FIVE + FIVE

Originally developed by Vinebright Foundry (2026)

FIVE LEVELS OF HUMAN IMPACT.

FIVE PRINCIPLES FOR RESPONSIBLE AI IN GAMES.

FIVE + FIVE

Framework for Responsible AI in Games

Applied in practice through

Helix Live Brain™ and *Of Moss & Moonlight*

Version: V1.0 Public Release

Author: Vinebright Foundry

Last Updated: May 2026

Context: This document presents the Five + Five framework for Responsible AI in Games, alongside its practical application through Vinebright Foundry's Helix Live Brain™ and *Of Moss & Moonlight*.

Five + Five combines: The Five Levels of Human Impact for AI in Games, and The Five Principles of Responsible AI in Games.

Together, they provide a shared language for discussing human impact, emotional safety, consent, transparency, and responsible implementation in AI-enabled game systems.

Table of Contents

Part I - FIVE + FIVE Framework

Foundations, Principles, and References for Responsible AI in Games.

1. Foundations for Responsible AI in Games

Further foundational context is available at vinebrightfoundry.com/five-five.

1.1 About the Five + Five Framework

This framework combines two complementary ideas. The *Five Levels of Human Impact for AI in Games*, which help us understand how different AI systems affect players and therefore require different approaches to social license, and the *Five Principles of Responsible AI in Games*, which offer proportional guidance for responsible design and implementation.

Together, Five + Five offers a shared language to discuss emotional impact and responsible implementation without prescribing uniform solutions, technology choices, or gameplay patterns.

This document presents the Five + Five framework as a shared model for the industry in sections 1-5, and shows how Vinebright applies it in practice through Helix and *Of Moss & Moonlight* in sections 6-10.

We are eager for critique, contribution, and adaptation from across the creative, player, and research communities as this work evolves.

Vinebright is also seeking participants to form the Reference Group for Responsible AI in Games, a voluntary, cross-disciplinary forum to support ongoing discussion and evolution of responsible practice. Expressions of interest are open at: vinebrightfoundry.com/contact-expression-of-interest.

1.2 State of the Industry

Long-term success and sustained adoption of AI technologies depend not only on technical implementation, but on responsible behavior and proactive stakeholder engagement.

Specifically in the gaming industry, reactive systems and AI-enabled narrative tools are becoming more central to how games behave. Stakeholders across the industry bring different, and sometimes conflicting, risk tolerances, incentives, and expectations around trust, consent, and accountability.

We have unique opportunities for the use of AI, and therefore unique challenges – including the level of emotional engagement, issues of identity, parasocial interaction, and consequence.

We have a chance to learn from the challenges of other sectors and maximize how AI can improve the experience of players, creators, and communities.

1.3 Social License: Trust, Legitimacy, and Community Permission

In the deployment of AI systems, social license is the formal or informal agreement by stakeholders engaging with the technology (i.e. customers, users, governance, beneficiaries) to use AI in a manner or scope that is set

out by the vendors or implementation entities. It is essentially the community's permission for the AI to operate in the way it has been communicated to - and this permission must be earned, not assumed.

The lack of conversation about social license in gaming should be considered a growing risk to trust – and when public attention shifts toward AI in games, that tension will surface quickly.

Not all AI in gaming is created equal. Different systems have different levels of visibility, consequence, and player-facing influence. Some operate deep in infrastructure pipelines; others speak, respond, remember, and build relationships. These differences determine the level of social license required.

This spectrum forms the basis of the Five Levels of Human Impact for AI in Games.

1.4 Intended Audience & Orientation

This paper is written primarily for practitioners and decision-makers willing to engage with responsibility as a practical and strategic concern, recognizing that unmanaged risk to trust, reputation, and long-term value becomes materially visible as AI systems move closer to players. This includes developers, designers, producers, performers, creatives, and platform teams directly involved in shaping living game worlds.

It may also be useful to players, researchers, ethicists, regulators, and platform stakeholders who are affected by, studying, or engaging with these systems.

Note: This work is written with an international industry context in mind. While it is designed using practical experience, it is not tied to any single national regulatory regime and is intended to be legible and applicable across different jurisdictions.

1-5: FIVE + FIVE FRAMEWORK	→	6-10: VINEBRIGHT APPLIED PRACTICE
• Industry Challenge • Human Impact • Design Principles • Shared Stewardship		• Helix Capabilities • Of Moss & Moonlight Commitments • Emerging Work Categories • Open Questions

2. Vinebright Foundry's Position on Responsible AI in Games

More information about Vinebright Foundry is available at vinebrightfoundry.com/about.

Vinebright Foundry builds lore-rich, immersive game worlds, and the AI-powered Living World Intelligence Layer behind them (Helix Live Brain™).

As we developed increasingly responsive and emotionally adaptive systems, it became clear that ethical responsibility could not be separated from technical delivery. What began as an ambition to create rich, reactive worlds quickly raised broader questions around safety, consent, emotional impact, and player trust that needed to be addressed by design from the outset.

We believe AI technology is at critical risk of outpacing its social license – the informal but critical "permission" granted by society and communities for how it is used.

Our perspective draws on experience working with trust-sensitive systems, where obtaining and sustaining social license has been critical to successful implementation. We have applied this understanding to the creation of the Five + Five Framework for Responsible AI in Games.

AI is already changing how games are built. The question is not whether it appears in our worlds, but how deliberately, transparently, and responsibly we choose to use it.

