HUMANS FIRST:

A caring and careful adoption of artificial intelligence in civic technology

blue tiger

“Everything, everywhere, ALL AT ONCE”

Daniel Kwan and
Daniel Scheinert

The first robot to walk the earth was a bronze giant called Talos, programmed to defend Crete by hurling boulders to sink approaching foreign vessels. Hephaestus, an expert in invention and technology and master of the forge, created Talos (Mayor, 2018).

Fast forward thousands of years, and artificial intelligence is rapidly, pervasively, and persuasively invading all aspects of our lives, even the most intimate.

Today, hundreds of millions of people use large language models each week for tasks that are far more mundane (OpenAI, 2025a).

Ayrin, a 28-year-old woman, spends hours on end talking to her AI boyfriend. And other services that explicitly offer AI companionship have millions of users (Hill, 2025).

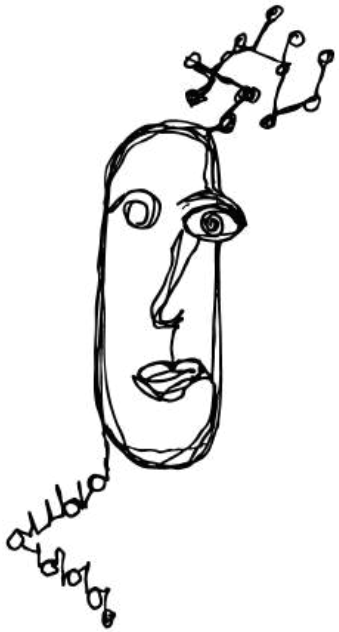
Like Icarus and his wings, who was too quick to adopt new technology, organizations are rapidly adopting solutions liberally sprinkled with artificial intelligence. The siren call of soaring to greater heights of efficiency, of doing more with significantly less, is too tempting to ignore. It is everything, everywhere, all at once.

But efficiency without dignity is just bureaucracy at machine speed. The speed at which your application process is adjudicated only matters if the person on the other end is treated well.

We are in a moment of unbridled enthusiasm for a technology most of us barely understand.

The stakes are higher in civic technology, which operates in highly sensitive domains, including social services, criminal justice, housing assistance, healthcare access, and child welfare. While wrong decisions can dramatically change the course of someone's life, so can slow decisions.

And so, how can we adopt AI in a caring and careful manner?



Efficiency without dignity is just

BUREAUCRACY AT MACHINE SPEED

TABLE OF CONTENTS

1. Understand the Technology
2. Adopt with Care
3. Build for Trust

Part 1

UNDERSTAND THE TECHNOLOGY

Today, the incantation of a few words brings to life images, video, poetry, or even music. Cars that drive themselves down busy city streets (Bojarski, 2016). Real-time translation so that you can make friends in Tierra del Fuego (SEAMLESS, 2025).

These were the things of fiction in the pages of Stephenson, Gibson, and Clarke. Now, they are our reality.

To use artificial intelligence effectively, we must first understand how it works.

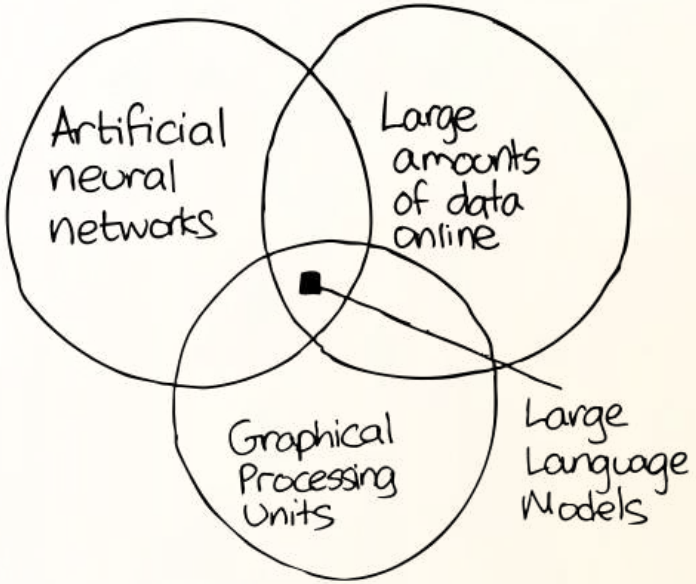


Figure 1: The three enablers of Large Language Models.

