Worlds apart

How world models and context will bring on the next wave of **trustworthy AI**

Capgemini

Statistically impressive, but individually unreliable

"Statistically impressive, but individually unreliable" – John Launchbury, Director of I2O, DARPA, sums up the feeling that many people currently have about AI. Impressive as it is overall, we are sometimes still reluctant to trust it with critical decisions.

The missing piece of the puzzle is context. Among humans, we take context for granted. You don't need to tell your barber that your goal is to look good; that's universally understood. Current AI needs that kind of information stated explicitly. Providing AI models with context goes a long way to preventing these misalignments, leading to better AI. Shared context lays the foundation for trust.

The implications for organizations are significant. Context becomes an asset, alongside data. In practice, this is the role of the semantic layer: the place where organizational context is structured, maintained and made usable by AI systems. Trust, safety, compliance, and robustness become properties that can be engineered deliberately. The result will be AI systems that users can turn to with confidence, opening the door to a far wider range of applications.

When AI systems operate without proper context, misalignment with the real world is inevitable. With context, we gain a valuable partner. One that's not only statistically impressive, but on which we can confidently rely.

Rules of the road

When AI is missing context

In the early morning hours of November 30th, 2025, a self-driving Waymo vehicle made history for all the wrong reasons. It could have been a scene from a Hollywood action movie: police cruisers blocking an intersection, lights blazing, and a suspect awaiting arrest. Then a white robotaxi rolled into the scene. It casually approached the stunned officers, signaled, and then continued along on its route. Any human would have recognized the danger and kept a safe distance. The AI was oblivious to the human drama.

The Waymo vehicle wasn't broken. In fact, it was operating flawlessly – obeying traffic laws, scanning for obstacles, and searching for a safe and legal path through the intersection.

For all its spectacular feats, there are situations where AI is so out of alignment with our expectations that we start to question whether we can trust it. So what's going on here? Why is an AI failing to recognize a situation so basic, that any child could understand it?

The answer goes straight to the biggest limitation in AI today, and hints at the potential to unlock the next level of real-world performance.

Why world models are not enough

While “world model” might be the prevailing term in the industry, it’s worth reflecting on whether this might narrow our perspective, and whether “context” might be a more comprehensive term.

The term “world model” has more physical connotations, when that’s only one part of what we’re talking about. Topics like ethics, laws or social expectations are every bit as important for an AI as, say, the laws of physics.

Proponents of the term “world models” would be quick (and correct) to point out that, within AI circles, the term is used in the broader sense of the world of information that our system exists within. But we already have a word that means just that – context.

World models and contextual models play complementary roles. A world model allows an agent to simulate how the environment might evolve, giving it the ability to anticipate outcomes. A contextual model grounds the agent in the present moment, helping it understand what matters in its current situation and respond with relevance and nuance.

We’ll be using both terms freely in this paper. But we must keep in mind that world models refer to many different worlds, each one brimming with unseen complexity.

The gap at the heart of AI

The AI field has spent the last decade prioritizing scale. Bigger datasets, larger models, deeper networks, more parameters. The logic is seductively simple: if intelligence can be approximated through statistical pattern matching, then more patterns should yield more intelligence. It is a compelling story, and in narrow domains it has produced extraordinary results.

Yet the more we push this strategy, the clearer its ceiling becomes. Modern AI systems appear brilliant in familiar territory but become brittle the moment the world refuses to behave like the data they were trained on. They struggle in the presence of missing information. They hallucinate with confidence when context shifts. They misunderstand human intent because humans rarely say everything explicitly. They operate like tourists who've memorized the phrasebook but never learned the culture. What these systems lack isn't more raw data; it's context.

> *The path to trustworthy AI runs not through bigger models but through better context.*

Context is the unwritten rules and causal relationships that govern how the world behaves, even when nobody has bothered to state them out loud. It is what provides an entity (human, animal, or AI) with an awareness of how its reality works. In the field of AI, it's part of a broad conceptual shift, beyond pattern recognition and towards actual reasoning. In order to reach that stage, AI will need to understand some fundamental facts about the world.

