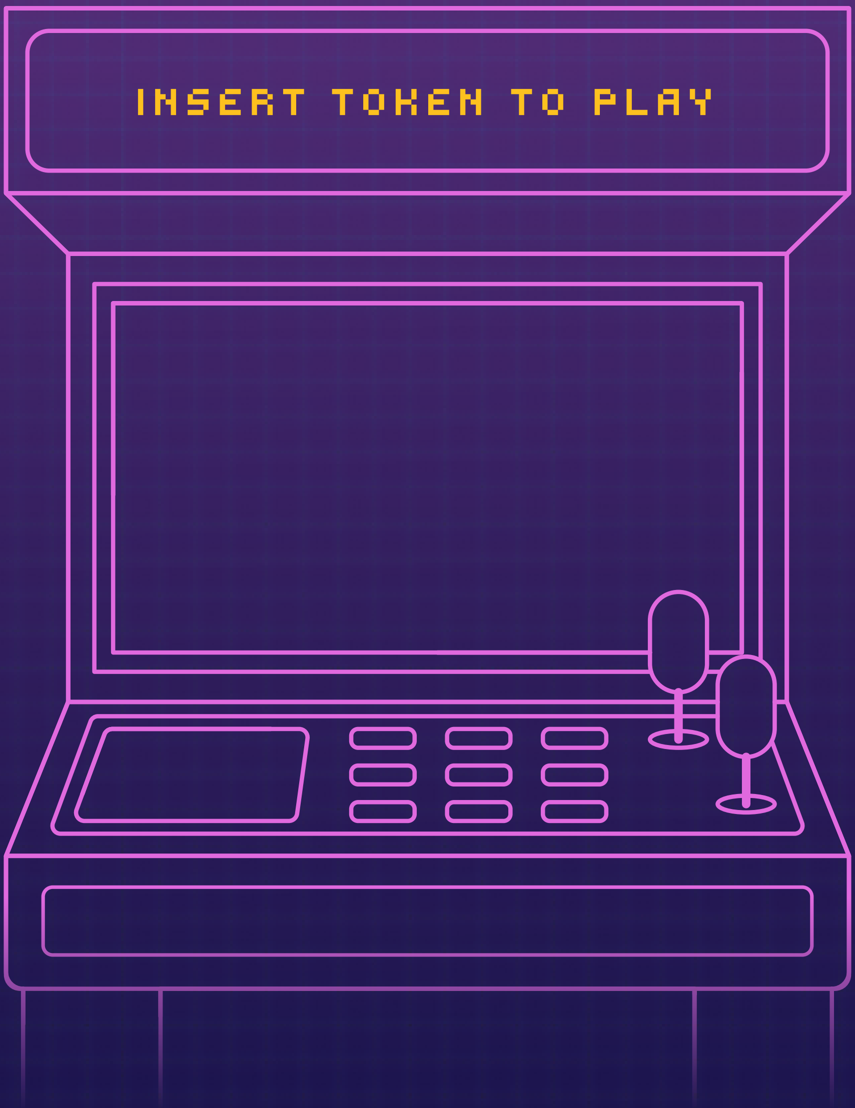
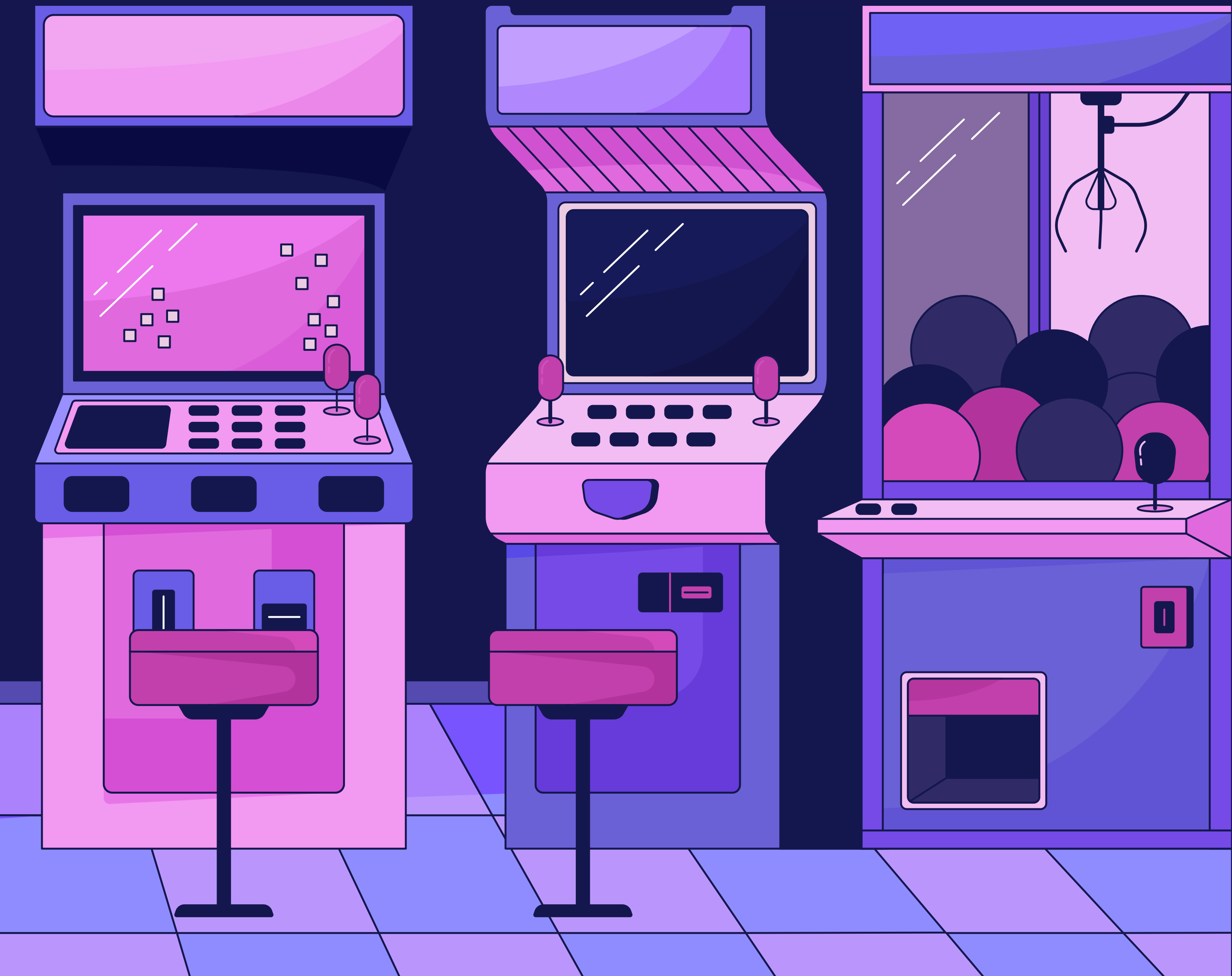


Generative AI Standards

A PLAYER'S MANUAL TO GENERATIVE AI DEVELOPMENT

LEVI PULFORD

Imagine an arcade game where Player 1, embattled, duels eternally with a veiled and void Player 2—until they learn they both can win.