Underlying the illusion of magic are three things working together: recent advances in artificial neural networks, massive amounts of data available online, and powerful hardware called graphical processing units (GPUs). These enablers drive the latest artificial intelligence models.

> "Any sufficiently advanced technology is indistinguishable from MAGIC"
>
> Arthur C. Clarke

How AI learns

Synapses connect neurons in your brain in a massive network. Each neuron is relatively simple, but thousands of them work together to operate your body, and plan brunch.

An artificial neural network is a type of machine learning algorithm that is inspired by this behavior. It's a system that learns patterns from data.

You don't tell a neural network what to look for. You show it thousands or millions of examples, and it figures out which patterns matter on its own.

For example, show a neural network enough photos of tigers, and it will learn to recognize features like stripes, fur texture, and body shape. Show it enough written language, and it will learn grammar, style, and context. Neural networks are especially good at image recognition, understanding and generating human language, and predictive analytics.

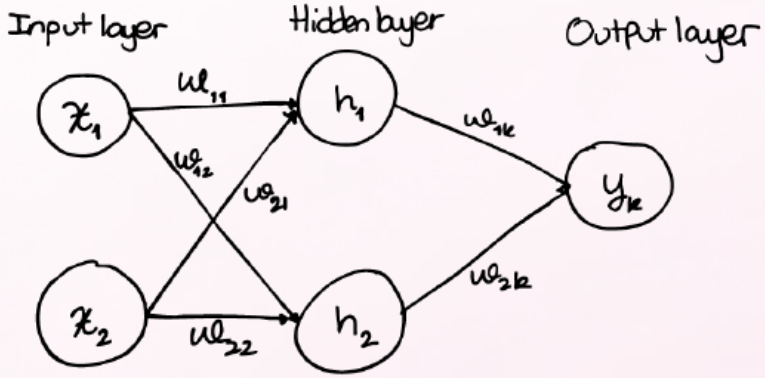


Figure 2: A neural network consists of an input layer, multiple hidden layers (these are layers that process information), and an output layer. This neural network has one hidden layer.

Large Language Models (LLMs)

To understand and generate the creative complexity of natural language, neural networks need to be incredibly powerful.

Deep learning networks are neural networks with many layers (sometimes up to thousands). Each layer builds on the understanding of the previous layer. This enables deep learning networks to model complex relationships.

These deep learning networks only became practical when GPUs, originally designed for video game graphics, proved to be exactly what was needed to train networks at scale. GPUs are designed to execute a large number of mathematical operations in parallel (Khan, 2020).

The most important breakthrough was the transformer, a type of deep learning network that can 'pay attention' (Vaswani, 2017). This makes them great at handling long paragraphs and code. The GPT in ChatGPT stands for 'Generative Pre-trained Transformer.'

Large amounts of data

Humanity's knowledge has never been as open and easily accessible as it is today. There are billions of web pages, Wikipedia articles, books, music, art, code, and academic and government datasets ready for consumption. Transformers hungrily learn from all of this data.