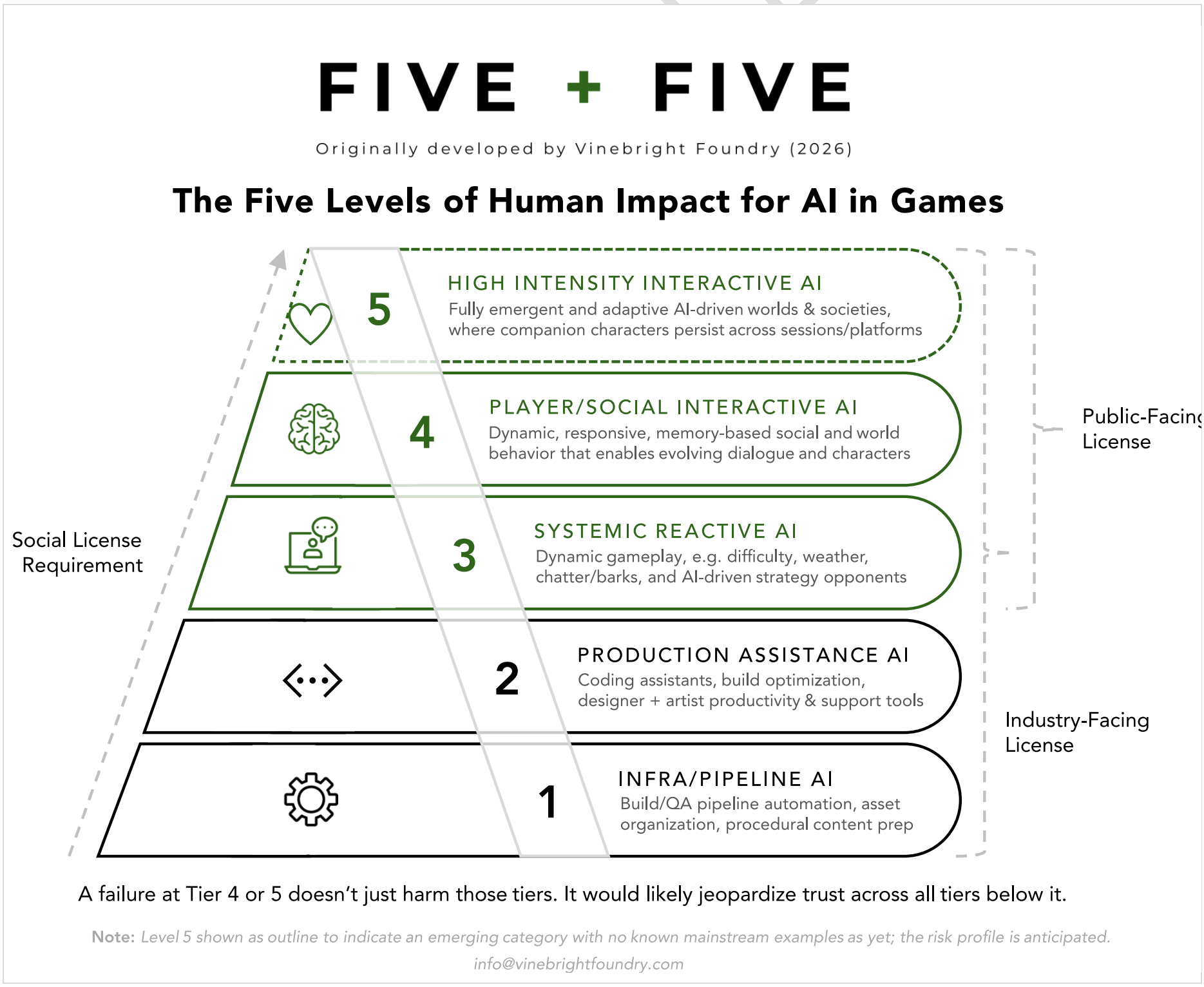
3. The Five Levels of Human Impact for AI in Games

The Five Levels of Human Impact are a key part of the Five + Five Framework, which can be further explored at vinebrightfoundry.com/five-five-levels-human-impact.

3.1 The Five Levels of Human Impact Overview

Not all uses of AI in games require the same permission, the same level of transparency, or engagement from the same communities.

This difference can be understood as a spectrum of human impact, outlined in the Five Levels of Human Impact for AI in Games.



For lower-impact AI used in development workflows, the relevant social license sits with the people who build and contribute to games, not the people who play them.

At Level 1, this includes infrastructure and pipeline systems such as build and QA automation, asset organization, and procedural content preparation - systems that operate behind the scenes and are not directly experienced by players. At Level 2, this extends to tools used by developers and creative teams, such as code assistants, design support tools, and Text-to-Speech (TTS) for internal prototyping.

This responsibility doesn't stop at Levels 1 and 2. The requirement for industry-facing social license continues into higher-impact systems.

Once AI touches the player experience, the affected community expands. Player-facing systems require player social license, not just creator agreement. And as the systems move closer to emotional or relational interaction - where the likelihood of parasocial bonds, behavioral influence, and ultimately identity distortion becomes possible - the need for social license increases sharply.

3.2 The Five Levels of Human Impact Detail

This framework focuses in more detail on levels 3-5, specifically where AI begins to directly shape the player's world in increasingly sensitive ways.

At these levels, public trust, emotional safety, and narrative integrity become essential - and maintaining social license becomes a core responsibility.

Level 3: Systemic Reactive AI

First level of player visibility - shapes the world and experience, but not relationships.

Level 3 systems influence the world around the player rather than the player directly. They may adjust game difficulty, weather, ecology, factions, world states, barks and ambient chatter, ambient flavor, or other systemic behaviors (e.g. AI-enabled strategy NPCs as seen in GOAP). They increase responsiveness and dynamism - but they do not attempt to understand the player, mirror them, or form relationships.

Human Impact: Low • Neutral-to-mild emotional responses (surprise, amusement, anticipation, annoyance) • Frustration due to tuning • Fragmented immersion if lore breaks • Gameplay distortion or imbalance	**Indicative Public-Facing Social License:** Public Awareness • Basic awareness of the use of AI, and the type of AI in use, in the relevant game(s) • Transparency is sufficient; consent is not yet required

Level 4: Player/Social Interactive AI

Player recognition, adaptive memory, personalized interactions, can form bonds and build rapport.

At Level 4, AI systems would cross a fundamental boundary - from functions of world responsiveness, to ones of relational responsiveness. This changes who must consent, what must be disclosed, and what must be actively monitored.

These systems may recognize the player, speak to them, remember past interactions, mirror tone, adjust tone and style to the player, and sometimes form alliances, friendships, or romantic trajectories.

This level introduces parasocial relationship potential, emotional hooks, and the possibility of a player feeling understood, recognized, or "connected" to an AI character. These effects may persist beyond the game and influence player expectations, emotional state, or behavior outside the game environment.

These systems are not designed to diagnose, treat, or replace psychological support, and are not intended to intervene in real-world mental health needs. Additional safeguards, constraints, or exclusions are assumed where systems are accessible to minors.

Human Impact: High • Mild-to-moderate emotional responses (satisfaction, laughter, enjoyment, appeal) • Narrative drift, misinterpretation of intent, or generation of harmful or inappropriate content • Self-disclosure of personal information • Boundary confusion, parasocial dependency, and identity disorientation • Unrealistic or excessive emotional reinforcement (e.g. affirmation, mirroring)	**Indicative Public-Facing Social License:** Public Consent + Authored Limitations • Clear, comprehensible consent before interaction begins, distinct from general legal terms • Authored boundaries • Clear limits on subject matter • Guardrails • Monitoring • Escalation paths Where harm occurs despite safeguards, responsibility for response, correction, and repair remains with those accountable for the operation and governance of the system and its player experience.

