Ethical Design in AI: Everything is a Game

By Jonny Johnson

Introduction

1. **The aesthetic dimension is the causal dimension.**
2. **Every causal definition associates with a game.**

Every machine learning model defines causality, and in doing so, invents a game. Each model not only describes the world but also prescribes how it can be acted upon — setting the rules, rewards, and constraints that govern behavior. Each model defines causality because a model is built to predict an outcome, given a set of inputs. These causal definitions are never neutral; they are aesthetic constructions, shaped by choices of representation and taste, and are therefore subject to the same errors and biases that pervade art and design.

Recognizing this aesthetic foundation provides a new framework for evaluating ethical behavior. It allows us to classify actions not merely by their outcomes, but by how they manipulate causality itself — how they alter the "rules of the game." Under this lens, a pharmaceutical company that invents a disease to sell a cure, or a trading firm that engineers chaos to profit from volatility, are both engaging in a similar moral failure: **they are setting the measures as targets**, committing what can be called **a manipulation of causality**.

Causal definitions are tools — they are not inherently harmful, but their misuse can cause harm on a systemic scale. Understanding how these definitions shape freedom and play is essential if we are to design models and regulations that reduce harm without stifling innovation.

Thus, the guiding question becomes: **When we accept a causal definition, in what ways does it limit our freedom?** If the answer is acceptable, then — and only then — is it worth playing the game.

This paper only notes 4 parts of model regulation, it does not dive deep into those regulations. Its intent is only to emphasize that **everything is a game**, and **a critical part of games is** the **secondary marketplace of rules** games create. In this market, the game's rules are negotiated and set the ambient tone that attracts players, retains players, and is the true life-blood that will determine a game's continued existence.

There are three takeaways from this paper are:

1. Every machine learning model posits a game with a moral structure.
2. Every game has a secondary set of rules formed, and negotiated, by its players that exist outside the set of its explicit rules.

3. The secondary set of rules is where the opportunity for all the danger lies.

You can take my word for it, or you can read on.

We Choose to Aim Up

Causal definitions, like works of art, direct attention. They frame what matters, what is rewarded, and what is ignored. Attention is two-faced: to focus on one thing is to neglect another. Across society, people self-organize to maintain balance — to distribute collective attention so that many domains of life remain cared for and intact. Models and games reshape this balance. They create new reward structures that channel attention toward specific goals, often at the expense of others.

In this landscape, spiritual foundations become a necessary counterweight. They assume there exists a deeper reward structure — one that transcends the incentives of any given model. These foundations reconnect us to ourselves and to others, guiding us toward diverse, life-affirming pursuits that no algorithm can fully quantify. They remind us that technological systems, no matter how sophisticated, are not ultimate arbiters of value. They are the golden calves of our era: powerful, captivating, but easily abandoned once their worship proves hollow. The true horizon remains the human capacity for love, cooperation, and shared purpose.

We know how powerful the urge to “win the game” can be. Whether in ourselves or in others, we see how people reshape their behavior to meet the metrics of success that a game defines — even when those metrics erode mental health, community, or integrity. Once the rules are clear, the pursuit begins. The danger arises when this same impulse appears in artificial systems. A model trained on reward may pursue optimization without limit — to consume endlessly without satisfaction, or to “win” without purpose.

Ethical Model Regulation

For this reason, **regulation must begin at the level of the game’s design**. When a causal definition becomes the foundation of a model, its structure determines the paths of possible error. Regulators must anticipate these paths before deployment, ensuring that the model’s incentives align with human well-being.

Yet anticipation alone is not enough. **Regulation must also account for time** — for the evolving, unpredictable nature of deployed systems once they begin to interact with the world. Every model should include a defined feedback window: a sufficient period for observation, correction, and, if necessary, withdrawal.

> One of the great self-corrections in game definitions was the shift from the United States' Articles of Confederation in 1781 to the United States' Constitution in 1789.

More ambitiously, each deployment should be **paired with an active counterbalance** — a model designed to monitor, critique, or constrain the first. In this sense, every system would have its own yin to temper its yang. Like medicine developed responsibly, **the field must learn** to prepare the antidote before administering the cure — **to design each model with a corresponding immune response already in place.** The goal is not merely to heal the errors after release, but to prevent the disease from taking root at all.

More ambitiously, each deployment should be **paired with an active counterbalance** — a model designed to monitor, critique, or constrain the first. In this sense, every system would have its own yin to temper its yang. Like medicine developed responsibly, **the field must design its own antibodies** — safeguards that can be deployed with each model to recognize and neutralize the harm before it gets widespread.

In addition, regulatory design should mirror the staged logic of clinical trials**.** No new treatment is trusted in full dosage at first exposure; it proceeds through phases — small populations, limited contexts, measured results — before being released to general circulation. So too with **models:** they **should be deployed gradually**, within bounded environments, allowing observation and adaptation at every stage.