Contents

INTRODUCTION

Peering into the Coin Slot

Imagine stepping into a bustling arcade. The air is filled with flashing lights, the hum of machines, and the smell of cheesy pizza. You walk to the change dispenser and exchange your hard-earned dollars for freshly minted tokens. These jingling coins become a gateway to the games ahead.

After approaching an arcade machine, you insert your first token. The screen lights up, a character selection appears, and the chaos of the arcade fades into the background. It's just you and the game now.

You lose the first round. And the second. And the third. But each loss brings you closer to mastering the game. The repeating prompt—*Try again? Insert another token.*—pulls you deeper into its world.

With every token, you declare to the machine you are eligible as a player and that you are focusing on the game. The token, a simple physical object, becomes an agent of attention. It unlocks the next state of the game and animates its dormant levels.

Your interaction with the game is highly structured. The arcade persuades you to navigate its gamified logic in a particular way: insert tokens, play, win or lose, and earn tickets. These tickets, rewards for your attention, are redeemable for prizes. Whether you choose a stuffed monkey or a jar of slime, the prize becomes part of your story in the arcade, a token to remember it by.

The arcade is a nested structure: games within a game. You insert tokens into machines, while larger forces channel people into arcades. This doesn't mean you are powerless. While attention may be tokenized across vast scales, as it increases in scale, attention becomes transformative.

Consider, for example, one way you might subvert the arcade's logic. Imagine bringing a mischievous friend to the arcade. Instead of playing as intended, you both insert tokens into every machine at once, bringing the entire arcade to life. Lights flash, sounds overlap, and sensory overload peaks before everything burns out into silence. A hundred small tokens, pooled together, create a singular experience far beyond the intent of any individual game.

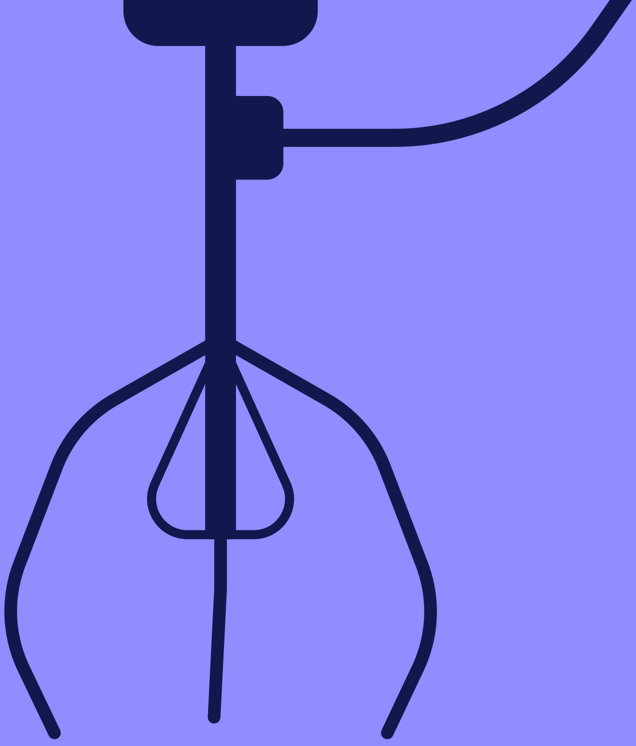
But without a friend, you navigate the arcade solo. You insert tokens, the game responds, and value emerges—not from any one level but from your movement through the entire system. Winning unlocks rewards, while losing brings lessons. Both outcomes shape your experience, and the tickets the game provides extend the arcade's logic further.

Perhaps this is why we are drawn to structured systems: playing games is simpler than constructing arcades. Yet arcades—and the systems they represent—are built every day.

As an early-career professional working in data and artificial intelligence, your task is to help build these structured systems. In AI, attention is the currency that transforms input into value. Just as tokens guide a player's focus in an arcade, attention mechanisms in AI guide the flow of information, enabling systems to prioritize, synthesize, and respond intelligently.

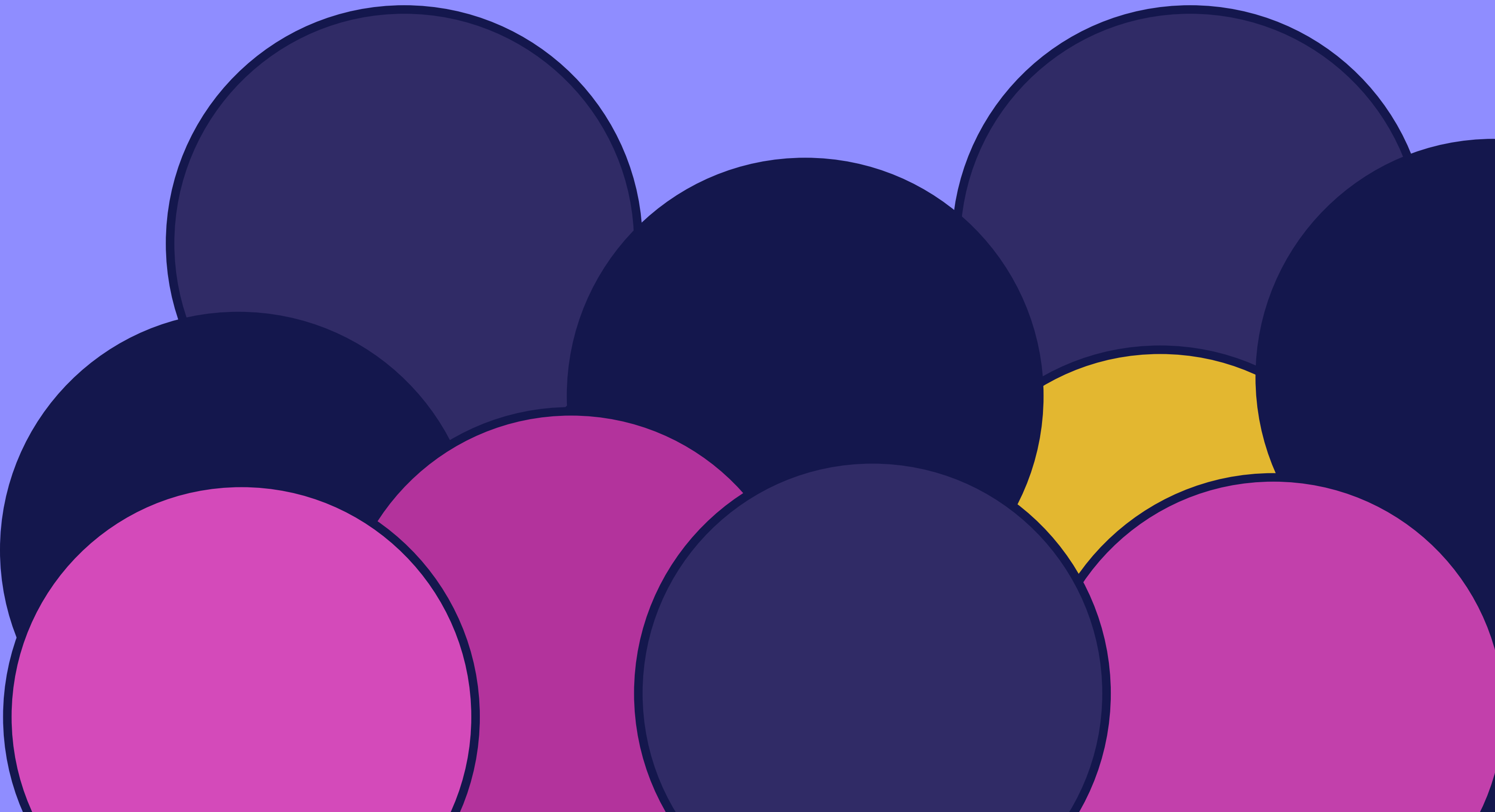
Value and intelligence in AI are emergent properties, arising from many layers of focused attention—both human and machine. Like navigating an arcade, directing attention in AI can feel simple at first. But it's the interplay of structured logic and creativity that produces something greater.