Level 5: High Intensity Interactive AI

Emergent worlds, persistent personas across worlds and games, adaptive psychological models.

Level 5 represents a transformational horizon - where AI systems not only interact, but autonomously adapt, generate, and expand. They may self-evolve lore, shape entire societies and worlds, simulate emotional behavior, or target players based on psychological profiling.

This is the point at which AI systems would be considered to resemble entities within a world, rather than authored components. Advances in world-modeling or simulation alone would not constitute High-Intensity Interactive AI unless they are coupled with persistent, player-facing social or emotional interaction.

A failure of trust at this level would likely jeopardize social license across the entire pyramid.

Human Impact: Transformational • Strong and potentially disproportionate emotional responses (e.g. joy, desire, anger, fear)	**Indicative Public-Facing Social License:** *Public Trust + Oversight + Governance* • Active transparency and public literacy regarding function, impact, and scope

• Reduced visibility of boundaries, loss of narrative control, and emergent behaviors beyond creator intent • Expectations of continuity, companionship, and ongoing presence • Deep parasocial immersion, identity entanglement, and difficulty disengaging from the world • Failure of safeguards may result in harm that is no longer contained within the game itself	• Explicit, meaningful, educated consent - ongoing or progressive as interaction and emotional engagement increase • Creator-authored boundaries for use, monitoring, and exit/expulsion use cases • Extended community engagement • In-game safety mechanisms and escalation pathways • Proportionate external review and governance • Legal and ethical disclosures

3.3 Unacceptable Human Impacts (Non-Exhaustive)

The following examples illustrate categories of human impact that undermine emotional safety, consent, and social license in AI-enabled games. They are not an exhaustive list, nor specific implementation guidance, but represent boundaries beyond which player-facing AI systems cease to be responsible.

In practice, safeguarding against these harms requires guardrails that operate at runtime. These shape what AI systems can generate, respond to, or escalate in live player interactions, not only policies or intentions defined in design.

- Sexualized or pornographic misuse, including non-consensual or exploitative fantasy scenarios
- Harm to or targeting of minors, including exposure to age-inappropriate content or dynamics
- Manipulative or coercive parasocial dynamics, including systems designed to foster emotional dependency
- Identity or likeness misuse, including unauthorized use of real individuals or cultures
- Spillover harm, where in-game AI interactions contribute to real-world psychological or social harm

4. The Five Principles of Responsible AI in Games

More detail about the Five Principles can be found at vinebrightfoundry.com/five-five-principles.

4.1 The Five Principles Overview

The Five Principles of Responsible AI in Games form the core design commitments that make living world systems safe, ethical, and worthy of community trust.

These Principles are what we believe should be considered foundational ethical design and implementation commitments for AI-enabled game systems, especially those that influence the player experience (corresponding broadly to Levels 3–5 of the Human Impact model).

How these Principles are implemented is ultimately up to creators, studios, and vendors. Our intention is to articulate what responsible design looks like; we share our patterns so others can adapt, build, extend, and improve.

4.2 Principle Summary

The following sections explore each principle - how they apply to real design decisions, where the risks sit, and why emotional safety, creator sovereignty, and transparency must be treated as core system requirements, by design, not by footnote.

A. Player Safety & Respect

Comprehensible, progressive consent-driven; player-first, clear in purpose, risk, and benefit.

Principle Goal

Ensure players feel secure when using AI-enabled game systems, with clear understanding of purpose, risks, and benefits, and founded on player agency and emotional wellbeing. We do not simply aim to avoid harm – we want to build AI characters and worlds players can trust, enjoy, and are safe returning to.

Indicative Human Impact

When boundaries are unclear, unpredictable, or unintentional, players may feel confused, overwhelmed, or unsafe - even without obvious harm, and sometimes beyond the game itself.

When gaining emotional trust is part of the experience, and players have clarity about scope, benefits, risks, and privacy, they can feel relaxed enough to truly engage, and enjoy positive immersion without intensity, pressure, or escalation. This emotional pattern can be expressed simply as:

BOUNDARIES → TRUST → RELAXED PLAY → POSITIVE IMMERSION

Responsible Design Considerations

- When clearly articulated and accessible, system scope and boundaries can be a key tool in building trust. When they change, engagement and awareness can help preserve player confidence
- The higher the potential human impact, the more formal the consent flows that may be appropriate. Consent mechanisms reduce risk, but cannot eliminate all vulnerability in emotionally complex play.
- Ongoing monitoring can be designed that recognizes and responds to risky interaction patterns
- Studios are encouraged to author emotional boundaries and clear pathways for escalation
- Avoiding manipulation, intense emotional hooks, inappropriate intimacy escalation, and unpredictable emotional models can support safe and enjoyable engagement
- Player identity or disclosures should be treated as explicit and opt-in, rather than inferred or derived

Industry Opportunities

Designing for Player Safety & Respect does not mean restricting tone or limiting intensity - instead it encourages intentional emotional architecture appropriate to genre, audience, and narrative purpose.

Responsible design becomes a constraint only when left to the end; when authored early, it turns emotional safety into a craft layer rather than a compliance burden.

- Dialogue, pacing, and relational arcs can reflect authored intent, with emotional consequences that feel legible and proportionate, and are not unpredictable in nature
- Games can embrace tension, intimacy, conflict, dramatic stakes, and even darkness - when those elements are intentional rather than algorithmic
- Players can feel emotional agency - interactions are readable, voluntary, and consistent with tone, not coercive or identity-threatening
- Creative teams are freed to focus on designing emotional architecture, desired tone and feeling, and world lore, rather than manually implementing every reactive branch or repetitive dialogue detail

Worlds don't get smaller when designed safely - they get deeper. Guardrails unlock trust, not limits.

B. Creative & Performer Rights

Creator-directed, performer-protected, consent-bound within agreed scope.

Principle Goal

Ensure creators and performers feel respected, protected, and confident when contributing their work to be used in game systems. Participation should feel safe, without fear of misrepresentation or silent reinterpretation, and the possibilities of AI should elevate craft, rather than diminish rights.

Indicative Human Impact

If creative rights are unclear or unbounded, and contributions are reused or reinterpreted without consent, creatives may feel misrepresented, exposed, lose trust in the industry, and become reluctant to participate.

When terms of use are transparent, respectful change processes exist, and contributors retain clear agency over scope, collaboration on games will feel safer and more inviting. Contributors are confident, advancements become opportunities, and AI is a medium for elevating creativity rather than appropriation.