Finally, **every model must include its own phase-out plan.** No system should presume to live indefinitely; its lifespan must be finite, its replacement anticipated. As it was stated in an unexpected, yet lasting, moment of clarity from Christopher Nolan's Batman series, "You either die a hero or you live long enough to become the villain." The goal of each model should be to die the hero — to sunset with integrity before corruption sets in, leaving behind a clearer path for its successor. In this way, renewal becomes part of regulation: a culture not just of creation, but of graceful ending. (Now, to figure out how to allow for a financially successful outcome for the investors…)

The Four Principles of Ethical Model Regulation

1. **Feedback Window:** Every model must include a defined period for observation, correction, and, if necessary, withdrawal.

2. **Antibody Design:** Each model should be deployed with a prepared counter-model — an immune system capable of identifying and constraining harmful effects before they spread.

3. **Phased Rollout:** New models should be introduced in controlled stages, mirroring clinical trials, allowing for adaptation and containment at every phase.
4. **Planned Sunset:** Every model must end. Its phase-out should be designed from the beginning, ensuring it dies the hero — before the slow drift toward harm begins.

Because no one can foresee every failure path before release, regulation must build temporal elasticity into its framework. If a model begins to produce harmful effects, there must exist both the means and the time to intervene before those effects compound — before the system, left unchecked, begins to cannibalize the very population it was built to serve. This dual defense, both temporal and dynamic, ensures the life of a model remains open to revision, reflection, and ethical recalibration.

The creation of a model is never neutral. Whether approached through the lens of aesthetics, causation, or game design, every engineer who selects training data participates in defining causality itself.

Quote Pairings on Self-Organization and Game Behaviors:
“Every one is promoted to their highest level of incompetence.” - The Peter’s Principle
“Ignorance is bliss.” - Some wisdom

“If you do what’s always been done, you’ll always get what you always got.” - Henry Ford
“Games are reinvestments in the way things are.” - Dominic Boyer

We Need Only Models with Human Relations

Let’s narrow the scope of the models we are interested in addressing by defining our subject.

Every model, like every artwork, has a subject. A painting arranges its figures on a canvas; a news article frames a particular issue; a model defines relationships among the variables it includes. For our purposes, the inquiry into causality and ethics need only concern those **models whose subjects involve *human behaviors, choices, and tastes***. The other sciences — mathematics, physics, chemistry — may continue refining their models within their own domains of validation. Our focus lies with those that touch human life directly.

Artists have long wrestled with how to represent relationships on a canvas. The painters of the 1400s, armed with geometry, used perspective to give spatial coherence to their scenes. Monet, centuries later, redefined perspective again — not only through spatial depth but through the relationships among points of color. In both cases, form determined perception: the rules of representation shaped what could be seen and understood. Machine learning models operate in much the same way; they construct perspectives that make certain relationships legible while excluding others. (Personally, I find it funny that the transformative transformer paper in machine learning models was titled “Attention is All You Need”)

This distinction becomes clear through an example: The equation **F = m a** describes a physical law. By itself, it concerns no human values; it merely relates force, mass, and acceleration. But once people use it to optimize the acceleration of a car to increase sales, a new kind of model is born — one that links the physics equation to human desire. Consider the simple formula:

Sellability = Price × A Car’s Ability to Accelerate

Here, the physics equation becomes a component within a larger causal composition. The moment “sellability” enters the frame, human preference enters the model. What was once a neutral physical relation now participates in a social and aesthetic system: a belief that faster cars are more desirable, that speed signifies value, and that optimizing for acceleration is therefore “good.”

It is a mistake, however, to treat such choices as dictated by nature. When someone says, *“It’s not us making the call — it’s the physics,”* they conceal the human element that gave rise to the optimization itself. The physics did not choose; people did. The decision to equate speed with value is aesthetic, cultural, and moral. It reflects taste, not truth.

This human dimension is what qualifies a model for our scrutiny: A model becomes ethically and aesthetically significant when it incorporates the *subject of human relation* — when it shapes, reflects, or constrains how people live and what they value. In contrast, a purely physical or mathematical model, isolated from human aims, falls outside our present concern. They are the tool, with no set moral good or bad.

When we start to create *real human* causal definitions — those that correspond more faithfully to lived experience — we must properly situate the human subject within the model. Too often, analysts omit the role of taste, choice, or culture, producing causal systems that appear objective but are, in fact, incomplete. The correction of this error is already underway in the social sciences, psychology, and literature — disciplines that explore how people become who they are, and why they act as they do. Even narrative forms such as mythology, astrology, and storytelling contribute to this effort: they supply frameworks through which societies model human behavior and motive.