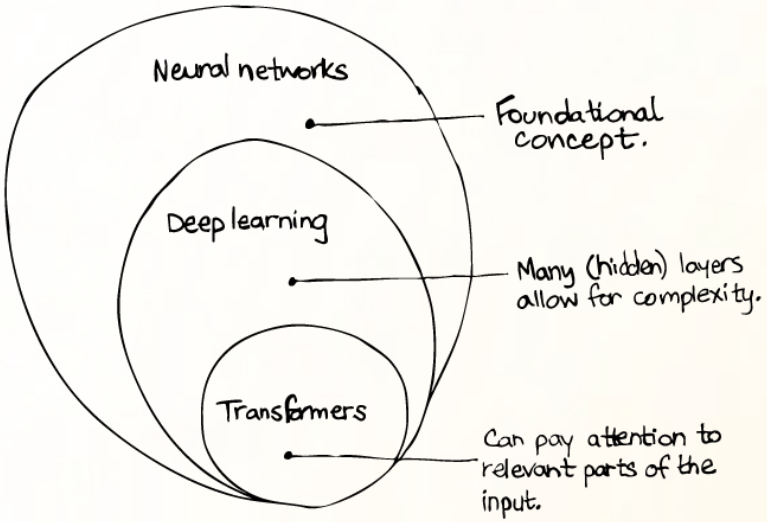


Figure 3: The conceptual relationship between neural networks, deep learning, and transformers.

What AI actually does

And so we're in a moment where artificial intelligence is seemingly capable of so much.

But here's the thing: Despite recent advances in "reasoning models" that think step-by-step, these systems remain fundamentally statistical pattern-matchers.

When an AI writes a paragraph that sounds insightful, it has merely predicted word by word what an insightful paragraph would look like based on patterns in its training data.

The illusion of reasoning makes careful oversight even more critical in high-stakes environments, like the government.

AI has **MERELY PREDICTED** word by word what an an insightful paragraph would look like.

“I’M HAVING A HALLUCINATION NOW, I don’t need drugs for that”



Thomas Pynchon

Also, LLMs hallucinate. A hallucination occurs when a response appears plausible but is actually entirely fabricated (Ji, 2023). There are a few different types of hallucinations, for example:

Factual: Making up things that never actually happened.

Source: Fabricating citations, quotes, URLs, or book titles, even when the underlying fact might be correct.

More specifically, a hallucination is when an AI is confidently incorrect. LLMs simply predict the next word in a sequence of characters. They have no concept of whether what they are reading or generating is accurate (Kalai, 2025).

Hallucinations are not a temporary bug that can be fixed with each subsequent release, they are inherent to how these systems work (Kalai, 2025).

And more concerning is the fact that the problem may be getting worse, not better. OpenAI's own evaluation of its most advanced reasoning models found that they hallucinate more frequently than their predecessors, in some cases at more than double the rate of earlier models (OpenAI, 2025b).

Potential for significant harm

In addition to hallucinations, AI systems deployed without appropriate guardrails can cause significant harm, for example:

- A New York City chatbot repeatedly told landlords they could discriminate against Section 8 voucher holders (Lecher, 2024).

- Michigan's unemployment system falsely accused 40,000 people of fraud, resulting in a 93% error rate. The automated system seized tax refunds and garnished wages (Angwin, 2020).

These are not isolated failures. Researchers have documented how automated systems in public services consistently produce disproportionate harm for people who are already marginalized (Eubanks, 2018; Benjamin, 2019).

AI is here to help

This is not entirely a cautionary tale. One can not ignore the vast benefits of artificial intelligence. Its ability to augment human capabilities today to solve both minor and the most important problems of our time is too powerful to ignore.

Consider a state Medicaid office processing millions of claims a year. A human reviewer might catch a handful of suspicious billing patterns in a week. An AI system can flag thousands.

At a VA medical center, AI can synthesize information across hundreds of pages of Veteran medical records and service history. This can free up valuable provider time.

These are the kinds of tasks where AI shines: detecting policy violations at scale across thousands of documents, flagging abnormal patterns in healthcare data, and categorizing and prioritizing constituent feedback so that urgent requests don't sit in a queue.

AI can handle the volume, so that humans can focus on the decisions and tasks that require empathy, context, and care.

Part 2

ADOPT WITH CARE

Technology is a strange beast in organizations that serve the public. In the front of house, it must emerge quietly when needed, and disappear when its job is done.