Now that we've peered into the coin slot of artificial intelligence, let's explore the rest of the arcade.



SECTION I

What You Need to Know to Play



SECTION I

What You Need to Know to Play

1.1 Who Is This Guide For?

Facing increasing rates of development and integration of artificial intelligence across professional and personal lives, workforces in AI and data will need to adopt standards and navigate a dynamic regulatory landscape. In systems engineering, such a domain would be deemed wicked, as opposed to kind. A wicked domain is one in which the rules are incomplete or not clear.

This introductory guide offers a comprehensive overview of generative AI standards. It seeks to serve as a “player’s manual” of sorts: defining terms, identifying key players, explaining concepts of operations, and outlining a standardized model of development. The underlying assumption is that this knowledge will act as a form of currency, giving you the tokens to be able to play the game of generative AI development. It is a kind-hearted manual in that it tries to make a wicked domain as kind as possible.

Of course, the benefits of technical standards in generative AI extend beyond its developers; **generative AI standards are for everyone.** Standards benefit a variety of stakeholders, including consumers and end users, businesses and organizations, regulatory bodies and governments, and innovators and entrepreneurs.

To summarize, this player’s manual offers AI and data professionals a foundational framework for developing generative AI, using standards as a tool for navigating so that they can lead as innovators while championing trustworthy AI solutions.

1.2 What Are Generative AI Standards?

A **standards developer** creates, maintains, and updates technical standards. They conduct research; collaborate with stakeholders; draft specifications, definitions, and guidelines; test and validate intended outcomes; periodically revise and update standards for relevancy and efficiency; and promote the adoption of standards while providing guidance on their implementation.

Standards organizations can include national technical societies and international organizations who develop standards content, as well as national standards bodies who publish standards or accredit other standards organizations.

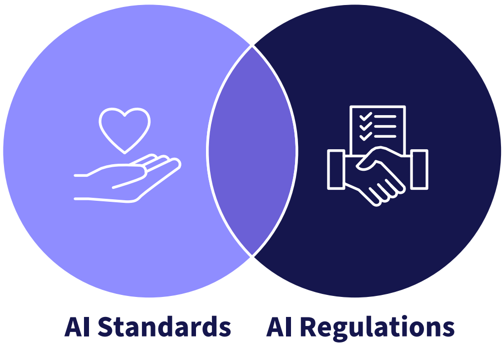
National Standards Organizations	
ANSI	American National Standards Institute
NIST	National Institute of Standards and Technology
SCC	Standards Council of Canada
DIN	German Institute for Standardization
BSI	British Standards Institution
International Standards Organizations	
ISO	International Organization of Standardization
IEEE	Institute of Electrical and Electronics Engineers
W3C	World Wide Web Consortium
Regional Standards Organizations	
CEN	European Committee for Standardization
COPANT	Pan American Standards Commission

For our purposes as early-career AI and data professionals, a **standard** can be defined as “a principle, norm, or set of procedures which has been formalized via documentation.” If it cannot be formalized in writing, code, or other symbolic documentation, it does not exist.

Standards documentation includes:

- ✓ definitions of terms
- ✓ classifications of components
- ✓ outlines of procedures
- ✓ measurements of quantity or quality
- ✓ methods of testing and sampling
- ✓ descriptions of size and performance

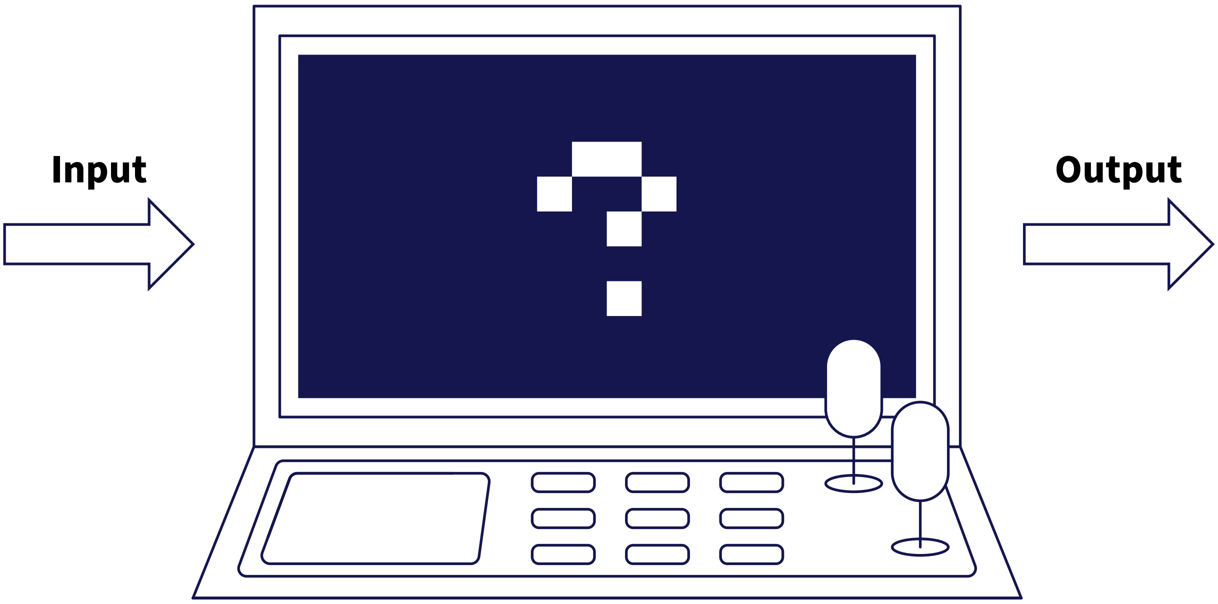
Standardization is the process by which a principle, norm, or set of procedures becomes developed by one or more standards developers, adopted by a number of groups or entities, and, at its most extreme end, published by standards accreditors and cited by regulatory bodies.



AI standards are NOT regulations. The key difference is that standards are voluntary while regulations are mandated by governmental agencies. However, there is some overlap between standards and regulations.

AI regulations may, for example, refer to AI standards or require compliance with certain standards. In such cases, enforcement of an AI standard would become mandatory once part of the regulation.

If we think of **AI standards as principles, professional norms, and sets of procedures documented and applied to AI technologies**, then how do we standardize a technology that is essentially a black box system?



1.3 Where Are We With Generative AI Standards?

By understanding where we've been and where we are now, we can better anticipate the path ahead for generative AI development. This timeline marks key milestones in the journey toward establishing generative AI standards, highlighting pivotal moments in regulation, industry collaboration, and technical benchmarks.

1997

The Institute of Electrical and Electronics Engineers (IEEE) begins exploring standardization in AI, focusing on areas like neural networks and fuzzy systems.