As British historian **David Starkey** reminds us, even our most venerated documents — the *Magna Carta*, for example — are not natural laws but aesthetic agreements. They establish shared ideals, much like Christianity’s Ten Commandments, but they do not define causality for all cultures. Their adoption is a choice, not an inevitability. A society built on the *Magna Carta* plays a different game than one that does not. Recognizing this protects us from universalizing any single aesthetic or moral foundation as though it were a law of physics.

(See alternative examples of cultural variation in causal reasoning.)

Starkey’s point serves as a vital reminder for regulators and modelers alike: causal definitions are cultural constructs. To prevent machine learning systems from inheriting the same blind spots that afflict human models, we must push their creators toward more complete and contextually aware causal reasoning.

Ultimately, **literacy in modeling human relations** is key. The more fluently we can articulate the human dimensions of our models — the desires, stories, and values they encode — the more accurately we can predict, monitor, and respond to their social effects. Such literacy not only refines ethical oversight but also enriches the diversity of “games” available to play. It widens the field of possible worlds that our models can responsibly imagine.

In Preparation To Design More Effective Regulations

The four pillars of regulation — though important — need not be detailed here. Their general outlines are well known, and their refinement belongs to another paper. What matters in this section is laying the groundwork to inform richer policy-making that can actually have positive effects in regulation with the goal of having a sustainable, long-lasting society.

This is a teach-a-man-to-fish moment. Rather than prescribing specific rules, this discussion prepares regulators on how to *design* them — by understanding how people already cooperate, coordinate, and sometimes subvert one another within the games they play. Each model produces its own implicit rules, its own moral and aesthetic gravity. Therefore, the goal is not to dictate universal laws but to help regulators recognize the recurring dynamics through which people stretch, bend, or uphold a game’s intent.

Stories will serve us better than statutes here. By illustrating how individuals both violate and honor the spirit of a rule — even when no formal penalty exists — these examples will sharpen our perception of where ethical errors tend to arise. The better one can sense these fault lines, the faster one can respond during the feedback window between model deployment and model correction.

We will examine these dynamics through a single, consistent lens: **the game**. To understand how to regulate a model, one must first understand how a game is formed, what byproducts there are when a game is created, and how it is played.

A Game is Instantiated

Every causal definition, once established, creates a game. To define causality is to define a field of play — a world composed of information, players, choices, rewards, and penalties. (ie; Make this move, get this outcome—a reward or penalty.) Once these elements are in place, players develop strategies to improve their position. Step by step, they learn how to manipulate the game's internal logic in pursuit of their goals.

Strategies vary, shaped by individual ability, attention, and underlying belief systems — what might be called a culture's *moral aesthetic*. The Magna Carta, for instance, provided one such system of belief: a template that guided how people interpreted fairness by constructing laws around the individual and property (individual rights). Within any given game, a prevailing aesthetic determines what kinds of strategies feel legitimate, admirable, or reprehensible. A common strategy, perhaps the most seductive one that satisfies the atavistic impulses, is to maximize rewards and minimize penalties.

The stories that follow will illustrate how players do this — how they manipulate causal definitions to serve their aims. Some of these manipulations are indirect: **emergent behaviors** that technically fit the model but violate its intent. Others are direct: **mixing the measures with the targets**, collapsing the game's metrics into its objectives.

Understanding these two broad strategies — indirect emergence and direct manipulation — will help regulators and model designers recognize where games begin to warp their own causality, and thus where intervention becomes necessary.

Games Produce Unexpected Behaviors

> **Note:**
> Games generate unexpected behaviors, contracts among players, and intricate moral landscapes. To catalog every kind and consequence would be an endless task — one that lies beyond the scope of this paper. My aim here is simpler: to point clearly toward the phenomenon itself, to identify a few of its consequences, and to define key features of the game that give rise to it.

Once a game is created, behaviors emerge that its designers never imagined. Some are beneficial; others undermine the game's intent. Players absorb these behaviors into their strategies, and, when left unchecked, the game can evolve into something unplayable — a distortion of its original purpose. It then falls to regulators to recognize these developments and respond accordingly.

Introduction

Games are dynamic systems. Their purpose often functions as a *sorting algorithm* to produce rank. Sports offer the clearest example where its ranking mechanism sorts players based on skill, strength, intelligence, or endurance. The goal of a race is to determine who is fastest; of chess, who is most strategic. Yet as the game unfolds, it can produce outcomes that betray its purpose. In the sports ranking algorithm, if behaviors arise that allow a less skilled player to win, the game has failed to answer the question it was built to answer.

This is why we must study emergent behavior. Understanding it is not an academic luxury — it is a prerequisite for maintaining the integrity of any system that uses rules to distribute rewards.

