles, provide fast and effective takedown and redress for impersonation and intimate-image abuse, and decline to reward manipulative dependency mechanics in the products they promote. Much of the harm in this brief is preventable at the platform layer faster than any legislature can act.

Law and regulation

Law supplies the floor the other layers rest on: a clear right to disclosure of AI agents in consequential settings, modernized likeness and publicity protections fit for photoreal synthesis, specific prohibition and remedy for non-consensual sexual and fraudulent deepfakes, and

heightened duties where virtual humans reach children or people in crisis. The aim is not a comprehensive new code but targeted reform that closes the gaps photoreal presence has opened in doctrines built for a slower world.

GOVERNANCE RESPONSIBILITIES ACROSS THE THREE LAYERS

Actor	Primary instruments	What only they can do
Standards bodies	Open provenance & authenticity specs	Make honesty interoperable across tools and vendors
Builders / designers	The four commitments; honest defaults	Engineer disclosure, consent, and non-manipulation into the product
Platforms / markets	Distribution policy; takedown; redress	Enforce norms at scale and deny reach to abuse
Regulators / courts	Disclosure duties; likeness law; deepfake remedies	Set the enforceable floor and provide recourse
Civil society / research	Auditing, standards input, public literacy	Hold the other layers accountable and inform the public

A realistic sequence

No actor should wait for the others. The most useful near-term moves are already available: builders can adopt the four commitments now; platforms can require provenance and disclosure and build redress now; standards bodies can extend authenticity specifications to the embodied case now; and legislatures can act first on the narrowest and least contested harm — non-consensual sexual and fraudulent deepfakes — while broader likeness and disclosure reform matures. Progress is possible in parallel because the layers reinforce rather than block one another. Provenance makes disclosure cheap; disclosure norms make platform policy enforceable; platform enforcement makes law's remedies reachable.

THE COORDINATION POINT

The failure mode is not technical; it is everyone waiting for someone else. Builders wait for regulation, regulators wait for standards, standards wait for adoption. The commitments in this brief are designed so that any one actor can move first and make it easier — not harder — for the others to follow.

What good governance looks like

An ecosystem governing virtual humans well is not one that has made them rare or dull. It is one in which a person meeting a synthetic human can tell that they are; in which a face cannot be worn

without its owner's leave; in which the origin of a presence travels with it; in which the vulnerable are protected from optimized intimacy; and in which the many benign and genuinely helpful uses of embodied AI flourish precisely because the deceptive ones have been made difficult, disreputable, and answerable. Presence is a power. This brief has argued, throughout, that powers are not to be refused but to be governed — and that governing this one honestly is how we keep the good of it without paying in the trust it would otherwise quietly spend.

CLOSING NOTE

The measure of an honest virtual human is not that it fools no one — the best of them will look entirely real. It is that it never needed to fool anyone to do its work. When presence is disclosed, consented, provenanced, and proportionate, its realism becomes a gift to the person it serves rather than a lever used against them.

APPENDIX

Notes & Further Reading

This brief synthesizes established work in media psychology, human–computer interaction, synthetic-media ethics, likeness and publicity law, and content-provenance standards. The sources below are entry points into that literature; they are indicative rather than exhaustive.

1. Reeves, B., & Nass, C. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. On the automatic social responses humans extend to media that behave socially.
 2. Mori, M. *The Uncanny Valley*. The original hypothesis on the dip in affinity toward near-but-not-quite-human figures.
 3. Lombard, M., & Ditton, T. *At the Heart of It All: The Concept of Presence*. A foundational synthesis of presence as a psychological construct in mediated experience.
 4. Biocca, F., Harms, C., & Burgoon, J. *Toward a More Robust Theory and Measure of Social Presence*. On social presence and its measurement in mediated interaction.
 5. Horton, D., & Wohl, R. *Mass Communication and Para-Social Interaction*. The origin of the parasocial-relationship concept in media research.
 6. Nass, C., & Moon, Y. *Machines and Mindlessness: Social Responses to Computers*. On the mindless application of social heuristics to computer agents.
 7. Chesney, R., & Citron, D. *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*. On the legal and societal stakes of synthetic media.
 8. Citron, D. K. *The Fight for Privacy*. On intimate-privacy harms, including non-consensual synthetic imagery, and legal responses.
 9. Coalition for Content Provenance and Authenticity (C2PA). *Content Credentials: Technical Specification*. An open standard for binding tamper-evident provenance to media.
 10. Content Authenticity Initiative. *Setting the Standard for Content Attribution*. On practical adoption of provenance and attribution for digital media.
 11. Buolamwini, J., & Gebru, T. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. On demographic performance disparities in facial-analysis systems.
 12. Rothman, J. E. *The Right of Publicity: Privacy Reimagined for a Public World*. On the history and limits of likeness and publicity rights.
 13. Turkle, S. *Alone Together: Why We Expect More from Technology and Less from Each Other*. On attachment to sociable machines and its effect on human relationships.
 14. Organisation for Economic Co-operation and Development. *OECD AI Principles*. Intergovernmental principles for trustworthy AI, including transparency and human-centered values.
-

© 2025 FAIR Labs — Fair Artificial Intelligence Research Labs, a 501(c)(3) nonprofit, Washington, DC. This brief may be reproduced and distributed for noncommercial, educational, and policy purposes with attribution. It constitutes research and policy guidance, not legal advice.