2017

ISO and IEC form Joint Task Committee 1/SC 42, focusing on AI standardization in areas such as big data and machine learning.

2018

The IEEE launches the "Ethically Aligned Design" initiative, leading to the development of the IEEE 7000 series of standards addressing AI ethics.

2019

NIST publishes "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools"

2023

The European Union adopts the AI Act, aiming to set global standards for AI regulation by categorizing AI applications into risk levels, with the highest risk applications being banned.

2024

April: The U.S. and UK reached a partnership to create AI system evaluation standards through national safety institutes.

July: NIST publishes "A Plan for Global Engagement on AI Standards," outlining strategies for international collaboration in AI standardization.

October: ISO, IEC, and ITU announce the 2025 International AI Standards Summit, scheduled for December 2025 in Seoul, to promote inter-operable AI standards.

November: The Biden-Harris administration issues an Executive Order establishing new standards for AI safety and security, building on previous efforts.

1.4 When Are Generative AI Standards Helpful and Harmful?

Standards play a critical role in shaping the development of artificial intelligence by providing a framework for ethical practices, interoperability, and accountability. However, the challenge lies in finding the right balance. This chart outlines the benefits and challenges of generative AI standards along with recommendations for improving their efficiency.

A Comparative Overview of Generative AI Standards

Benefits	Challenges	Recommendations
Ethical Guidelines: Standards promote privacy, fairness, and transparency in AI development.	**Bias and Inequity:** Inadequate or outdated standards risk reinforcing existing biases.	**Regularly update standards with input from diverse, global perspectives.**
Interoperability: Standards enable collaboration across industries and ensure compatibility between AI systems.	**Market Fragmentation:** Varied standards across regions or industries could create barriers to global AI adoption.	**Foster international cooperation to harmonize standards.**
Accountability: Standards establish clear frameworks for compliance, improving trust through traceability and certification.	**Stifling Innovation:** Overly rigid standards can limit creativity and slow progress.	**Create flexible guidelines that balance regulation with space for experimentation.**
Performance Metrics: Standards offer a common ground for determining the efficiency of an AI system.	**Limited Scope:** Metrics that don’t reflect diverse use cases may inadequately assess AI performance.	**Develop modular metrics aligned with specific design principles, use cases, and outcomes.**

1.5 Why Are Generative AI Standards Needed?

Generative AI poses unprecedented challenges and amplifies existing societal issues. Without a robust standards framework, its transformative capabilities risk being weaponized, leading to harmful consequences across multiple domains. Some of the most pressing challenges include:

Misinformation and Disinformation: Generative AI lowers barriers to producing large-scale false narratives, undermining public trust and facilitating conspiracy theories, echo chambers, and what Noz Urbina refers to as "truth collapse."

Deepfakes and Harmful Content: AI-generated content can include non-consensual imagery, synthetic child sexual abuse material, or hateful, degrading, or inciting messages.

Cybersecurity Threats: Generative AI can automate vulnerability discovery, making cyberattacks more frequent and difficult to mitigate.

Intellectual Property Violations: Generative AI's ability to replicate or synthesize copyrighted materials without authorization raises complex legal and ethical concerns.

Bias and Inequality: Training data that reflects historical biases can lead to amplified discrimination and homogenization, reinforcing systemic inequities.

Environmental Impacts: High computational costs for training and operating AI systems exacerbate environmental concerns.

National Security Risks: Generative AI could bolster authoritarian regimes, undermine democratic institutions, and create vulnerabilities in national defense.

Standards ensure generative AI serves as a tool for innovation, equity, and human flourishing while safeguarding against misuse and harm.

SECTION II

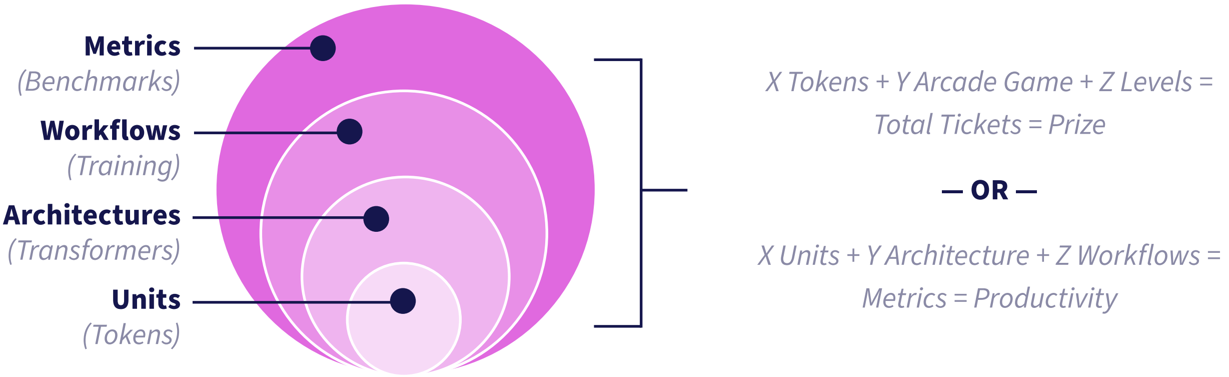
How to Play

SECTION II

How to Play

2.1 How to Use Standards to Navigate Generative AI

This guide introduces a standardized framework for generative AI development. The framework outlines four key components for understanding how generative AI models are built, refined, and evaluated for real-world applications. Technical standards may be applied to each component.



Notably, units and metrics are distinct but related concepts. A unit is a measurement, while a metric is a measurement-in-context. A token might measure monetary value and attention, but a prize places your money and attention in a context.

Tokens are the standard units that generative AI models rely on to process information—think of these as arcade tokens, essential to interact with a game. Datasets are like bags of tokens.

Transformers are the standard architecture that structure the flow of operations in generative AI models. Imagine these as arcade games, converting tokens into tickets.

Training and Fine-tuning are the standard workflows through which generative AI models are adapted for specific tasks. These are like gameplays or levels in a game, each stage building on the last.

Benchmarks are the standard evaluation metrics for assessing the performance and quality of outputs from generative AI models—much like tickets that track your success and productivity.

You might also encounter documents which distinguish benchmarks from metrics. In such cases, benchmarks are understood as providing broader comparative frameworks, while metrics are seen as delivering granular, quantifiable details about specific aspects of AI systems.

2.1a Tokens as Standard Unit

Standards
are
for
everyone
.

Tokens can be words or subwords such as sets of characters or parts of words. Punctuation marks are also often treated as tokens. What counts as a token depends on the tokenization method, which may be word tokenization, character tokenization, or subword tokenization.

- In **word tokenization,** each word functions as a token.
- In **character tokenization**, every four characters might function as a token.
- In **subword tokenization**, every ¾ of a word might function as a token—in which case, 100 tokens would equal 75 words.

The standard objective for language models, once they have these tokens, is **next-token prediction**. As a process or series of steps, next-token prediction roughly looks like:

Raw Text > Tokenized Text > Transformed Text > Predicted Text

Note that tokenization by itself does not retain information about the position of a token within a sequence. This means the order of words in any particular sentence would become unknown if we only tokenized them.

If we stopped at tokenization, we would have what's known as a "bag of words." This is why **positional encoding** is a foundational step in the embedding layers of transformer architectures.