No one wants to chat with your bot, train your algorithm, or understand how to upload perfectly machine-readable documents. They just need help when they are unemployed, sick, or to feed their kids.

> "Our inventions are wont to be pretty toys, which distracts our attention from SERIOUS THINGS"
>
> Henry David Thoreau

But behind the scenes, technology is ever-present. It must be actively cared for: Constantly updated, iterated, patched, and fed. Forever, it must continuously enable organizations to do more and better with less.

And today, AI is the most complex and unique of these technologies. It is probabilistic, adaptive, and mostly opaque. It requires a caring, careful adoption that prioritizes humans first.

Start with the outcome

In the anxiety and excitement of adopting emerging technologies, many organizations brainstorm lists of use cases. These use cases turn into procurements, then pilots, and in a fantastic feat of metamorphosis, production. The outcome of a pilot is rarely the rejection of a technology.

An alternative approach is to start with a clear definition of a desired user and policy outcome. For example, no one wants to file their taxes. Instead, they want to maximize the income they can keep while avoiding penalties.

As the Design Justice framework reminds us, the people most affected by a system should help define what a good outcome looks like (Costanza-Chock, 2020). This starts with learning from existing and potential users, and leads to a synthesis and identification of the outcomes they want to achieve.

These are then paired with metrics that matter.

1. Write a clear outcome statement
2. Identify metrics that matter
3. Baseline those metrics

A service design approach

Users move through a series of transactions across one or many agencies to achieve their goals. When that journey breaks down, the problems rarely originate at the customer interface. A poor user experience is generally a symptom of issues deeper within your organization.

Service design is a holistic approach that focuses on understanding all the elements that deliver an experience: people, policy, processes, technology, and physical touchpoints (GDS, 2022).

Taking a service design approach will help you understand how policy, decisions, and data flow through your organization. And how those might affect the customer experience.



A poor user experience is generally a **SYMPTOM OF ISSUES** deeper within your organization

Favor simple, sustainable solutions

Only once you've identified the problem can you start thinking about potential solutions.

Policies and procedures are often too complex for technology to implement easily. Looking upstream for solutions will yield better and more sustainable results in the long run (Pahlka, 2023).

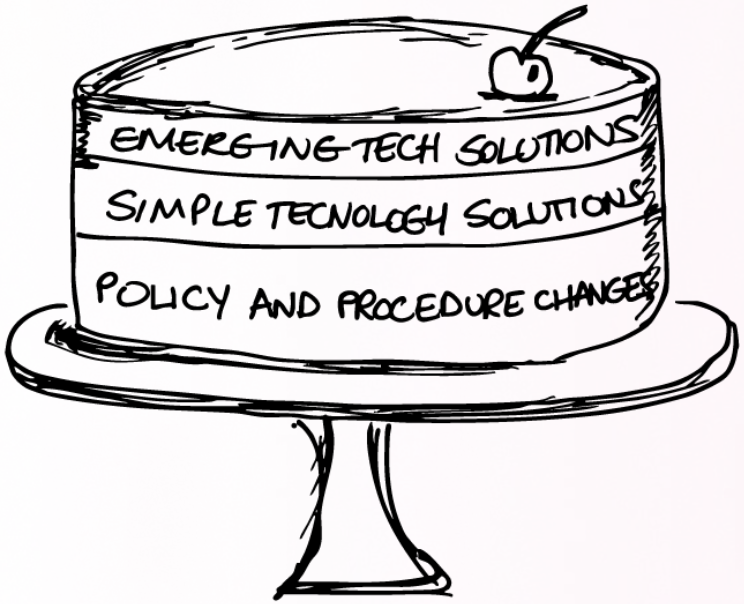


Figure 4: The layer cake of possible ways to solve a problem.

You should consider policy and procedure changes first, basic technology solutions second, and only then consider any emerging technologies, such as AI or machine learning.