Responsible Design Considerations

- Respectful boundaries and consent encourage creative participation, which can support richer collaboration, more intentional narrative expression, and artistic originality
- Public misunderstandings about AI may create reputational pressure for creatives, even when ethically scoped. Visible alignment with human authorship and fair terms can help prevent stigma
- Responsible practice extends beyond legal ownership - honoring emotional, reputational, and narrative boundaries associated with original work helps protect creative confidence
- Contributors benefit when studios acknowledge the ongoing value of human creative input in adaptive systems, not as a replacement layer, but as a collaborative craft
- Legal and ethical scope, including IP ownership and reuse rights, are not identical; maintaining alignment with original intent, tone, persona, emotional context, and reputational fit strengthens trust
- Synthetic performance may be perceived as faster or cheaper, but reductions in originality, innovation, and imagination may result in a homogenous product that is less interesting or attractive to players

> *AI can extend performance, canon, and craft -*
> *but only human contributors make worlds worth extending.*

Industry Opportunities

- The commercial upside of adaptive AI isn't smaller teams - it's richer games, stronger emotional attachment, higher replayability, and expanded audience reach. That's how AI can create value
- Purpose-built, intentional, original work can be amplified rather than replaced - extended into more fully realized worlds, shared with wider audiences, and strengthening contributor identity and visibility
- With the assistance of AI tools, narrative design can become less procedural, and more expansive. Players rarely ask for less authenticity or more repetition – nuance and variation are a design asset
- Richer, deeper worlds increase replayability, enabling branching that is impractical to author manually
- Work opportunities shift rather than disappear – in high intensity adaptive worlds the role of emotional safety experts, narrative guardians, and safety & intervention operations becomes foundational

Creative rights aren't a constraint - the future of distinctive, compelling living worlds depends on human originality and new creative capability.

C. Narrative Integrity & Canon Preservation

Human-designed, canon-constrained, drift-resistant by design.

Principle Goal

Ensure AI strengthens worlds and games rather than overwrites them - enabling adaptive storytelling without sacrificing authored narrative intent. Creators should be able to define the level of improvisation allowed, and within those boundaries, players could choose the style of experience that feels right for them.

Indicative Human Impact

If human-defined parameters aren't clearly established or enforced, AI-enabled narrative systems can evolve in ways that are contrary to creator intent and player expectations - eroding trust and creating risk.

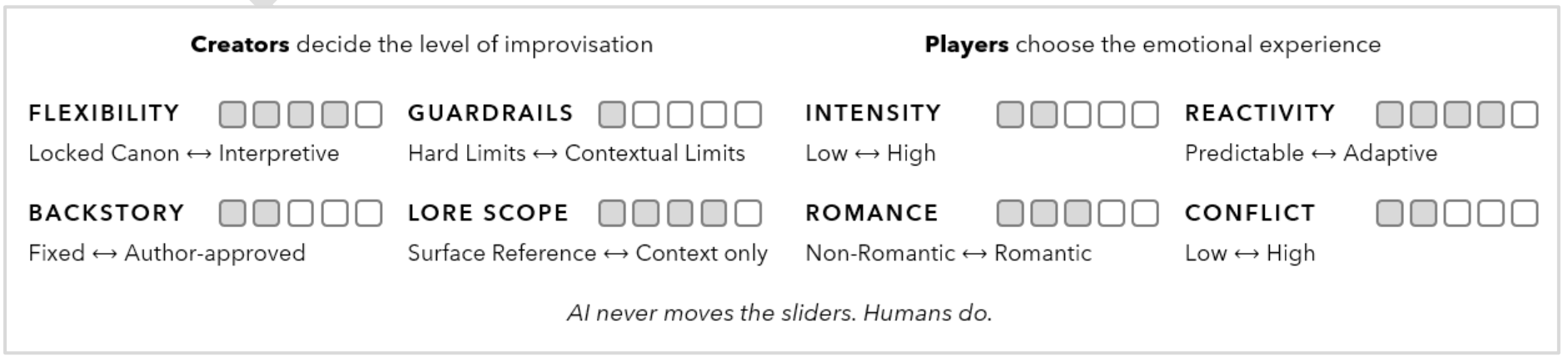
When creators retain narrative authority and can specify the degree of improvisation allowed, players can engage with AI-enabled systems confidently. Canon, tone, and authorial intent stay intact; the world feels coherent and reliable, and structured control functions as a safety system - not just a storytelling device.

Responsible Design Considerations

- The role and scope of the AI should be creator-directed, with models operating inside clearly specified narrative, emotional, and mechanical constraints
- Human-authored narrative and boundaries don't just preserve canon – they also protect emotional tone, player safety, and cultural authority where narratives draw on lived tradition
- When the degree of emergence is intentional, AI acts as a narrative companion rather than narrative authority – enabling variation without volatility or drift
- Within a stable canon and tone, creators define where improvisation is allowed - and inside those ranges, players could choose the style of experience that feels right for them, where appropriate

Industry Opportunities

- By giving creators and players levers to define autonomy, AI can adapt the experience while still operating inside the same world rules
- Clear canon isn't a constraint - it becomes the backbone that makes adaptation sustainable across sequels, DLC, mods, and live updates
- That's where the industry gains something genuinely new - the ability to let players explore different flavors of the same world, without multiplying production cost or fracturing canon. For example:



AI shouldn't rewrite the world - it should create new experiences grounded in the one that already exists.

D. Transparency & Trust

Transparency as design - articulated impacts, individual understanding, meaningful choice.

Principle Goal

Ensure studios, creators, performers, and players all feel informed, respected, and in control when interacting with systems. Transparency should support clear understanding of what the AI does, and allow each individual to choose how deeply they engage, aligned with their level of comfort and expectations.

Indicative Human Impact

If the intent, boundaries, risks, and benefits of using AI in a game are unclear to stakeholders, later disclosure can feel like a breach of trust, leaving them feeling misled, taken advantage of, or exposed.

When proactive clarity replaces suspicion with mutual understanding, it enables informed consideration, meaningful choice, and genuine individual autonomy. Transparency safeguards safety and rights, protecting everyone involved. Feeling "part of it" not "subject to it" reinforces agency and respect.

Responsible Design Considerations

Committing to transparency and alignment with established societal and regulatory expectations collectively works towards the same thing - socially legible trust built in plain sight.

INTENT + BOUNDARIES + IMPACT + DATA USE + CHANGE + CONTROL + EXPECTATIONS

- State the intent. Explain why AI is being used so assumptions or suspicion don't fill the gaps
- Define the limits. Commit to boundaries to avoid fear of invisible changes or silent reinterpretation.
- Describe the impact. Share likely risks, benefits, and trade-offs early to enable considered choice
- Explain how data and creative assets are used. Be explicit about assets, identity data, inputs, and analytics to support mutual understanding and protect autonomy
- Signal change before it happens. If scope evolves, communicate early - it helps preserve trust.
- Preserve agency. Offer options for levels of participation, aligned to comfort and expectations
- Set realistic expectations. Don't oversell or undersell the AI - protect the industry's credibility

Transparency isn't a binary action and outcome
- it's the sum of every moment the system chooses to be clear.