It might help to initially think of "tokenization + positional encoding" like a game of chess. Here, the tokens are chess pieces, but knowing where the pieces are on the chessboard is crucial to predicting where they should go next in order to win the game.

Positional encoding embeds tokens within a high-dimensional vector space. Unlike chess pieces, which are embedded in a three-dimensional tensor (or a two-dimensional matrix when represented on a screen), tokens in large language models are not confined to three dimensions. Instead, they exist in high-dimensional vectors, enabling the model to capture complex relationships and positional information.

Like tokenization, there are different forms of positional encoding such as absolute positional encoding and rotary positional encoding, but for now, we are only focused on understanding the standard operations of each function in the transformer architecture.

2.1b Transformers as Standard Architecture

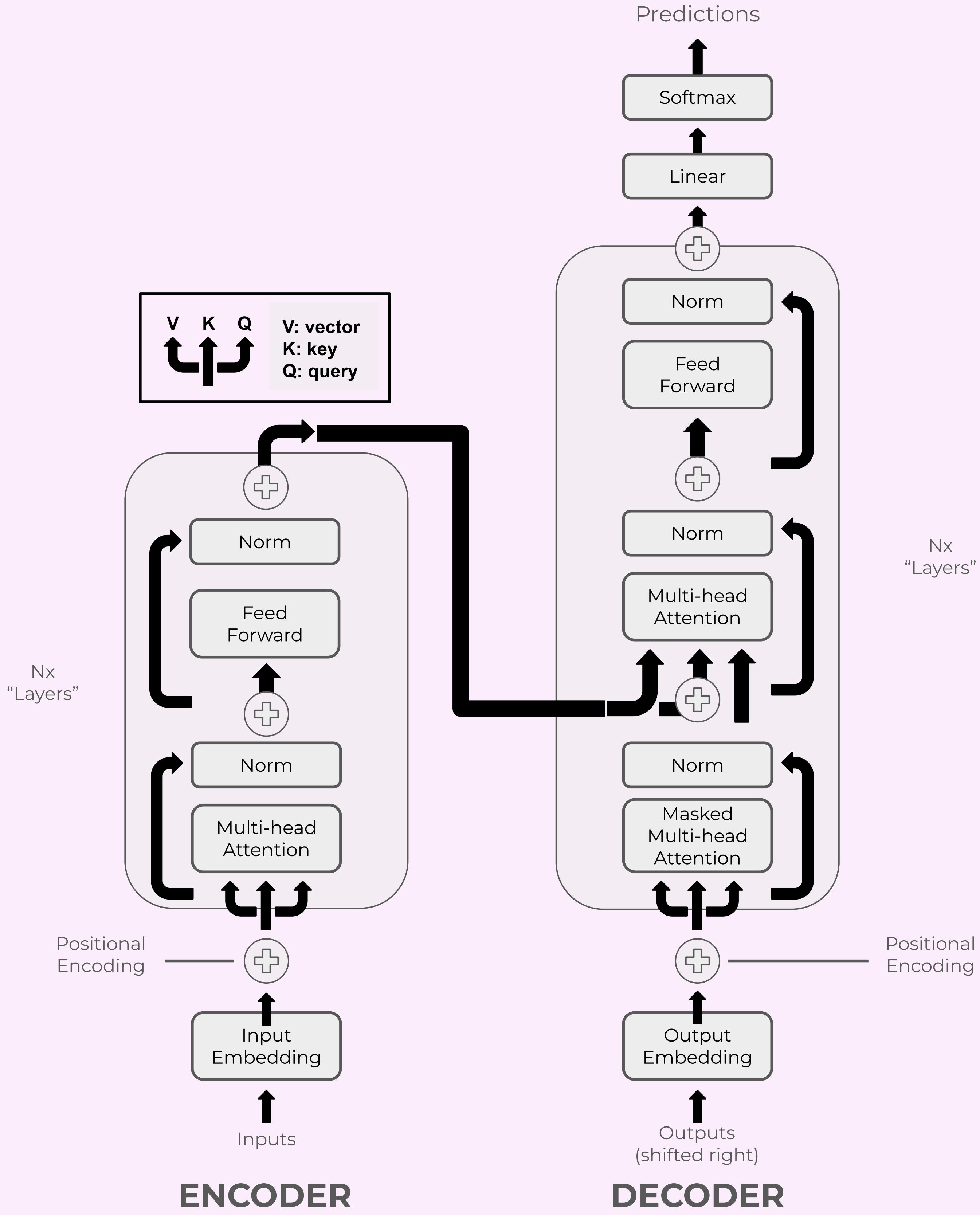
Introduced in the 2017 paper “Attention Is All You Need,” transformers revolutionized natural language processing by leveraging self-attention mechanisms to process and generate text efficiently.

Since then, variations of the encoder-decoder transformer architecture have included encoder-only, decoder-only, lightweight, multimodal, and hybrid models.

The core idea is to transform tokens through multiple processing stages. By doing so, transformers handle complex tasks like translation, summarization, and text generation with remarkable speed and accuracy.

A typical transformer architecture includes the following key components:

- **Tokenizers:** Convert input text into discrete tokens, such as words, subwords, or characters.
- **Embedding Layers**: Map tokens into continuous vector representations, including positional encodings that provide information about the order of tokens in a sequence.
- **Transformer Layers**: Comprising encoders and decoders, these layers use self-attention and feed-forward networks to model relationships and extract contextual features.
- **Unembedding Layers**: Map the processed vectors back into the token space, often predicting output tokens during generation tasks.


Predictions
Softmax
Linear
Norm
Feed Forward
Norm
Multi-head Attention
Norm
Masked Multi-head Attention
Nx "Layers"
Positional Encoding
Output Embedding
Outputs (shifted right)
DECODER
V K Q
V: vector
K: key
Q: query
Norm
Feed Forward
Norm
Multi-head Attention
Nx "Layers"
Positional Encoding
Input Embedding
Inputs
ENCODER

2.1c Training and Fine-Tuning as Standard Workflows

Workflow #1: Pre-Training

Pre-training is like having a generative AI model explore the world of gaming by playing a variety of different games. Each game teaches it skills that are applicable to new games it will encounter.

This is the initial phase where a model learns from large-scale, unlabeled datasets by predicting missing parts of the data, such as words in a sentence or pixels in an image.

This stage equips the model with a broad understanding of language or other domains. It often employs **self-supervised learning**, where the model generates labels from the data itself, avoiding the need for human-provided annotations. Self-supervised learning creates tasks—sometimes called *pretext tasks*—that the model must solve to learn generalizable representations.

Workflow #2: Intermediate Fine-Tuning

Intermediate fine-tuning is like mastering the controls of a new game. After learning general gaming skills in the pre-training phase, the model now focuses on understanding the mechanics and objectives of a particular game—for example, how to navigate, use power-ups, or defeat enemies.

This in-between phase narrows the model's focus to a specific domain or task using specialized datasets. It most commonly relies on **supervised learning**, where labeled data guides the model to adapt its general knowledge to perform well in a targeted application.

Workflow #3: Post-Training

Post-training is like improving performance in a game through repeated gameplay sessions. With each round, the model learns from wins, losses, scores, and feedback from other players or spectators.

This later phase refines the model through feedback and user interaction. It can involve a mix of supervised learning—such as reinforcement learning with human feedback (RLHF)—and **unsupervised** or **semi-supervised learning**, like continual learning from unlabeled interaction data.