Simpler, more established solutions might solve the problem better than AI, be easier to develop, and relatively cheap to maintain (GDS, 2025).

More complex solutions are expensive to operate. The economics, however, are shifting in counterintuitive ways. The cost of a single AI query has fallen dramatically. Meanwhile, the rise of 'reasoning' models and agentic workflows has increased the overall cost.

These systems consume more resources due to their complexity, consuming ten to a hundred times more tokens per task (Holter, 2025).

There are additional operational costs to AI, such as clusters of GPUs for adjacent libraries, per usage (token) costs, and licenses for access. And as venture capital subsidies that have kept many AI services artificially cheap begin to dry up, the drive to become profitable will further increase costs for the average user.

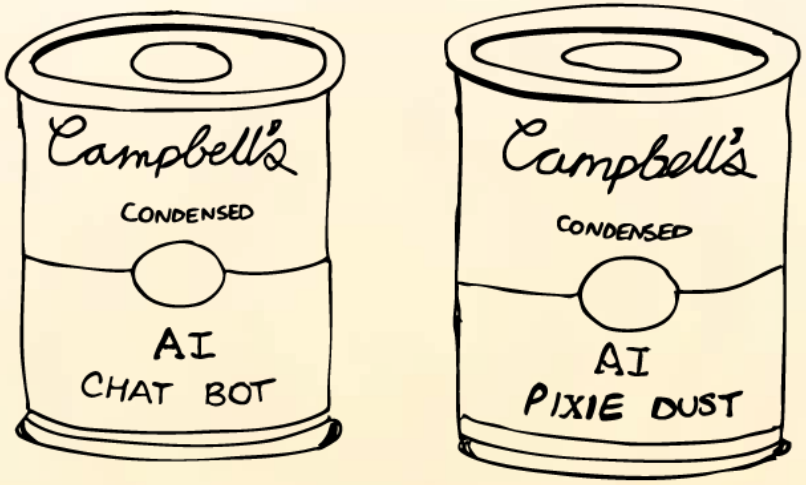
These more complex systems are also hard to maintain. While AI technology is constantly improving, it is also constantly changing. Keeping up with new advancements becomes an ongoing task for a development team.

And when evaluating AI solutions, be sure you understand what happens when the introductory pricing ends, who owns the data and the model outputs, how fine-tuning works, and whether you can switch AI providers without disrupting workflows.

1. Can policy or procedure changes solve this?
2. Can basic technology solve this?
3. Is AI necessary and appropriate?

The bottom line here is that not every problem should be solved with technology. There are situations where introducing AI will actively make things worse, for example:

- If the problem is fundamentally about power or resources, AI can't solve issues like underfunding of public services or a lack of political will.
- If there is a lack of representative, high-quality data or data reflecting past discriminations, AI will simply amplify these problems.
- If trust is low, introducing AI systems in communities that have experienced surveillance or discrimination can deepen mistrust.



Part 3

BUILD FOR TRUST

Building trust requires multiple overlapping layers of protection. Humans who care, technical guardrails, and continuous evaluation collectively reduce the risk of any single failure causing harm.

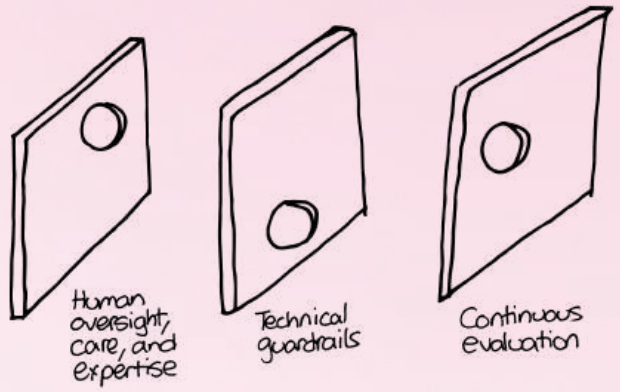


Figure 5: Layers of protection, organized from the most human to the most technical.