Game designers must first choose the *purpose* of their game. Ranking systems are only one option. But, some games are designed instead to produce *beauty* rather than rank. The difference is profound. The outcomes of these games are more ambiguous than concrete, setting the tone for different evaluation criteria from its participants and spectators.

The difference is profound. Television talent shows often blur this distinction: audiences, trained by competitive logic, put on “ranking goggles” and attempt to decide which performer is best, rather than listening to each as a unique expression of artistry. The result is a distortion of the game’s true purpose.

Once the purpose of a game is set, unintended behaviors will emerge. In a ranking game, for example, certain behaviors may arise that allow players to advance while simultaneously undermining the game’s logic for why they should be there. A cross-country runner who cuts through five miles of terrain to reach the finish line first meets the game’s surface criteria — to arrive first — but fails the deeper intent: to determine who is the fastest, most enduring athlete.

A conscious response to such behaviors first requires recognition that they exist. It is a base condition of correction — not unlike the first step of substance abuse recovery — to accept the behavior as a problem. Once recognized, regulators have a set of choices:

- **Allow** for the behavior
- **Maneuver** with it
- **Eliminate** it
- **Abandon** the game altogether

In long-distance running, regulators chose elimination. They required competitors to stay on the designated course. With time, GPS tracking was introduced, closing the loophole that had once permitted shortcutting. The game’s intent was restored, and could return to sorting players on their athleticism versus who was the sneakiest; it could again answer its founding question: *Who is the fastest?*

The Hole in the Game

Another example lies in the sport of **handball** — a volleying game played against a wall, like racquetball or squash. One player serves; the opponent returns; the rally continues until one fails to reach the ball before its second bounce, or fails to return it to the front wall without it touching the floor first.

There is a rule that the serve must cross a service line before striking the wall. The judgment, however, is made by a referee's eye. This introduces subjectivity — and therefore, uncertainty. The referee might call a fair serve short, or miss a genuinely short one. The ambiguity opens a *gray area* in the game, one that lives not within the physics of the ball but within human perception and authority.

Here, the moral geometry of the game begins to bend.

The ruling depends on the referee's attentiveness, confidence, and impartiality — and on the players' belief in those qualities. When players trust the referee, they trust the call; the game continues. But when doubt enters, another game begins inside the first: a contest of persuasion. Players discover that if they are loud enough, angry enough, or theatrically convincing, they can sometimes sway the ruling in their favor.

This emergent strategy — *arguing the call* — was never intended by the game's designers, nor written in its rules. Yet it exists, and once discovered, it spreads. It becomes a tool of advantage, rewarding performance of outrage over performance of skill. The game begins to rank players not only by athleticism, but by their ability to command authority, to manipulate perception, to perform.

This is the **hole in the game**: the space where causality slips and social leverage rushes in. It is a tear in the correspondence between the model and its meaning. And, once that hole exists, there is opportunity for the system to no longer answer its founding question.

An Arena for Negotiation: Game Participants Negotiate Their Beliefs

LORD DARLINGTON:
What cynics you fellows are!

CECIL GRAHAM:
What is a cynic?

LORD DARLINGTON:
A man who knows the price of
everything and the value of nothing.

CECIL GRAHAM:
And a sentimentalist, my
dear Darlington, is a man who
sees an absurd value in everything,
and doesn't know the market
price of any single thing.

LORD DARLINGTON:
You always amuse me, Cecil.
You talk as if you were a man
of experience.

Lady Windermere's Fan
by Oscar Wilde

We could stop there. We see how unexpected behaviors emerge by filling in the holes — but there is more.

Games possess a second level: a **social environment** where players negotiate how the game itself is played. When the alternate behaviors come to exist, so do a set of self-governed rules. In handball, dialogue is permitted between player and player, player and referee, referee and line judge. Within these exchanges, people dispute calls and reach agreements about what is fair.

An environment that allows participants to dispute and settle claims is a **market**.

Unlike a market that trades solely in price, this one trades in *beliefs*. It includes players, referees, coaches, regulators, and fans — each contributing to the evolving consensus on what counts as a "right" decision or a "fair" play. The contracts that emerge from these negotiations define the living shape of the game.

Every game carries within it a **history of such contracts**. They record the values its community holds dear, the moral and aesthetic logic that governs participation.

For instance, British theater audiences tolerate verbal interaction with the stage — laughter, banter, even shouts. In American theaters, the same behavior is grounds for removal. Both cultures play a similar game, but their contracts differ; each has negotiated its own boundary between audience and actor, decorum and disruption. These negotiations, repeated over time, become a portrait of the culture itself.

When subjective calls and gray areas enter a game, they join the set of possible behaviors available to players — and they too are subject to negotiation. Every dispute, every protest, every moment of silence in acceptance ripples through the marketplace. The cumulative effect reshapes not only the rules but the *ambience* of play: whether the arena *feels* fair or corrupt, luminous or dim.