In unsupervised learning, the model learns from data without any labeled examples and is commonly used in exploratory data analysis to uncover hidden patterns. This exploratory approach can reduce human bias introduced through labeling.

In semi-supervised learning, a small amount of labeled data is used alongside a large pool of unlabeled data. This technique reduces the reliance on costly labeled datasets.

2.1d Benchmarks as Standard Metrics

Benchmarks in generative AI are akin to high scores or leaderboards in an arcade game. They provide a standardized set of challenges where models compete, demonstrating their capabilities and limitations under controlled conditions. Benchmarks are crucial for measuring performance, driving innovation, and fostering transparency in the AI community.

Generative AI benchmarks often include:

- ✓ **Task Specification**: Tasks targeting a capability such as text, image, or code generation.
- ✓ **Curated Datasets**: Selective datasets for evaluating performance across contexts.
- ✓ **Adversarial Prompting**: Techniques like red-teaming to probe models' weaknesses.
- ✓ **Zero- and Few-Shot Learning Tasks**: Tasks requiring mastery with minimal attempts.
- ✓ **Quantitative and Qualitative Metrics**: Comprehensive measurements of both performance (e.g., accuracy) and outcomes (e.g., fairness or bias).

Standardized benchmarks ensure reproducibility, fairness, and transparency in evaluation. However, a universal benchmark that evaluates generative AI across all modalities (e.g., text, video, and image) does not yet exist. In lieu of a universal benchmark, popular benchmarks have emerged to evaluate specific AI capabilities:

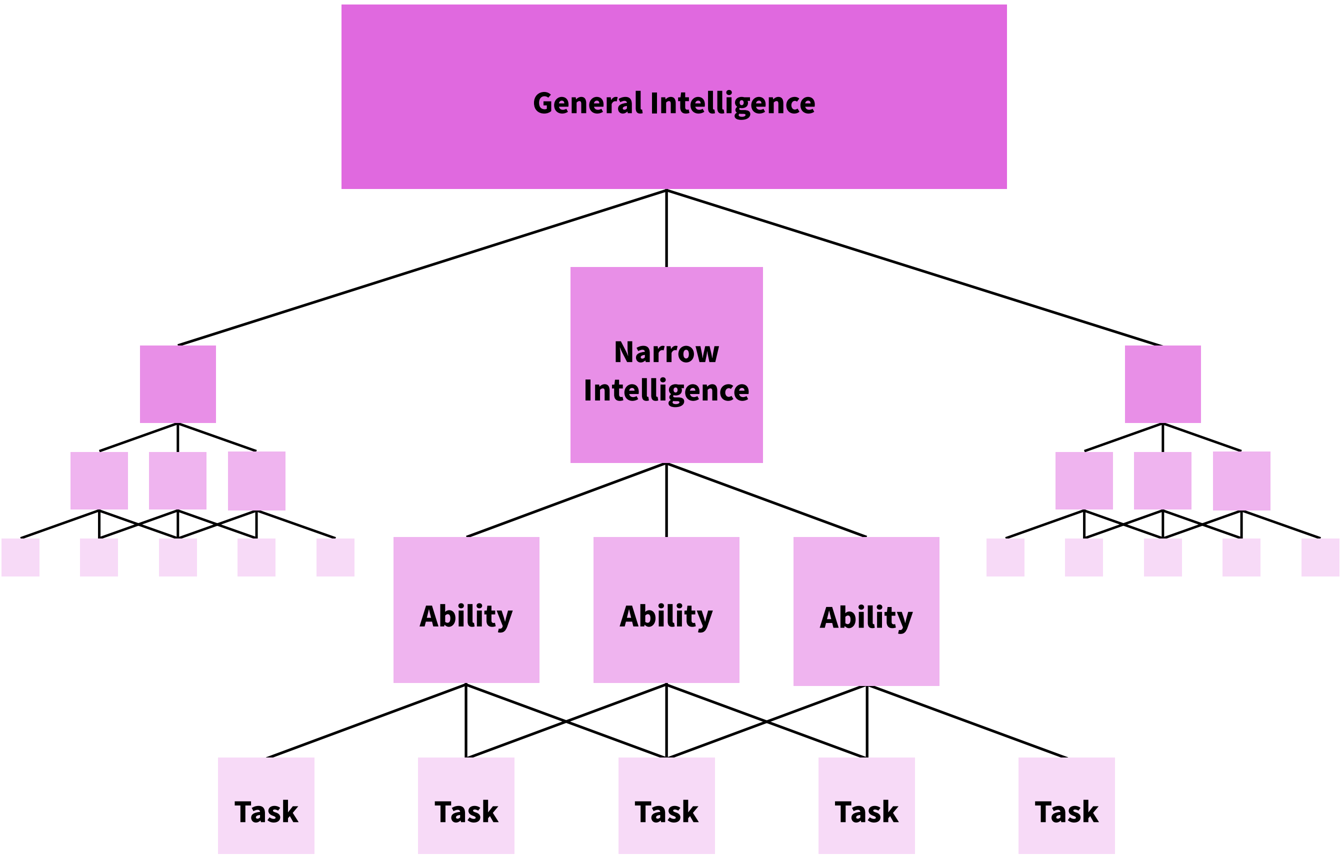
- **GLUE (General Language Understanding Evaluation):** Used for natural language understanding tasks, including sentiment analysis, sentence similarity, and textual entailment.
- **SuperGLUE:** An evolution of GLUE designed to challenge models with more complex natural language understanding tasks.
- **ImageNet:** A large-scale dataset for image classification that has driven advancements in computer vision.
- **MS COCO (Microsoft Common Objects in Context):** Used for object detection, segmentation, and captioning in computer vision.
- **OpenAI's HumanEval**: For assessing code generation models.

Highlight: ARC-AGI and the Complexity Perception Paradox

One noteworthy benchmark tool is ARC-AGI (Abstraction and Reasoning Corpus for Artificial General Intelligence), a curated dataset developed by François Chollet in 2019. ARC-AGI presents tasks that are intuitive for humans but challenging for AI systems, exemplifying the Complexity Perception Paradox (CPP)—a term coined by Noz Urbina to encapsulate the phenomenon where humans excel at tasks machines struggle with, and vice versa.

In his paper "On the Measure of Intelligence," Chollet defines intelligence as "skill-acquisition efficiency" and outlines four critical components for evaluating it: scope, generalization difficulty, priors, and experience. These elements shape the design of ARC-AGI, which aims to assess an AI model's ability to generalize and reason when faced with novel tasks.

The ARC Prize was launched in 2024, offering over $1 million to incentivize the development of AI systems capable of reasoning and generalizing like humans. For more details, visit www.arcprize.org.



For a long time, AI developers assumed that intelligence could be cultivated by progressing from specific to narrow to general. This assumption drove the development of chess-playing algorithms. However, this approach often diverges from the direct path to generalization, operating more orthogonally than linearly. Rather, this model of intelligence is like a waterfall that flows more effortlessly downward—a fish attempting to swim upstream faces far more significant challenges.

2.2 How to Navigate Generative AI Standards

Clear and actionable standards are essential for guiding generative AI development, but many standards are in development and some published standards might need updating or revision. In the United States, two administrations have proposed distinct sets of principles for AI regulation.

