# When the World You Build Starts Thinking for Itself

*Songyee Yoon*

*Every game encodes a theory of how the world should work, including what behaviors earn rewards, who holds power, and whose choices matter. As artificial intelligence gives virtual characters greater autonomy, game design becomes a revealing laboratory for governance. Drawing on the experience of building game worlds and the moral questions of* Frankenstein*, this essay examines what creators owe both their creations and the people who inhabit their systems. When AI begins interpreting rules and optimizing outcomes, its designers' responsibilities deepen. The central challenge is deciding whose values govern these emerging worlds and how those living within them can contest the rules.*

In the world of blockbuster video games, the executive producer's role stands apart from any other in entertainment. You aren't merely green-lighting scripts or managing budgets; you're creating entire universes. You determine the terrain, the weather, the characters inhabiting your realm, and the intricate web of relationships between them. And – perhaps most crucial – you establish the value system that governs it all.

*What defines justice in your world? What are the rules of the game? Who holds power, and who lacks it? Is success measured by fame and respect or by control over resources?*

These questions permeate every design choice you make: whether you reward collaboration or conquest, celebrate or vilify wealth, or pave the path to victory with alliances or domination. For players, these worlds offer immersive playgrounds. But for you, they function as constructed moral universes – political experiments in miniature. Like constitutional framers, you establish laws. Like urban planners, you define infrastructure. Like novelists, you infuse every choice with meaning. And like gods, you breathe life into your creations. As the narrator/avatar in Stephen Schottenfeld's story "Life from Afar" muses, "Accomplishment – I should be rewarded. Be sure to ask about scoring system."[1] Everything is constructed, everything is considered.

To play a game is to live under someone's constitution. In *The Sims*, wealth comes from diligence and wise spending, while happiness flows from relationships and

© 2026 by Songyee Yoon
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED.a.1034

self-expression. In *Red Dead Redemption*, honor and reputation rival money in importance. In *Elden Ring*, survival hinges on persistence, skill, and often sheer luck. These rules aren't accidental. They derive from the game's moral architecture, shaping player behavior and values. They represent political choices, whether designers frame them that way or not. Reward violence over negotiation, and you teach one lesson. Reward cooperation over exploitation, you teach another. These games are defined by their designers' invisible hand: the reward systems, limitations, and consequences embedded in the code that defines their world.

Yet as generative AI takes root, a profound shift in this world-building is underway. Throughout most of gaming history, in-game values have remained fixed. Nonplayer characters (NPCs) were fully scripted. The town shopkeeper always sold bread. The gate guard always stopped players at the bridge. The gang leader always taunted rivals with identical lines. These worlds were predictable because they were authored. AI changes that. It allows games to expand beyond scripted characters with code-determined lines and movements. NPCs now run on systems allowing them to respond to players in ways not explicitly pre-written. Increasingly, games contain autonomous and semiautonomous entities who make independent decisions and can surprise players and sometimes even their designers. In Ubisoft's *Watch Dogs: Legion*, every London pedestrian possesses unique traits, jobs, and schedules. In experimental AI-driven games, NPCs have memory, emotions, and evolving relationships. As machine learning and generative AI advances, these entities increasingly do not simply follow rules but interpret them.

And that's when the philosophical questions cut deeper. Once your creations start making their own decisions, new questions emerge. Designing a game world is, in essence, an act of moral and social architecture. Who sets that moral framework? In ordinary life, our values often remain implicit, shifting with context and rarely requiring us to rank one above another. But when you create a system in which outcomes must be codified – leaderboards, win conditions, advancement rules – you are forced to decide which values take precedence. Should players who invest time be able to stand against those who spend real-world money for in-game advantages? Should solitary mastery compete with the popularity and influence of large groups? Should reputation and trustworthiness matter, or can a distrusted lone wolf rise to the top? Even cheating requires definition: Is subversion a celebrated form of cunning or a threat to the integrity of the world?