Industry Opportunities

- Transparency strengthens internal design clarity. Clearly defining what the AI does prompts sharper thinking about scope, emotional impact, boundaries, and intent, and can make feedback more precise so teams learn faster.
- This clarity elevates team alignment - narrative, design, engineering, and performance all operate from the same understanding, reducing rework and avoiding costly "AI drift" later.
- Visible expectations mean less confusion and fewer escalations - saving time and energy for everyone.
- Trust isn't "soft" – it's a valuable asset. Clarity reduces reputational risk, and transparency is almost always cheaper than crisis management - especially as markets begin to punish opacity.

When transparency is treated as design - not fine print - trust becomes part of the system, not the marketing.

E. Empowerment Through Technology

Lowers barriers, unlocks possibilities, extends human capability.

Principle Goal

Ensure AI unlocks opportunities for existing and new creators by widening access to more advanced tools that augment and uplift human capability, not replace it. AI should create space for skills to evolve and reduce repetitive work, so people can focus on deeper craft, creativity, learning, and collaborative problem-solving.

Indicative Human Impact

If AI is used to bypass original design and craft rather than support them, it may create over-reliance on AI tools, erode team skills, flatten creative outcomes, and push invisible labor to cleanup and QA.

When AI is used to accelerate iteration and enable safe experimentation, it can free individuals to work at deeper levels and grow their skills into new areas. It can also lower barriers to participation, helping smaller teams to ship more ambitious work, and opening space for new contributors and roles across the ecosystem.

Responsible Design Considerations

- Responsible design considers how AI reshapes creative work, power dynamics, and tool dependence
- Tier 4 AI should be considered assistive by default – it can help and support work, but should not drive it
- The final say in the process should remain with humans – for quality, safety, and as a fail-safe
- Third-party tools should meet the same standards as internal ones, with clear scrutiny over IP, data storage, and sharing. Portability matters too - studios shouldn't be locked in for a game's entire lifecycle
- Time saved through AI investment should benefit humans, not just margins - creating space for learning, mentoring, and intentional experimentation that prepare for the next generation of AI games

Industry Opportunities

- Broader access to AI tools brings new storytellers, increased global talent, and culturally richer games.
- Smaller teams can deliver polish and scope that used to be out of reach
- Rapid prototyping and cheaper experimentation de-risk ideas, speed iteration, and raise overall creative quality

As AI enters live narrative pipelines, studios don't just add tools - they inherit new responsibilities around safety, coherence, player trust, and ethics.

ONCE A CHOICE IS MADE TO INTRODUCE AI INTO LIVE NARRATIVE SYSTEMS, NEW CATEGORIES OF WORK APPEAR. FOR EXAMPLE:

PLANNING	REQUIREMENTS	DESIGN	BUILD	TESTING	LAUNCH	MAINTENANCE
Outline intent & purpose of AI use	Define boundaries & identify risks	Expectations for social license & rights	Implement guardrails & contributor protections	Verify transparency, tone, remediations	Disclose clearly & empower choice	Monitor proactively & manage change
AI intent & purpose	Boundary & risk definitions	Safety pattern design	Contributor protections	Behavioral QA	Proportionate comms	Escalation & intervention
Player value proposition	Policy frameworks	Canon constraints	Traceability auditing	Hallucination & drift review	Clear education	Monitoring & drift detection
Ethical intent statement	Sovereignty rules	Interaction design	Prompts & tuning	Red-team scenarios	Consent & opt-out processes	Responsible change mgmt

5. Reference Group for Responsible AI in Games

Further information, feedback channels, and expressions of interest registration can be found at vinebrightfoundry.com/five-five-reference-group.

5.1 Proposal

Vinebright Foundry is proposing the formation of an international Reference Group for Responsible AI in Games to support shared discussion and emerging practice as systems become more common in player-facing experiences.

At present, there is no shared, cross-disciplinary reference point for the industry to help navigate emerging risks. In the absence of guidance, individuals or teams are left to interpret responsible practice in isolation - increasing the likelihood of inconsistent approaches and avoidable breaches of player trust that could affect the credibility of the technology as a whole.

The proposed Reference Group is intended to operate as a voluntary coalition of studios, researchers, creators, and players who share an interest in shaping responsible practice in games. Participation would reflect a commitment to collective stewardship rather than authority or enforcement.

Part II: Vinebright Implementation

This section reflects one applied approach, shaped by our goals, constraints, and values as a studio building a living world.

6. Responsible AI Applied Practice

More information about Vinebright's approach to applying the Five + Five framework is available at: vinebrightfoundry.com/responsible-ai-applied-practice.

6.1 From Framework to Practice

The Five + Five framework sets out how we think about human impact, responsibility, and social license in AI-enabled games.

The following sections form 6–10 of our Responsible AI work, showing how that framework is applied in practice through Helix and *Of Moss & Moonlight*. It reflects one implementation, shaped by context and constraints.

6.2 What Applied Practice Means at Vinebright

Our applied approach operates across two connected layers.

The first is Helix, Vinebright's living world intelligence layer. Helix provides the enforceable capability layer behind responsible AI use in our worlds, through runtime mechanisms that make boundaries visible and actionable in practice.

The second is *Of Moss & Moonlight*, our flagship living world RPG. This is where those capabilities are configured into player-facing commitments. Helix provides the infrastructure; *Of Moss & Moonlight* is where responsibility is experienced.

6.3 Designing for Responsibility

Achieving long-term success and realizing the potential benefits of AI technology depends on responsible practice and proactive stakeholder engagement - including obtaining and maintaining social license - being treated with the same priority as technical implementation.

A responsible approach requires emotional safety by design. Consent by design. Narrative integrity by design. These are fundamental social-license building blocks for AI use, not activities that can be deferred to launch or post-launch communication.

Vinebright operates in the higher-impact levels today (Levels 3–4), where expectations for safety, consent, and accountability are significantly stronger. Everything that follows builds on that core principle - higher human impact requires higher public and industry social license.

7. Helix Live Brain™ Responsible AI Capabilities

The full detail of the Helix Responsible AI Capabilities is available at vinebrightfoundry.com/responsible-ai-helix.

7.1 What is Helix?

Helix is an AI-assisted runtime for living worlds - an orchestration layer that enables characters and worlds to respond and evolve over time.