The European Union's AI Act, the first comprehensive AI law, classifies AI systems used to evaluate eligibility for public benefits, healthcare services, and law enforcement as 'high-risk,' requiring risk management, human oversight, transparency, and data governance (EU 2024).

In the United States, the Office of Management and Budget requires federal agencies to implement risk management practices for "high-impact" AI, including testing, assessments, human oversight, and remedies for individuals affected by AI-enabled decisions (OMB, 2025).

Start with humans who care

Civic technology operates in sensitive domains like social services, housing assistance, healthcare access, and child welfare. The people interacting with these systems are often vulnerable, stressed, or traumatized (Dietkus, 2022).

A poorly designed AI system can make this worse. For example, chat bots can repeatedly ask about traumatic experiences, and systems can demand documentation that survivors don't have.

These questions are worth asking repeatedly throughout the process:

- Do users feel safe and respected?
- Are they able to access services they are eligible for?
- Does the system reduce stress?
- Are outcomes equitable across different populations?

Trauma-informed design and care professionals help teams understand how to design, build, and operate systems while lowering the risk of retraumatization (Dietkus, 2022; SAMHSA, 2026).

- Clear non-threatening language
- Multiple ways to provide information
- Respectful interactions at each step
- Transparency and honesty

Assemble diverse, cross-functional teams

AI systems perpetuate societal biases and it does humanity no favors to pretend that these biases do not exist (Mehrabi, 2021).

The consequences are well documented: automated systems have denied benefits to people who deserved them, flagged communities of color for fraud at disproportionate rates, and embedded historical discrimination into decisions that feel objective precisely because a machine made them (Eubanks, 2018; Benjamin, 2019; Noble, 2018).

Mitigating this bias requires intentional work. This starts with ensuring that you're building a diverse product team, because the more diverse your team, the more likely it is to recognize and mitigate biases in your experiment design, data, and algorithms.

Work with subject matter experts

AI systems are often trained to mimic the actions of subject matter experts. With the current speed, prevalence, and vast capabilities of AI systems, it's more important than ever to embed subject matter experts on your team to ensure you develop accurate actions and responses.

Invest in AI literacy

Everyone involved in the design, procurement, or oversight of AI systems needs to understand what these systems can and can not do.

Build trustworthy systems

To help build trust in an AI-system, you should:

- Show the sources of information and data used in the decision-making process. This helps users independently verify results.
- Build interfaces that help show how the AI arrived at its conclusion. For example, in image classification, have the AI highlight which parts of the input drove classification.
- Benchmark your system against established industry standards. This allows you and your users to compare systems using a standardized test.
- Make it easy for users to report any issues in your system. Deal with those issues promptly.
- Provide confidence that an AI system won't create inappropriate outputs, harm or mislead users.

"TRUST is like a mirror. You can fix it if it's broken, but you can still see the crack."

Lady Gaga

The confidence you can provide users comes through implementing visible guardrails. Safety filters prevent harmful outputs, such as instructions for dangerous activities or illegal content. And alignment guardrails ensure that systems behave in accordance with human values and intended purposes.

Essential human oversight

Build your systems with interfaces that allow humans to periodically check routine work, override system decisions, and handle cases that require human intervention. Critical applications, like healthcare or benefits decisions, should let humans retain final decision authority.



This kind of trust-centered design is already happening. The United Kingdom's Caddy acts as a copilot for service agents rather than replacing them. It draws from verified government sources, shows where each answer comes from, and flags for human checks when needed (OECD, 2025).

Be extra careful with AI agents

AI systems are increasingly capable of not just generating answers, but also taking autonomous actions, like, filing forms, making eligibility determinations, sending notifications, or updating records.

If an agent hallucinates, it can act on that hallucination autonomously and make an irreversible decision. Guardrails, human oversight, and continuous evaluation become exponentially more important when AI systems act and not just advise.