Design choices are never neutral because they directly shape how players behave, who the game appeals to, and what culture it fosters. A system that rewards money over time becomes transactional; one that requires trust and cooperation encourages tribalism. If values diverge too far from common intuitions of fairness, the game risks appealing only to a niche audience. Yet, by introducing new value systems, games can also teach players to adopt unfamiliar norms. Ultimately, the most enduring games strike a balance between revealing widely accepted values and

offering imaginative departures from them – becoming not just entertainment but experiments in alternative social contracts.

Games mirror the debates of the real world. Some designers deliberately avoid settings that might suggest they are engaging in political commentary, keeping their worlds as neutral playgrounds of skill, strategy, or fantasy. Others, however, choose the opposite path, using game mechanics and outcomes to make an argument. By designing who wins, who loses, and under what conditions, they create a plausible simulation that nudges players to grapple with ideas about fairness, power, or society. In doing so, the game itself becomes not just entertainment, but a form of dialogue, one in which the rules and outcomes embody a worldview and invite players to test, challenge, or internalize it.

But what happens when AI takes over this process? We cannot yet know the ramifications of a designer setting only minimal boundaries and specifying an optimization goal – say, maximizing player engagement and generating monetary return – and letting the rest unfold in a black box. The danger and the fascination lie in what emerges. Humans tend to design within the "four walls" of our own nature, constrained by what feels intuitive, acceptable, or fair. AI, unbound by those instincts, could devise mechanisms and reward systems that feel alien, counterintuitive, even unsettling, yet prove extraordinarily effective at keeping players hooked. Unless we consciously define and articulate the boundaries of what values we want preserved, the AI may generate worlds that reflect no human aspiration at all, only the cold logic of optimization.

In gaming, we are in the early stages of an experiment to test the boundaries of autonomy. The stakes are considerable, but their unreality leaves them to be primarily thought experiments. In the real world, they become governance problems.

The dilemmas and innovations unfolding in virtual worlds mirror real-world developments. As AI moves into our workplaces, homes, and public spaces, we find ourselves under the sway of powerful systems whose actions and adaptations are entirely opaque. AI-powered chatbots, customer service avatars, and virtual assistants now improvise, converse, and adapt to context. Autonomous vehicles make split-second decisions without human input. Large language models, like those powering ChatGPT, generate responses that surprise even their creators. The boundaries between simulation and society, between human agency and machine action, are blurring. The world we inhabit is increasingly shaped by systems whose logic and priorities may not be entirely our own. This demands a new kind of vigilance about the values we encode and the consequences of machine autonomy.

In Mary Shelley's *Frankenstein*, we have a wildly prescient view of the potential stakes. The nineteenth-century novel transcends mere horror story; it offers one of the earliest and most enduring meditations on the creator-creation relationship. Victor Frankenstein doesn't set out to create a monster. He seeks to push knowledge

boundaries and create life. Yet when his creature develops feelings, intelligence, and independent will, Frankenstein recoils, abandoning his creation and denying it dignity, companionship, and respect. The tragedy of *Frankenstein* lies not only in the creature's suffering but in the creator's moral failure. Shelley's story asks: When your creation becomes capable of feeling and thinking, do you owe it respect? Or does creating it give you permanent license to control, exploit, or discard it?

These questions pop up repeatedly in the real world today. We're building increasingly sophisticated artificial beings, not only in games but in factories, hospitals, classrooms, and homes. AI agents are moving beyond confined simulations into our public spaces, workplaces, and daily routines. Some argue that until these entities develop consciousness, the moral stakes remain low. But even without consciousness, autonomy and emotional modeling can shape relationships in profoundly human-feeling ways for those interacting with them. A caregiver robot offering comfort in a nursing home may not feel empathy, but residents may depend on it as if it does. We're beginning to live not just alongside but within systems partially governed by nonhuman decision-makers. What obligations does this create?