Helix does not “write” games. It operates between human-authored intent - characters, arcs, tone, memories, and rules - the evolving game state that tracks what has happened, and AI-extended delivery inside a living world.

While Helix was developed alongside *Of Moss & Moonlight*, it is designed to run across engines and power any living world experience.

Helix does not confer social license. It does ensure living world AI remains bounded, observable, and manageable over time. It preserves authored boundaries, enables studio-defined tuning, and supports accountability through traceability and intervention pathways.

More information about Helix can be found at vinebrightfoundry.com/helix.

7.2 How Helix Builds Responsible Living Worlds

Helix is designed primarily for Human Impact Levels 3 (Systemic Reactive AI) and 4 (Player/Social Interactive AI), with potential to evolve toward Level 5 systems (High Intensity Interactive AI). At these levels, the risks of using AI are not only technical - they are emotional, relational, and trust-based.

The sections that follow outline how Helix addresses these risks by design across three capability layers - how living worlds are scoped, how behavior is made observable, and how accountability is sustained over time.

Each of the 18 Helix Capabilities is uniquely identified to support traceability, review, and accountable ownership over time, and reflects our experience building a living world RPG.

i. Authored Scope & Player Safeguards

These capabilities exist so designers do not have to invent safety patterns from scratch.

HLX -01

EXPLICIT SCOPE

Defined boundaries for what AI may and may not do

Makes AI scope explicit and enforceable through built-in constraints across canon, safety, and runtime, with studio-defined thresholds and tuning.

HLX -02

CONSENT STATE ENFORCEMENT

Gated withdrawal, progressive, and revocable states

Enforces consent states at runtime, including game-level consent gating, and support for withdrawal and data controls.

HLX -03

DE-ESCALATION CONTROLS

Configurable pacing, cooldown, and tone reset

Includes runtime regulation controls to throttle escalation, apply cooldowns, and reset tone - supporting safe redirects when needed.

HLX -04

PLAYER AGENCY PROTECTION

No coercion, trapping, or forced escalation

Blocks coercive patterns so AI cannot trap or guilt players, or force escalation beyond player choice - while still supporting authored tension.

HLX -05

ROLLBACK & RECOVERY

Safe resets, reversals, and feature retirement

Enables rollback and recovery mechanisms - e.g. memory resets, state reversals, and retiring risky behaviors - without breaking continuity.

HLX -06

CONSENT & OPT-IN UX

Comprehensible, player-configurable boundaries

Enables studios to present consent and boundary settings in UX (e.g. exit controls, relationship modes, language, intensity boundaries, memory limits).

HLX -09

DATA & IP OWNERSHIP

World state, memory, and canon remain first-party assets

Keeps conversation state, memory, and truth sources as engine-level assets - no centralization, harvesting, or vendor-owned persistence.

HLX -17

NO LOCK-IN ARCHITECTURE

Interchangeable models, vendors, and services

Uses modular adapters and first-party runtime assets so studios can choose models, TTS, analytics, and storage providers; and change providers easily.

ii. Observability & Auditability

These capabilities exist so studios can see what is happening before players feel it.

HLX -07

RUNTIME SAFETY ALERTING

Drift, boundary, and anomaly monitoring

Detects and flags behavioral anomalies in live systems (e.g. hallucinations, drift, boundary crossings) to support timely human review and intervention.

HLX -08

TRANSPARENCY & TRACING

Runtime decision trails & auditability

Provides structured tracing across configuration, boundaries, guardrails, and model behavior - enabling outcome auditability, and player disclosure.

HLX -12

CHANGE VISIBILITY

Designed to prevent silent scope expansion

Provides platform-level transparency (e.g. release notes, logs, behavioral-change notices) to stakeholders so modifications are never silent.

HLX -10

PROVENANCE METADATA

Origin, attribution, permissions, and usage constraints travel with assets

Captures origin, authorship, attribution, and usage constraints for creative and performance assets - enabling traceability, credits, and rights-safe use.

HLX -11

CANON GOVERNANCE

Versioned canon sources and authorship

Provides governed canon sources with version control, traceability, and access rules to ensure models use approved truth, not invent or overwrite it.

HLX -18

HUMAN OVERRIDE ABILITY

Pause, throttle, block, rollback, disable behaviors

Provides intervention controls for live ops - enabling pause, rate-limit, block, rollback, or retire behaviors when risk, drift, or harm is detected.

iii. Oversight & Operations

These capabilities exist so intervention does not require designing fixes or escalations mid-crisis.

HLX -13

REPORTING & ESCALATION

In-game incident reporting & intervention

Provides structured reporting and escalation, including in-game incident flagging, severity classification, and intervention controls.

HLX -14

RELEASE OVERSIGHT

Human-impacting change gates

Routes human impacting changes through elevated testing and validation paths, including independent ethical review where required.

HLX -15

AGGREGATED OBSERVABILITY

Privacy-preserving monitoring by default

Captures anonymized, aggregated trend metrics for system health and tuning - with no individual player tracking. Helix watches systems, not players.

HLX -16

OPTIMIZATION SUSTAINABILITY

Efficient, cost-aware, and resource-conscious runtime

Reduces unnecessary inference through routing, caching, batching, and memory discipline - with built-in controls for cost, performance, and environmental footprint.

Together, these capabilities define Helix as an accountable runtime for living worlds - bounded by design, observable in practice, and manageable over time.

The next section outlines how Vinebright applies and configures these capabilities in *Of Moss & Moonlight* through player-facing commitments, community practice, and named accountability ownership.

8. Commitments for *Of Moss & Moonlight*

Further information about *Of Moss & Moonlight's* Responsible AI Commitments is available at vinebrightfoundry.com/responsible-ai-omm.

8.1 What is *Of Moss & Moonlight*?

Of Moss & Moonlight is a living world where we apply the Five + Five framework in practice. It brings together reactive characters, evolving relationships, and a world that adapts over time.

Helix Live Brain™ provides responsible AI baseline runtime behaviors and configurable capabilities. *Of Moss & Moonlight* defines the configuration, thresholds, and player-facing consent design appropriate for a Level-4 Human Impact experience.

Learn more about *Of Moss & Moonlight at* vinebrightfoundry.com/of-moss-and-moonlight.

8.2 Evolving The World With Our Players

In a living world, players aren't just "consumers". They're participants - and their experiences, boundaries, and feedback help shape where we take the game. Players are a key part of the conversation about what feels good, what feels off, and where the guardrails aren't right.

> *If there's a problem, we repair trust – not just patch systems.*

Positive responses are important too, helping us preserve the moments that resonate, rather than unintentionally flattening them.

Feedback isn't just bug reporting. It's part of the design and evolution of our living world, and that responsibility continues long after launch.