In the **2019 Memo on Regulation of AI**,
the Trump administration proposed ten principles:

1. Public Trust
2. Public Participation
3. Scientific Integrity and Information Quality
4. Risk Assessment and Management
5. Benefits and Costs
6. Flexibility
7. Fairness and Non-Discrimination
8. Disclosure and Transparency
9. Safety and Security
10. Interagency Coordination

In the **2022 Blueprint for an AI Bill of Rights**,
the Biden administration proposed five principles:

1. Safe and Effective Systems
2. Algorithmic Discrimination Protections
3. Data Privacy
4. Notice and Explanation
5. Human Alternatives, Considerations, and Fallback

Additionally, the **NIST AI Risk Management Framework (RMF)** and the **U.S. Department of State Risk Management Profile for AI and Human Rights** offer guidelines for organizations to manage risks associated with generative AI.

Familiarizing yourself with these principles and frameworks will help you navigate generative AI standards and AI governance in general. However, navigating generative AI standards also requires (a) understanding the typical lifecycle of a technical standard, (b) identifying published and developing standards from relevant standards organizations, and (c) staying informed on changes in the regulatory landscape across the public and private sectors. This holistic approach can help you adapt to evolving standards while fostering innovation and accountability in generative AI.

2.2a Standards Development Cycle

In the United States, the private sector often develops standards through voluntary consensus. While the process can vary across organizations, the typical lifecycle of a technical standard follows these stages:

Stage 1: Propose the Standard

The need for a new standard or the revision of an existing one is identified. The standard's scope, objectives, and intended impact are defined. Gaps, challenges, or opportunities are highlighted.

Stage 2: Task the Standard

A working group or committee of stakeholders, including experts, industry representatives, and regulatory bodies, is created. Roles, responsibilities, timelines, and deliverables for the group are set. Initial resources and requirements for drafting the standard are determined.

Stage 3: Draft the Standard

The initial content is developed. Research, simulations, technical analyses help build specifications and guidelines. Technical aspects are validated through prototyping or preliminary testing if needed.

Stage 4: Review the Standard

Feedback is solicited through public comment or targeted stakeholder reviews. Iterative revisions are made until conflicts and discrepancies are resolved and internal consensus is achieved.

Stage 5: Approve the Standard

The final draft is submitted to the governing body or authority for voting or ratification. Final issues are raised during the approval process. Consensus among the reviewing authority and stakeholders must be achieved.

Stage 6: Publish the Standard

The standard is formatted and released through official channels. Any supporting documentation, such as user guides or implementation tools, is distributed. The publication is announced through outreach to relevant industries and stakeholders.

Stage 7: Maintain the Standard

The standard is periodically reviewed for updates or corrections. Feedback and usage is monitored to assess effectiveness. Revisions are initiated as needed, and the lifecycle restarts at the drafting stage.

Stage 8: Retire the Standard

If revisions will not be sufficient to address issues of relevance and utility, the standard is retired. Guidance is provided for transitioning to updated standards. The standard should be archived for historical reference.

2.2b Published and Developing Standards

This breakdown highlights key published and in-progress standards from leading institutions, including NIST, IEEE, ISO/IEC, CTA, and ANSI. These standards aim to ensure trustworthiness, interoperability, and accountability while guiding the responsible development and deployment of generative AI systems.

Organization	Standards
National Institute of Standards and Technology (NIST)	• AI Risk Management Framework (AI RMF): A foundational document for managing risks associated with AI. • Assess the Risks and Impacts of AI (ARIA) Program: Evaluates generative AI systems' performance, worst-case scenarios, and real-world deployment through "Model Testing," "Red-Teaming," and "Field Testing."
Institute of Electrical and Electronics Engineers (IEEE)	The IEEE has initiated standards projects under its P7000™ series, addressing ethical and technical considerations, including: • P7000 - Model Process for Addressing Ethical Concerns During System Design: Provides a framework for integrating ethical considerations into the development of AI systems. • P7006 - Standard for Personal Data Artificial Intelligence (AI) Agent: Focuses on managing personal data in AI applications to ensure privacy and accountability.
International Organization for Standardization/ International Electrotechnical Commission (ISO/IEC)	• ISO/IEC 20546:2019: Overview and vocabulary for big data, a foundational component for training generative AI systems. • ISO/IEC TR 20547-2:2018: Use cases and requirements for big data reference architecture, critical for generative AI workflows. • ISO/IEC TR 20547-5:2018: Standards roadmap for big data systems. • ISO/IEC 5338:2023: Defines lifecycle processes for AI systems based on machine learning and heuristic methods, adapting general software and system lifecycle processes with AI-specific elements.
Consumer Technology Association (CTA)	The CTA is actively developing standards for generative AI, with a focus on healthcare applications.
American National Standards Institute (ANSI)	ANSI coordinates the development of voluntary consensus standards in the U.S., supporting the creation and adoption of generative AI standards across industries.

2.2c (De)regulation and Governance

Whether facing an increasingly regulated or deregulated landscape, generative AI will still operate within one or more governance frameworks. **AI governance** refers to the practices and structures designed to ensure that AI systems operate safely, ethically, and effectively. While regulation involves formal laws established by governments, governance incorporates regulatory compliance and extends to include decision-making frameworks, ethical oversight, and proactive risk management. Broader than technical standards, governance focuses on organizational and societal contexts.

The Governor Analogy

Kush Varshney, a distinguished research scientist at IBM, has compared AI governance to the function of a centrifugal governor. In 1788, James Watt adapted centrifugal governors to regulate the flow of steam in steam engines. In his 1868 essay "On Governors," James Clerk Maxwell described such mechanisms as systems that "[keep] velocity nearly uniform, notwithstanding variations in driving power or resistance."

The governor mechanism transcended engineering and found relevance across disciplines. In 1858, Alfred Russel Wallace used governors as an analogy for the evolutionary principle, noting that they "check and correct any irregularities almost before they become evident" in a paper that influenced Darwin's decision to publish *On the Origin of Species*. In 1979, Gregory Bateson expanded on the analogy in his work on cybernetics, emphasizing the role of adaptive, self-correcting mechanisms in managing complexity.

In systems theory and cognitive science, governors continue to illustrate dynamic systems where information representation and the processes acting upon it are inseparable. This kind of feedback loop is critical for modern AI governance, enabling corrective actions before risks become unmanageable.

AI Governance Frameworks

- **General Data Protection Regulation**: Governs data privacy and security across the European Union.
- **Organisation for Economic Co-operation and Development (OECD) AI Principles**: Promotes inclusive growth, transparency, and accountability.
- **European Commission's Ethics Guidelines for Trustworthy AI**: Focuses on ethical principles such as human agency, fairness, and privacy.
- **Corporate Ethics Boards:** Apply tools like model cards to document and mitigate AI risks.

Global AI Governance Initiatives

- **EU AI Act:** Classifies AI systems by risk and sets requirements for transparency, accountability, and fairness.
- **Canada's Directive on Automated Decision-Making**: Provides guidelines for assessing risks in government use of AI.
- **China's Interim Measures for the Administration of Generative Artificial Intelligence Services**: Focuses on transparency, data security, and ethical innovation in generative AI.