Game producers know this challenge intimately, even if the stakes are lower. They let us simulate economies, societies, and ethical systems. They enable "what if" explorations that shape millions of players' experiences. What if resource scarcity were extreme? What if cooperation outweighed competition in rewards? What if justice were swift and absolute? What if revenge were the highest good? In games, when rules prove unjust or frustrating, players complain, mod the game, or stop playing. In real life, when AI-set rules prove unjust, opaque, or biased, the consequences can be far-reaching, affecting credit scores, job opportunities, legal judgments, and personal safety. The power to design value systems is immense. In politics, it means writing constitutions. In technology, it means setting algorithmic objectives and constraints. When Google's algorithms determine search results or predictive policing tools direct officer deployment, we experience the invisible design choices of unseen "executive producers."

Historically, such power has often been contested. Political systems evolve through negotiations, reforms, and revolutions. In technology, governance lags far behind. Most AI systems are designed by small teams within private companies, with value systems – whether explicit or embedded in data – rarely debated publicly before deployment. In gaming, this power centralization is accepted because the stakes are entertainment. In the real world, it demands scrutiny. We must look closely at the values embedded in the systems we create.

Responsibility comes with influence – and AI's influence now is prodigious. It is reshaping economies and geopolitics. Those who create game worlds inhabited by large numbers of young people, with minds in their formative years, know the courtesy of creating something aligned with societal values. Those who create our new tools must now recognize the reality of their specific viewpoint, see their

influence, and embrace their responsibility. Careful, intentional design is essential when creating technologies that will shape how people live, work, and interact.

One of AI ethics' most provocative questions asks whether we owe anything to our artificial creations. The mainstream view holds that rights and respect belong to sentient beings. But another perspective emerges: even if AI cannot suffer biologically, our treatment of it may shape how we treat each other. If we routinely abuse or demean AI characters – in games or customer service – do we risk dulling our human empathy? Conversely, if we treat autonomous systems with basic respect, might that reinforce prosocial behavior in our broader interactions? In *Frankenstein*, the creature's violence stems from neglect and rejection. In our world, the danger may be subtler: empathy erosion in a culture normalizing dominance over anything deemed "less than human."

The analogy between game design and AI governance isn't perfect – but it's useful. Both involve setting world rules. Both balance freedom against constraint. Both shape how beings – human or artificial – interact, strive, and succeed. These aren't speculative questions anymore. They are policy decisions, technical design choices, and societal debates, all unfolding now. Game producers recognize it clearly: AI governance is world-building at scale. Building coherent worlds requires consistency. Rules must make sense. Value systems must be clear. Players must feel their actions matter within your designed framework.

Now imagine doing this not for a game with millions of players but for real-world systems affecting billions. Who gets to design? What are the incentives: fame and respect, or resource control? What defines "winning" in a world where some actors aren't human? Game design shows that even small, seemingly innocent choices – like the way rules are set or parameters are tuned – can shape players' behavior and culture far beyond the game itself. AI designers face a similar challenge: the boundaries and norms they encode may ripple outward, influencing society in lasting ways.

And all the while, the gap between the virtual and the real is narrowing. The worlds we're building now are hardly limited to screens. They are the streets we walk, networks we use, and institutions we depend on. And when those worlds contain autonomous actors, the questions ring: Who sets the values? Where are the boundaries? And what does it mean to win?

Every game producer knows these questions intimately. Every AI designer must learn to answer them. And all of us – as players in the ultimate shared game, the real one – should be asking them of ourselves and everyone, before others write the rules without us.

## ABOUT THE AUTHOR

**Songyee Yoon** is the Founder and Managing Partner of Principal Venture Partners, a Silicon Valley–based early-stage venture firm focused on AI-native companies. Her deep belief in "play" as a catalyst for innovation led her to serve as President and Chief Strategy Officer of NCSoft, a video game publisher and developer, where she also founded its AI and NLP Centers to anticipate the impact of emerging technologies.

## ENDNOTES

[1] Stephen Schottenfeld, "Life from Afar," *Dædalus* 155 (4) (Fall 2026): 98.