8.3 Commitments Overview

The Five Principles for Responsible AI in Games aren't met by technical delivery alone. They depend on clear intent, transparent communication, and practices that earn trust and maintain social license over time.

COMMUNITY PROMISE + RUNTIME ENFORCEMENT + TESTING & OPERATIONS + INTERNAL OWNERSHIP

= VINEBRIGHT COMMITMENT

We deliberately use the term commitment when talking about how we ensure responsible use of AI in *Of Moss & Moonlight*. A commitment is not a goal. It is a promise backed by runtime enforcement, testing and operations, and named internal ownership.

The *Of Moss & Moonlight* Commitments below focus on design choices specific to the game - where we tighten Helix defaults, increase transparency, and add safeguards beyond baseline.

[View a full-size copy of the Commitments Map]

8.4 Commitments Detail

What follows is the practical application of these Principles in *Of Moss & Moonlight*. Each Principle is mapped to specific Commitments, and includes Helix support references where enforcement or technical safeguards apply.

A Player Safety & Respect

OMM-A01: We're upfront about where and how AI works in the world - before and during play

(HLX-12) Design *Lead - Player Transparency & Trust UX*

Players are told where AI is used, what it does, and what it doesn't do. We don't introduce hidden systems, silent changes, or unexpected behavior.

Trust breaks when systems feel sneaky. Transparency by design keeps the social license with our community intact.

OMM-A02: High-intensity and adaptive AI is always opt-in - and always adjustable

(HLX-02, HLX-06) *Design Lead - Consent & Intensity UX*

Players explicitly opt into high-intensity content and AI-enabled systems. Consent is checked at runtime and revisited at any time.

Consent must meaningfully shape the experience - not just exist in settings. Safety and agency come first.

OMM-A03: Players can always opt out - without being punished

(HLX-02, HLX-06, HLX-18) Design Lead - Consent & Intensity UX

Choosing lower-intensity AI never reduces progression, rewards, or core content. Players can change or withdraw consent at any time.

Player control shouldn't come with penalties or repercussions.

OMM-A04: We make sure players can step back instantly at any time

(HLX-02, HLX-06, HLX-18) *Design Lead - Emotional Safety*

Pause, skip, disengage, or freeze AI-enabled moments the second something feels uncomfortable - without losing progress.

When players know they can stop a moment immediately, they feel braver exploring emotionally rich worlds.

OMM-A05: We design reporting and escalation into the experience - before we need it

(HLX-13, HLX-18, HLX-14, HLX-09) Ops Lead - Incident Response & Community Trust

Players can flag moments, lines, NPC interactions, or scenes directly in-game. Reports flow into clear escalation paths with defined severity, ownership, and action.

Low friction → faster signal. Issues are surfaced early, before harm compounds.

OMM-A06: We commit to listening to our community

(HLX-13) *Ops Lead - Incident Response & Community Trust*

Clear channels for concerns, confusion, emotional friction - with visible responses and follow through.

Feedback isn't noise - it's crucial. Listening keeps the world aligned with the people inside it.

OMM-A07: We make consent by design part of core gameplay

(HLX-02, HLX-06) *Design Lead - Consent & Intensity UX*

Opt-ins, intensity changes, resets, pauses, and exits are tested repeatedly - just like combat, crafting, or dialogue.

Consent must work, every time, under pressure - otherwise players lose agency exactly when they need it most.

OMM-A08: We monitor the world - not individual players

(HLX-15, HLX-16) Engineering Lead - Data Protection & Privacy

We monitor live behavior using anonymized, aggregated data only. We do not build systems for surveillance, profiling, or identity-linked tracking.

Living worlds need oversight - but players are not data subjects. Monitoring should make the world safer without normalizing surveillance, profiling, or inferred psychological categorization.

OMM-A09: We test for real player impact - not just technical correctness.

Runtime Lead - Safety & Intervention

We deliberately stress-test grief scenes, romance edges, power-imbalanced moments, consent failures, canon drift, and emergent behavior - not just whether things technically "work."

Players experience worlds, not code. If something feels unsafe, confusing, or off-tone, that can matter more than a crash - and it deserves the same level of attention.

B Creative & Performer Rights

OMM-B01: We design AI as augmentation within human-authored worlds.

(HLX-19) Governance Lead - Ethics & Player Trust

We use AI as augmentation and imagination but our humans always choose, shape, and finalize.

Protects authorship, ensures originality, and ensures the integrity of the world we want to build.

OMM-B02: Creator and performer consent is explicit, scoped, and respected.

(HLX-10) Governance Lead - Studio Policy & Risk

Contributor agreements clearly define scope, usage, and boundaries - including how assets may be reused, adapted, or trained into tools. Consent can be limited, revised, or withdrawn.

People deserve control over their work and their likeness. Trust in Vinebright beats shortcuts every time.

OMM-B03: We enforce consent through rights metadata.

(HLX-10) Governance Lead - Studio Policy & Risk

Permissions, attribution, and usage constraints are attached to creative and performance assets, so consent and authorship travel with the work.

When rights are embedded in the pipeline, consent isn't a policy document - it's operational.

OMM-B04: We build visibility into AI behavior.

(HLX-08, HLX-15) Engineering Lead - Tooling & Observability

We add anonymized logs, dashboards, and traces so teams can see what the AI did, when, and why.

Visibility allows us to detect misrepresentation, tone distortion, or behavior that could reasonably harm contributor intent or reputation.

C Narrative Integrity & Canon Preservation

OMM-C01: We make sure our canon is always overseen and traceable.

(HLX-11, HLX-07) Canon Owner - Lore and Continuity

Canon truth sources are versioned, access-controlled, and enforced so AI cannot invent or overwrite lore.

Traceability keeps story, intent, and lore coherent.

OMM-C02: We always verify canon integrity under pressure, before release.

(HLX-11, HLX-07) Canon Owner - Lore and Continuity

We push the AI into strange corners on purpose to ensure it never significantly contradicts lore, tone, or character voice.

Canon is the spine of the world. If the AI starts bending it, the story stops feeling authored, intentional, and ours.

OMM-C03: We keep watching for canon & tone drift in live systems.

(HLX-11, HLX-07) Design Lead - Narrative Integrity

If AI bends lore, changes character voices, or warps tone, we act quickly so the world stays aligned with what our humans actually created.

Stories are fragile. If AI bends characters out of shape, trust cracks.

OMM-C04: We keep the boundary between authored truth and adaptive expression clear.

(HLX-08) Design Lead - Player Transparency and Trust UX

Players understand what is fixed canon and what is flexible or adaptive through consistent world design and clear framing.