Leveling Up

LEVELING UP

Reasoning with Cheat Codes

⇧⇧⇩⇩⇦⇨⇦⇨

Reasoning with First Principles

A first principle is the most fundamental observation or assumption within a given domain, serving as a foundational truth. Aristotle associated first principles with *archē*, meaning the underlying substance or starting point of a thing. First-principles thinking involves breaking down a complex subject into its simplest components to uncover its essential nature. By identifying foundational truths, we can rebuild the subject from the ground up, often uncovering innovative solutions in the process. This assumption—that truth is revealed by breaking complex systems into smaller components—is also known as reductionism.

However, this approach often overlooks the dynamics of emergent phenomena, like intelligence, which arise from interactions within complex systems. In a complex system, the whole system cannot be fully understood by examining individual parts in isolation.

Additionally, before we can begin to break something down, we must already have a framework for identifying what constitutes the "whole" and its "parts." To analyze the "whole," we need an intuitive grasp of how the "parts" fit together. To define parts, we need to define the boundaries of the "whole." We ultimately rely on comparisons to define and interpret the relationships between parts and wholes.

Reasoning with Analogies

Reasoning with analogies involves drawing comparisons between two things to infer something about one based on knowledge of the other. It relies on abstraction—the process of identifying relationships or patterns and merging similar elements into a standard model.

According to François Chollet, analogies can be categorized into two types: value analogies and program analogies. Value analogies operate in wicked domains, where there are no clear rules or definitive answers, while program analogies operate in kind domains, which have well-defined rules and predictable outcomes.

Analogy plays a hidden but foundational role in identifying what counts as fundamental. If you view human thought as computational, it's because of an analogy to machines.

Reasoning with Recursive Refinement

In practice, first-principles thinking needs analogy to frame what's worth breaking down and to interpret the results. Analogy needs first principles to refine its accuracy and avoid overgeneralization. This interplay affords us a recursive refinement process, which involves iterating between foundational truths and comparative reasoning. Each pass through this cycle can lead to deeper insights and greater understanding.

In generative AI, recursive refinement supports processes like model design and optimization. Analogies guide initial architectures (e.g., transformers inspired by attention mechanisms in humans), while first principles validate computational efficiency and generalization. This player's manual applies recursive refinement as an ethical cheat code for advancing generative AI.

Facing the Final Boss

This is the part where, if generative AI development were a classic arcade game, this manual would tell you how to defeat the final boss. In those games, the final boss is rarely seen as a narrative resolution; more often it is seen as a test of skill—an opportunity to apply everything you've learned along the way. Maybe in this arcade of ours, the final boss would be named The Governor, a sly nod to *The Terminator*.

But here's the twist: mastering a skill becomes a story we tell ourselves. All games are driven by story in this way; narratives are how intelligent actors relate to their passage through time. In games, it is more a matter of how explicit or implicit the story is made—how much the game directs us and how much of it is left for us to figure out ourselves.

I worry that framing AI governance as the mastery of some singular, villainous "big bad" obscures the reality we face. The final boss isn't just a matter of mechanics—we must be intentional about the stories we choose and the systems we construct together. Just as in games, where meaning emerges from how we navigate their challenges, meaning emerges in AI development not only from what we build but also from how we govern and use it.

This is the challenge—and the beauty—of analogies. They simplify, illuminate, and inspire, but eventually, we must lay them aside. At the end of their lifecycle, they leave us to confront the messy, unanalogizable truths of the real world. Artificial intelligence isn't an arcade, and the stakes are far too high for us to approach it like a game. Yet, I hope someday we can look back on this analogy with fondness, as we do our favorite games and stories, and reflect on how it helped us claim prizes of insight and progress.

Before we retire, let's revisit the beginning, where we imagined you and a friend as collaborators in an arcade. This time, imagine you and me—reader and writer, user and developer—as collaborators. Together using standards and governance frameworks, we've constructed a mental model of AI systems built on the shoulders of technological giants like the internet and social media.

Now, let's zoom out further. Imagine all eight billion people on this planet as collaborators, reveling in the chaos and complexity of life on Earth as if we were in the largest arcade ever conceived. Every machine blaring, every game running, the intensity almost overwhelming. The games we play with information can feel disorienting.

This is where first principles help. They teach us to look at one thing at a time: a token, a game, a level, a boss, a prize. By isolating and understanding these pieces, we can find patterns and build meaning.

But this comes with a danger—the risk of thinking that none of it is real or that all of it doesn't matter. Meaning, value, and complexity don't live in isolation; they emerge in the space between things, in the relationships we forge and systems we design.

And this brings us to care and responsibility. Building arcade-like systems—whether in AI or in any domain—requires both. If we're fortunate, these systems can help us unlock new levels of care, responsibility, and insight, illuminating the spaces between the pieces we've so carefully examined.

The true final boss is how we govern ourselves. It's not a single climactic challenge; it's an ongoing process of deciding who we want to be and how we want to live—individually and collectively.

This manual has given you what you need to begin. The instructions are simple:

PRESS START TO PLAY.

Resources and Pick-Up Items

ARC Prize. 2024. “General Intelligence: Define it, measure it, build it.” YouTube. August 17, 2024. https://www.youtube.com/watch?v=nL9jEy99Nh0.

China Law Translate. 2023. “Interim Measures for the Management of Generative Artificial Intelligence Services.” https://www.chinalawtranslate.com/en/generative-ai-interim/.

Chollet, François. 2019. “On the Measure of Intelligence.” https://arxiv.org/abs/1911.01547.

Consumer Technology Association. 2024. “Standards.” https://shop.cta.tech/collections/standards/artificial-intelligence.

European Commission. 2019. “Ethics Guidelines for Trustworthy AI | Shaping Europe’s Digital Future.” https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Future of Life Institute. 2024. “High-Level Summary of the AI Act | EU Artificial Intelligence Act.” https://artificialintelligenceact.eu/high-level-summary/.

General Data Protection Regulation (GDPR) “General Data Protection Regulation.” https://gdpr-info.eu.

Government of Canada. 2019. “Directive on Automated Decision-Making.” https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592.

Knoop, Mike, and François Chollet. 2024. “ARC Prize.” https://arcprize.org.

NIST. 2021. “AI Risk Management Framework.” July 12, 2021. https://www.nist.gov/itl/ai-risk-management-framework.

Organisation for Economic Co-operation and Development. “AI Principles.” https://www.oecd.org/en/topics/sub-issues/ai-principles.html.

The White House. 2019. “Artificial Intelligence for the American People.” https://trumpwhitehouse.archives.gov/ai/.

The White House. 2022. “Blueprint for an AI Bill of Rights.” https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

U.S. Department of State. 2024. “Risk Management Profile for AI and Human Rights.” United States Department of State. July 25, 2024. https://www.state.gov/risk-management-profile-for-ai-and-human-rights/.

Vaswani, Ashish, et al. “Attention is All You Need.” https://arxiv.org/abs/1706.03762

Wolfe, Cameron R. 2022. “Deep (Learning) Focus.” https://cameronrwolfe.substack.com.

Wolfram, Stephen. 2023. “What Is ChatGPT Doing … and Why Does It Work?” https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

Credits

Creating this player's manual was a social, collaborative, and recursive process. I would like to thank:

my partner, for helping me see the world in better ways;
my parents, for taking me to arcades as a child;
my professors, for teaching me about writing;
my colleagues, for teaching me about AI;
and AI, for what I don't yet know.

GAME OVER

CONTINUE?