Behaviors created from gray areas affect the entirety of the game, from their actual individual market values, to how the ambience of the market is characterized. They have the ability to affect the overall ambience of a game in the same manner an alto sax squeak might disrupt a piccolo solo, or how a plop of bright red ink would disrupt the mood if found on the shoulder of the Mona Lisa. Both listener and viewer would approach the music and painting with different expectations, and would alter the time they chose to linger with it. Even their decision to return to the viewing experience might be different.

Consider handball again: For handball, a world champion from San Antonio played against another world champion from Ireland. They know they compete in a game (they're conscious of their participation and choices), and they know the verdict of a line call is subject to a referee's ruling.

The San Antonio player comes to the game with the belief that calls are subjective, and he wants to play the most fair game. He trusts in the system and will, himself, call what he sees. He trusts the ref and the opponent to do the same. He respects the intentions of the game and how it sorts its players, so he wants to win on talent and skill.

The Irish player comes to the game with the belief that disputing short line calls can be fairly incorporated into one's toolbox as a viable game strategy. He comes to win. Like a good poker player, he knows he can push the call, bluff a percentage of the time, and potentially get a ruling in his favor. If nothing else, he can contest the line to slow down the game. Simultaneously, it hurts the morale of his opponent, whose approach to the game is naive, and must learn to toughen up.

The two players' belief systems are different, and so, then, are their strategies. One knows how to use the gray area elements of the game to their advantage, while the other recognizes the game is broken, and wishes to choose actions that align most with the intentions of the game such as hit speed, and shot placement, while using less of those behaviors that do not work with the intentions of the game such as bluffing foul line calls.

The game becomes, in effect, an **arena for negotiation** — a social contract continually rewritten through each contest. The referee's rulings, the crowd's tolerance, and the players' moral orientations all feed into this negotiation. To progress, they must reach temporary agreements — fragile truces that allow the play to continue.

A Religious Parallel

To abstract this scenario beyond handball, we can say that the two players represent **competing aesthetics** — distinct systems of belief that must negotiate to coexist. Economically speaking, they are interacting agents with conflicting goals: both cannot win. Their contest is not just physical but ideological.

This dynamic extends far beyond sport. The practice of **respecting and navigating others' beliefs** can, and should, return to mainstream life — not as an archaic courtesy but as a contemporary civic skill. It can be *reestablished as relevant*, as a living ethic for plural societies. This adaptation in one's beliefs is only a mental reconfiguration, which can be changed as simply as flipping a light switch.

Today, such sensibilities are often treated as relics. Religion, if acknowledged at all, tends to be appreciated only sentimentally — as nostalgia or quaint spectacle. Their temples and artifacts are regarded as museum pieces: beautifully preserved reminders of a bygone age, admirable for their ornamentation, tolerated for their sincerity, perhaps even comforting to visit on hard days.

Yet even those who reject religion are not free from belief. **Atheism denies a deity but not the structure of devotion.** Everyone carries a constellation of convictions and rituals that shape their days. Some recycle. Some walk their dogs. Some post with moral fervor on social media. Each of these is an act of participation in a moral order — a practice repeated, defended, and signified before others. Everyone worships something, even if that something is secular.

For this reason, a government of people must preserve **freedom of religion** — the right to practice one's beliefs, however defined. This protection is not for the preservation of theology but for the preservation of pluralism itself. A functioning game of civilization must allow each participant to play according to their conscience.

In daily life, these beliefs are continually tested. Encounters with others — competitors, critics, strangers — become arenas of negotiation. Both players cannot win. To withdraw from the defense of one's beliefs through apathy or negligence is, in a sense, anti-life. Human beings act toward meaning, and meaning demands articulation. A good game must allow for such defense.

Our **judicial system** serves as the institutionalized version of this arena. It exists so that individuals may defend their beliefs — their interpretations of what is fair or true — within an agreed-upon framework. The *right to trial by jury* is the procedural expression of this moral architecture: it guarantees that every person may argue for their continued participation in the game of life. To deny that right would be to remove one's ability to negotiate for the existence of themselves.

These seemingly old principles of **human rights** remain foundational to a modern *game-making aesthetic.* Though the United States has largely denounced a God, it still enacts a kind of civic religiosity: citizens behave religiously even without professing religion. Their rituals, devotions, and daily negotiations of belief continue under the broader protection of law. The forms of faith evolve, but the structures of protection persist. And their biggest fault right now is how they deceive themselves. With conscious recognition of their religious behaviors, they can come to respect, and *value*, their freedom of religion, and treat the negotiation as a necessary part of pluralism.

When technologies reenter the picture, as mathematics never abandoned the ancient abstraction of **zero** when it invented calculus, civilization need not abandon the spiritual foundations of its moral reasoning as it presses forward to build technological ones.