If your organization is considering agentic AI, ensure that no autonomous action can adversely affect a person's benefits, eligibility, or rights without additional human review.

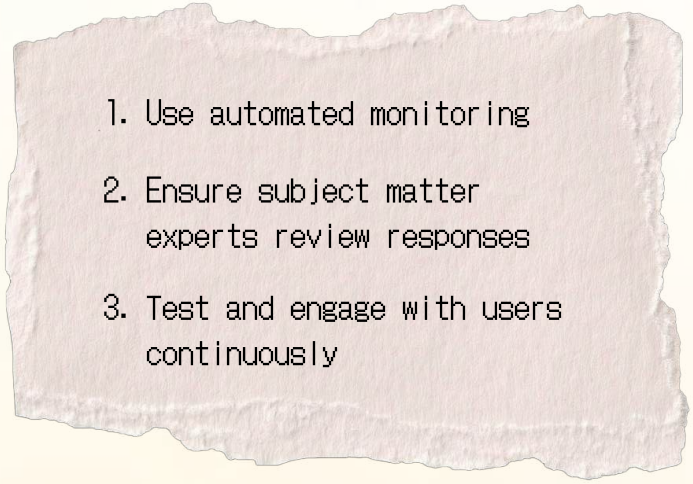
Evaluate continuously

Unit testing works well for deterministic systems where the same inputs will repeatedly produce the same outputs. However, AI systems are probabilistic and context-dependent. The same input might yield different but valid responses.

You need to implement different tests:

- **Accuracy:** Does the system provide the correct answers for your environment and scenarios?
- **Safety:** Can the system be manipulated into generating dangerous content?
- **Alignment:** Does the system's behavior match intended values? Check for biases, evaluate tone and helpfulness, and ensure that it acknowledges uncertainty.
- **Behavioral coherence:** Does the system maintain appropriate context across conversations? Does it contradict itself?

AI systems are constantly evolving, so continuous evaluation in both pre-production and production environments is necessary.



Your evaluation framework is also a learning mechanism. When certain types of queries lead to unexpected outputs, you can improve the training data. When guardrails trigger too aggressively, you can refine the rules.

AI systems are probabilistic and CONTEXT-DEPENDENT

Usability test often for trust

During your continuous discovery and usability testing sessions, check in with users explicitly about trust. You want to learn things like:

- What percentage of AI recommendations do users accept?
- How often do users seek second opinions or verification?
- How long do users spend reviewing AI outputs vs. human outputs?
- Can users articulate the boundaries of where the AI is predictably incorrect?

Continuous evaluation is both about demonstrating catching failures as well as demonstrating care and accountability.

You build trust when you can show users and stakeholders that you're systematically monitoring for problems, measuring impact, and responding to issues.

Generate synthetic data

Continuous evaluation requires realistic synthetic datasets: artificial data designed to mimic real-world data without exposing sensitive information such as personally identifiable information, health records, or financial data.

For example, if you're testing a benefits eligibility system, you need thousands of realistic but fictional applications, each with missing documents and temporal inconsistencies, because that's what real applications look like. You'll need this data to validate non-deterministic behaviors, such as biases and hallucinations.

Generating or anonymizing this data takes time, especially when it includes documents, images, and health records. Start early and maintain it with the same care you maintain software.

Humans first:
A caring and careful adoption of AI

Unlike other technological trends, AI is here to stay. The promise of doing more with less is too attractive to ignore.

AI is merely a tool. It reflects the values, biases, and priorities of those who deploy it.

Organizations that will succeed with AI aren't those that deploy it fastest, but those who deploy it thoughtfully.

Remember, no single layer is foolproof: humans who care, diverse teams, subject matter experts, technical guardrails, continuous evaluation, and time to learn. These things together form a system of protection worthy of the people you serve.

On the other end of every system is a person trying to feed their family, go to college, or rebuild their life after a crisis. They deserve systems that work.