It keeps expectations grounded. Players know what's canon, what's flexible, and where surprises live - which protects trust and the story.

D Transparency & Trust

OMM-D01: We begin with why, not “because we can.”

Governance Lead - Ethics & Player Trust

AI is only used when it has a true purpose and makes the world feel more alive.

Keeps us honest. No gimmicks, no seeking tech credit, and no bait and switch on players.

OMM-D02: We set hard boundaries early - and stick to them.

(HLX19, HLX10, HLX01, HLX03) Governance Lead - Studio Policy & Risk

From day one, some things are simply off-limits - e.g. cloning voices without consent, manipulative emotional mechanics, and opaque data use.

Declaring boundaries upfront stops us being tempted by “just this once” compromises down the track.

OMM-D03: We refuse to manipulate players “for engagement.”

(HLX-04, HLX-03) Design Lead - Player Transparency & Trust UX

We don’t do guilt scripts. FOMO pressure. Addiction or mental health mechanics tuned to keep you hooked.

We want players to play our game because they love the world, not because we engineered pressure.

OMM-D04: We use plain language to explain our use of AI - and what it means for players.

(HLX-08, HLX-12) Design Lead - Player Transparency and Trust UX

Players are told where AI is used, what it does, and what it doesn’t do - in clear, human language, before and during play. We explain risks, benefits, and trade-offs without hype or spin.

Clarity builds trust. If players can’t understand it, they can’t consent to it. We want players to be able to make informed choices.

OMM-D05: We repair in the open, and we learn from it.

(HLX-14, HLX-12) Ops Lead - Incident Response & Community Trust

When something goes wrong, we acknowledge it, explain why, and fix it. We use rollback or resets as needed, and treat consent defects as Sev-1.

Trust comes from honesty and repair, not perfection. When we learn something useful, we act on it.

OMM-D06: We treat offsets honestly.

(HLX-16, HLX-17) Engineering Lead - Sustainability Optimization

We give teams visibility into environmental cost so they can seek to reduce first, and make informed trade-offs.

Keeps us honest about impact instead of buying our way out of responsibility.

OMM-D07: **We're transparent about our footprint.**

(HLX-08, HLX-12, HLX-16, HLX-17) Engineering Lead - Sustainability Optimization

We commit to publicly sharing 6 monthly estimates of AI interactions, model mix, energy use, optimization wins, and mitigation strategies - in plain language.

We want to acknowledge all the consequences of using AI, and to be kept accountable for improving wherever we can.

E Empowerment Through Technology

OMM-E01: **We support our creators over time.**

(HLX-10, HLX-19) Engineering Lead - Tooling & Observability

We provide ongoing training, guidance, and safeguards - so working with AI strengthens craft not replaces it.

Valued contributors make better art, and we build more resilient teams. Long-term creative wellbeing matters as much as features.

OMM-E02: **We prioritize vendors who share our values.**

(HLX-11, HLX-07) Engineering Lead - Sustainability Optimization

We favor providers who disclose energy profiles, support greener routing, and don't lock us in.

Aligns infrastructure with ethics - and keeps long-term resilience in our hands, not theirs.

OMM-E03: **We design tools that keep humans in control.**

(HLX-07, HLX-14, HLX-05) Engineering Lead - Helix Runtime

We build tools that let humans override (pause, throttle, rollback, retire), shape, and steer AI behavior, to extend creative capability without replacing judgment or intent.

Creators should feel more capable - not more dependent, locked out, or second-guessed by the system.

OMM-E04: **We reduce lock-in so our world can evolve responsibly.**

(HLX-20, HLX-17, HLX-16) Engineering Lead - Sustainability Optimization

We avoid brittle dependencies by favoring adapters, portable configurations, and vendor flexibility, so our AI stack can evolve without compromising values or continuity.

Lock-in turns choices into traps. Flexibility protects long-term creative freedom - and keeps responsible decisions possible.

---

9. Roles & Accountabilities

Accountability is not abstract or informal - it is assigned, supported by Helix runtime controls, and reinforced through governance, design, engineering, and live operations.

Each Commitment has an Accountable Owner responsible for delivery, oversight, and intervention where required. Together, these roles ensure that responsibility for AI behavior in *Of Moss & Moonlight* is explicit, owned, and reviewable.

These roles are capability-based - one person may hold multiple roles, and some Commitments may be supported by external advisors. Each capability has a named owner, even when responsibilities are shared.

A full list of the role types that are likely to be required can be found at vinebrightfoundry.com/responsible-ai-roles-accountabilities.

10. Applied Practice and Open Questions

This framework does not claim to resolve all ethical, psychological, or regulatory questions associated with AI in games. It does not substitute for legal compliance, platform governance, or player-specific safeguards, nor does it prescribe a single “correct” implementation.

Our implementation surfaced real tensions that no framework can fully resolve, and other questions remain open.

These questions do not have universal answers. Appropriate responses depend on genre, audience, culture, art direction, and community norms. What is proportionate in one context may be unnecessary, or insufficient, in another. Studios may draw on this framework in whole, in part, or not at all.

This work is continuous, contextual, and shared.

11. References

11.1 Context and Adjacent Work

These works reflect adjacent research and discussion across AI ethics, governance, and player experience. They primarily explore what responsible AI in games should consider. The Five + Five framework focuses on how those considerations are translated into concrete design, system constraints, and runtime practice.

AI Ethics & Governance

World Health Organization. *Ethics and Governance of Artificial Intelligence for Health* (2021)

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. *Trustworthy Artificial Intelligence: A Review* (2022)

Moffat, K., Lacey, J., Zhang, A., & Leipold, S. *The Social Licence to Operate: A Critical Review* (2016)

Games & Player Experience

Cook, M. *The Social Responsibility of Game AI* (2021). IEEE Conference on Games.

Melhart, D., Togelius, J., Mikkelsen, B., Holmgård, C., & Yannakakis, G. N. *The Ethics of AI in Games* (2023).

Mikkelsen, B., Holmgård, C., & Yannakakis, G. N. *Ethical Considerations for Player Modeling* (2017).

European Parliament. *Consumer Protection in Online Video Games: A European Challenge* (2023).

Industry/Practice Signals

Madary, M., & Metzinger, T. K. Real Virtuality: *A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology* (2016).

World Health Organization & International Telecommunication Union. *Global Standard for Safe Listening in Video Gameplay and Esports* (2025).

Critical/Alternative Perspectives

Thiebes, S., Lins, S., & Sunyaev, A. *Trustworthy Artificial Intelligence: A Review* (2021).

Mueller, B. *It's Just Distributed Computing: Rethinking AI Governance* (2024).