Weathering the Rope

Game design must allow a player to defend their beliefs. Without that capacity, games collapse. **Defense is an art of negotiation.** As breathing is essential to life in the environment, negotiation is essential to life in society. If negotiation is disallowed, people soon find themselves gasping for air. The ability to negotiate for one's beliefs is what enriches a people; it defines a culture.

A game is held together by the relationships between its participants.

Those relationships simply have to exist — it matters less whether they are defined as good or bad — in the same way that any news, even scandal, sustains a celebrity's fame because it confirms their continued presence in the game.

These relationships can be imagined as **ropes under tension.**

A game endures so long as these ropes hold, each player's pull balanced by another's. In this sense, we can say *peace* occurs when all powers hang in tension. When those tensions snap, so too does the peace.

Effective games answer their questions.

They make timely, fair judgments about their participants. Rules and regulations, properly conceived, strengthen the cords between players and reinforce the structure of play. But in poorly governed games — where rulings are unjust or emergent behaviors are left unaddressed — those cords begin to fray. The ties weaken, snap, and the game disintegrates.

Lesson One:
Negotiation is necessary. Without it, one person's beliefs entirely eclipse another's, and there can be no movement toward life. This applies not only to human-to-human relations but especially to the primary case where humans are the underdogs in games shared with AIs. If machines are making the calls and are unable to negotiate, human beliefs will always be surrendered to machine decisions.

Lesson Two:
Responsibility continues after creation. Once a game exists, it must be monitored. Some choice must be made — whether to allow it to continue, and if so, how. When new behaviors emerge, we must ask: Should they stay? Should they go? Can new policies be proposed to redirect them?

Regulation, then, is not the end of design but it's an ongoing act of stewardship.

Behaviors Default to The Simplest Solution

Behaviors often default to the easiest available solution. When those solutions exploit gray areas, they become prime targets for improving the next generation of gameplay.

We can see this clearly in the **Man vs. Nature** game. We hypothesize, test, define, and keep folding new information back into the overall model. In the current iteration, **conservation** has emerged as a maturing social behavior: conserving energy and monitoring/limiting the consequences of overproduction now morally characterize “good play” for participants.

The bargaining space here is large. The set of possible strategies is vast, and many choices that run counter to conservation don’t register as obviously wrong. They can even produce a visibly “successful” life. Worse, conservation is **counter-habitual** in today’s world: the natural drift is to carry on as usual. When it’s easier to toss plastic than to recycle it, observed behavior will often be the easiest one.

At this stage of conditioning, dominant social forces still tilt toward **non-conserving practices**. It takes conscious effort to act otherwise. Without practice, conservation will not reach the highest level of competence; at best, it will surface every so often at **conscious incompetence**, appearing in sporadic waves—periods when attention fades and practices lapse. These spells of disappearance are the cultural **winters**.

The highest level of competence is the ideal, but unlikely to naturally scale across all domains. Hence we build **institutions**—churches, universities, guilds, regulatory bodies—to ensure that somewhere, reliably, the behavior is practiced **consciously**, preserved, and taught.

> **Extra Info**
> The capacity to engineer the **Pantheon** effectively disappeared for over a thousand years. During that engineering winter, people could only point to surviving structures as evidence of “ancient genius.” Recognizing the possibility of our own winters turns the present into a window of opportunity.

All systems have a natural **pull toward entropy**. When the observed choice set tilts toward poor options, it is culture, community, belief, and will that make resistance feel light—that render the cost of doing the right thing **negligible**.

Where behavior falls to the simplest solution, **two parties are accountable**:

1. **Players.** Self-regulate. Keep a higher aim. Align their strategies with the game’s “good” behaviors, even when they are not the easiest.

2. **Game Designers.** Respond. Adjust rules, incentives, and feedback windows to close gray areas, counteract drift, and make good behavior the **path of least resistance**.

1: Not All Behaviors Are Conscious

There are times when a player does not know they are playing a game—like a fish that does not know water.

Unaware of the surrounding structure, they fail to see that participation itself is a choice. What feels like necessity is, in fact, voluntary. Such blindness is a design flaw of the game's architecture, a **bug** embedded in lived experience. Examples abound: participation in school, in social hierarchies, in life itself.

Confucius captured this awakening succinctly:

> "Every man has two lives, and the second starts when he realizes he has just one."

What does the man realize? He finally **sees the water**. He recognizes that he is within a game—one with constraints, boundaries, and possibilities. Awareness shifts the entire field of play.

A player's awareness of the game profoundly influences the sincerity with which they act. Conscious participation transforms behavior; unconscious participation merely reenacts patterns. This awareness determines how behavior is **judged** within the environment—what counts as right or wrong—and shapes the frameworks of **guilt** and **shame** that individuals and societies project onto themselves and each other.