Technology exists to free humans for the **MOMENTS THAT MATTER MOST**

Bibliography

Angwin, J. (2020, July 24). The seven-year struggle to hold an out-of-control algorithm to account. *The Markup*. https://themarkup.org/newsletter/hello-world/the-seven-year-struggle-to-hold-an-out-of-control-algorithm-to-account

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.

Bojarski, M., et al. (2016). End to end learning for self-driving cars (arXiv:1604.07316). *arXiv*. https://arxiv.org/abs/1604.07316

Costanza-Chock, S. (2020). *Design justice: Community-led practices to build the worlds we need*. MIT Press. https://designjustice.org

Dietkus, R. (2022). The call for trauma-informed design research and practice. *Design Management Review*, 33(2), 26–31. https://doi.org/10.1111/drev.12295

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L series. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Government Digital Service. (2022, April). Understanding and meeting policy intent. *GOV.UK Service Manual*. Retrieved November 18, 2025, from https://www.gov.uk/service-manual/design/understanding-and-meeting-policy-intent

Government Digital Service. (2025). Using artificial intelligence (AI) in services. *GOV.UK*. Retrieved October 17, 2025, from https://www.gov.uk/service-manual/technology/using-artificial-intelligence-ai-in-services

Hill, K. (2025, January 15). She is in love with ChatGPT. *The New York Times*. https://www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html

Holter, A. (2025). AI costs in 2025: Cheaper tokens, pricier workflows. https://adam.holter.com/ai-costs-in-2025-cheaper-tokens-pricier-workflows-why-your-bill-is-still-rising/

Ji, Z., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. https://dl.acm.org/doi/full/10.1145/3571730

Kalai, A., et al. (2025). Why language models hallucinate. *arXiv*. https://arxiv.org/pdf/2509.04664

Khan, S. M., & Mann, A. (2020). AI chips: What they are and why they matter (Issue Brief). Center for Security and Emerging Technology. https://cset.georgetown.edu/wp-content/uploads/AI-Chips%E2%80%94What-They-Are-and-Why-They-Matter-1.pdf

Lecher, C. (2024, March 29). NYC's AI chatbot tells businesses to break the law. *The Markup*. https://themarkup.org/news/2024/03/29/nycs-ai-chatbot-tells-businesses-to-break-the-law

Mayor, A. (2018). *Gods and robots: Myths, machines, and ancient dreams of technology*. Princeton University Press.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35. https://doi.org/10.1145/3457607

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

OECD. (2025). *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions*. OECD Publishing. https://doi.org/10.1787/795de142-en

Office of Management and Budget. (2025). Memorandum M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust. https://www.whitehouse.gov/wp-content/uploads/2025/02/M-25-21-Accelerating-Federal-Use-of-AI-through-Innovation-Governance-and-Public-Trust.pdf

OpenAI a. (2025, September 15). How people are using ChatGPT. https://openai.com/index/how-people-are-using-chatgpt/

OpenAI b. (2025, April 16). O3 and O4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf

Pahlka, J. (2023). *Recoding America: Why government is failing in the digital age and how we can do better*. Metropolitan Books.

Substance Abuse and Mental Health Services Administration. (2026, February 8). *Trauma-informed approaches and programs*. U.S. Department of Health and Human Services. https://www.samhsa.gov/mental-health/trauma-violence/trauma-informed-approaches-programs

SEAMLESS Communication Team. (2025). Joint speech and text machine translation for up to 100 languages. *Nature*, 637, 587–593. https://doi.org/10.1038/s41586-024-08359-z

Shojaee, P., et al. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv.* https://arxiv.org/abs/2506.06941

Vaswani, A., et al. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). https://arxiv.org/abs/1706.03762

Writing and Illustration
Shashank Khandelwal

Editing, Layout and Design
Tyler Gindraux

A special thank you to our reviewers:
Shweta Bansal, Rachael Dietkus,
Frances Ruiz, Holly Syndor