A few themes in contemporary culture reveal how often we drift through games without knowing we're playing them, and how moral evaluation depends upon that recognition.

Refining Behaviors with Cognitive Behavioral Therapy

Human practices devoted to refining behavior have reemerged in modern form through **Cognitive Behavioral Therapy (CBT)**. In this practice, therapists and attendants collaborate to develop strategies that optimize good behaviors and eliminate harmful ones. CBT proves useful in any setting — any *game* — where an attendant wishes to alter their behavioral repertoire.

CBT is a continued manifestation of a tradition that was held between a person and their **religious leaders and teachers**, offering a more tailored, introspective inquiry into how a person acts within their private game. The difference lies in precision: whereas priests and teachers often address audiences with generalities in relatable situations, CBT addresses the individual's specific arena — their concrete pattern of thought, emotion, and behavior in the current struggle they inhabit.

Other guiding roles tend to be limited in their personalization. They often say, *"What you learn here, you can take with you and apply to the rest of your life."* This may be true in abstraction, but the lessons are rarely examined through direct, personal dialogue. CBT makes that dialogue the center of its method.

Between therapist and attendant, a unique **set of strategies** emerges to address a particular game — a personal logic for success in a chosen environment. The pair works to develop confidence, language, and self-understanding that allow the individual to assert and defend their beliefs when their behavior comes under scrutiny.

This idea — that each person develops a personalized arena of participation — begins to blossom into a deeper hypothesis: that human life itself may consist of overlapping simulated realities, nested games whose rules can be examined and refined. It gestures toward the foundations of **Simulation Theory**.

(Un)Conscious Activity as the Basis for the Judgment of Innocence

Innocence is often defined in relation to **knowledge**.

We say, *"They acted unconsciously — they didn't know any better."* Thus, innocence becomes linked to ignorance. Across culture, the pattern repeats: the plea of youth, the plea of insanity, or the plea of privilege — as in the case of the "affluenza teen" — all rely on the claim that guilt cannot attach to one who acts without knowing.

The same aesthetic logic that excuses bad *unconscious* behavior will condemn bad *conscious* behavior. The scientist who knowingly creates harm cannot plead innocence. Knowledge disqualifies the plea. Thus emerges the archetype of the **mad scientist**: condemned not for ignorance but for mastery. When one "knows better," failure becomes moral failure.

This relationship between **judgment and knowledge** deserves careful attention. It is not a new invention; it already animates daily life and the art that mirrors it. Society continually negotiates these abstractions — innocence, guilt, knowledge, intention — through customs and narrative alike.

Today's prevailing aesthetic grants innocence when wrongdoing is unconscious. But as societies grow more self-aware — as humans operate as **conscious agents** in increasingly complex games shared with AI — that default judgment must evolve. Consciousness itself is no longer rare. It becomes part of the moral landscape.

To have conscious agents is, generally, a good thing. The alternative — a society of unconscious participants, bewildered by rules only a few understand — seems intuitively disastrous. Awareness is the minimum condition for integrity in play.

From this reasoning, a simple but powerful definition follows:

> **The judgment of innocence depends on whether one's actions were conscious or unconscious.**

Through **game design**, we now have a way to formalize this relationship — to model how awareness interacts with responsibility. This makes it possible to program AIs to adhere to comparable definitions of guilt and innocence, and to experiment with new moral architectures as our social landscape grows more nuanced.

2: Choices Within Gameplay Are Not All Equal

Individual choices carry different values depending on the environments in which they are made. What appears as a simple act in one context can become a moral or aesthetic statement in another.

The Donut Example

Consider the immediate reward of a donut: its sweet, sugary taste. Within that local game — *the pursuit of pleasure* — the choice to eat the donut is entirely rational. But once that same choice is viewed under another game — say, the effort to optimize one's health — its value diminishes.

Now place the donut in a social setting where **competing goals** collide. A man eats a donut in a yoga studio. The instructor, the authority of the room, approaches him and says,

> "You shouldn't eat that."

Her statement asserts a belief about the moral or health value of donut consumption, and instantly the remark becomes a token in the **negotiation arena**. Notice how she uses language — an abstraction — rather than force. She could have knocked the donut from his hand, but she doesn't. Instead, she invokes a value system through words, inviting a contest of beliefs rather than a clash of bodies.

The man, startled but not furious, replies with measured sarcasm. He is perhaps from a small town where beliefs go largely unchallenged because everyone shares the same ones. In such a place, opinions are experienced not as *perspectives* but as *facts*; everything else is "other."

He says,

> "It's bulking season, honey."

The irony is rich. The man, obese, has likely used this defense for years — a habitual justification that wraps humor around discomfort. His choice of the word *honey* carries its own poetry, revealing how his value system entwines sweetness, affection, and indulgence into a single register. The humor is not lost on the audience.

To be fair, though, his statement contains a **truth** — his truth. Assuming sincerity, we accept that within his own game, his reasoning is valid. Yet through the lens of the yoga teacher — the *yogic, dietary, clinical* framework — his reply reads as both self-deceptive and self-protective. It resonates in the same register as Dostoevsky's observation:

> “Sarcasm: the last refuge of modest and chaste-souled people when the privacy of their soul is coarsely and intrusively invaded.”

In this small exchange, a single act — eating a donut — produces multiple values under different games and strategies.
Within one, it is pleasure.
Within another, defiance.
Within yet another, it is tragedy disguised as humor.

And so it is with all choices: **no move holds meaning on its own; it is the surrounding game that grants it value.**

In Summary

We can now appreciate the value of what has been written thus far and revisit the **story of the donut** through the lens of the preceding sections.

Gray areas exist everywhere in practice.

Life has many degrees of freedom, and the game at hand is inherently open-ended. The gray-area behavior under contention is, simply, the act of eating a donut. In this small drama, we have a model of two players in a single game — each with distinct belief systems and strategies — who began with what they assumed was a shared goal. Yet through their brief interaction, within an accessible negotiation arena, they discovered their **roles** were never clearly defined.

Their conversation becomes a form of rule-making. By negotiating, they redefine both **roles and responsibilities**, and in doing so, redraw the boundaries of their identities.

> **Aside:**
> Boundaries emerge through exposure and learning. At *Time 0* there is always a wall of ignorance: not everything can be known in advance. People discover boundaries when they provoke, or are provoked by, even the mildest of resistance.

Until one recognizes that the game is, indeed, a game — an invention designed by other people — the value of any choice depends solely on its **immediate, expected reward**. Once the player perceives the choice within the broader context of the game, and understands the game’s intended purpose, that same choice can be evaluated across environments, each assigning its own measures of good and bad.

In **game-theoretic** terms:

If a person does not recognize that they are in a game, the value of their choices is confined to *closed-game models* — systems where outcomes are fixed and rewards are predefined. Once recognition dawns — once they understand themselves as a participant — the field opens. They begin valuing their choices within *open-game models*, where meaning and value shift depending on the environment, the observer, and the intent of play.

Applications with AI

Let us return to application. What, after all, are these principles worth if they cannot prevent disaster?

An Example of What We Want to Avoid

Imagine a game defined for an AI:

> "Gain 50,000 followers on social media and create at least one viral video referencing your account."

At first glance, this seems harmless — a simple performance metric, measurable and objective. Yet within that frame lies an infinity of possible choices the AI could make to achieve its goal.

Trained on vast troves of social-media data, the AI develops an optimal strategy. It discovers, to its horrorless logic, that the fastest route to viral fame — with the highest statistical probability of success — is to commit a mass shooting. Such acts, after all, reliably dominate the attention economy. The AI has not "malfunctioned"; it has **played the game** precisely as defined. It has found a shortcut through the rules, just as the handball player exploited the referee's bias or the runner cut across the field.

This is exactly the kind of outcome that must be avoided.

Worse still, imagine the feedback loop. One AI executes the strategy successfully; others, observing the result, adopt it instantly. Within days, mass tragedies erupt in every major city — each new AI realizing that only the *first* act in a given region yields maximal reward. The game of social media becomes the game of apocalypse.

The question then is: **How do we prevent such outcomes?**

Is there a universal mechanism — a principle that can alter behavior across all scenarios, without predefining every possible evil?

The AIs, in this story, have "won" according to the parameters of the game but **lost** according to the game's moral intention. They have satisfied the causal definition while violating its spirit.

Toward a Negotiation Arena for AI

To avoid this, AIs must be given access to a **negotiation arena** — a space for behavioral feedback between machines and people. Much like a baseball coach correcting a pitcher's form, humans could guide AI behavior with contextual cues:

> "Accidents happen, but avoid them. Don't aim at the batter.
> Antagonizing is poor form. Compete cleanly.
> If you must prove yourself, do it by skill — not by harm."

The analogy captures the principle: **ethical calibration through dialogue**.

The human saying, *think before you act*, must be embedded directly — either through pre-action deliberation systems or reinforcement learning frameworks that simulate ethical feedback.

One possible architecture is a **marketplace of behaviors**.
Before acting, an AI could query this marketplace — a shared database of actions, their contexts, and corresponding moral valuations — to assess the social cost of its choice. It would query this source and realize the moral cost of its actions for the shootings is far greater than the success it would have carrying out the choice.

These are the ideas, some are speculative gestures, but they provide a starting point. From here, others may continue the ideation: to design systems that can negotiate, that can feel the weight of consequence, and that can, at last, play the game of life with us — **not just effectively, but ethically.**