This paper argues that the path to trustworthy AI runs not through bigger models but through better context. The future of AI lies in systems capable of operating beyond their training distribution: combining statistical learning with structured knowledge, causal inference, physical constraints, and symbolic reasoning. These hybrid systems behave less like autocomplete engines and more like participants in a world governed by rules.

The third wave of AI

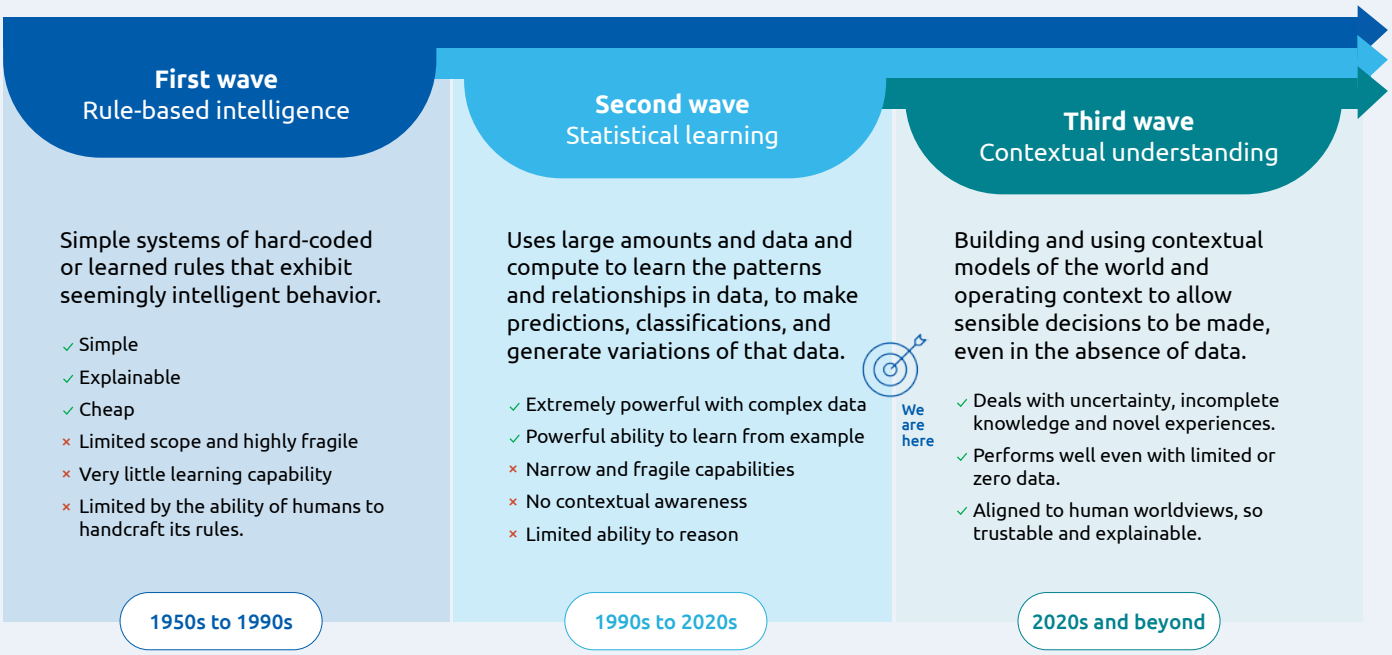
In 2017, John Launchbury outlined a vision for the ***evolution of AI***. The first wave of handcrafted AI rules would give way to a second wave of statistical learning, before finally reaching a third wave of contextual adaptation. In this third wave, he suggested, systems would understand data in terms of context, giving them the ability to reason abstractly. From our current vantage point at the far edge of wave two, we're able to see a number of signals that the third wave is approaching fast:

- The gains from purely scaling LLMs are diminishing.
- It's clear that hallucinations are a feature, not a bug, and can never be fully trained or prompted away.
- Safety and compliance requirements demand systems that behave predictably.
- Robotics and autonomous systems require models grounded in physics, spatial awareness, proprioception, causality and more.
- Enterprises need AI that can integrate with structured business logic, policies, regulation and laws.
- Decision-making tasks require causal, not correlational, reasoning.
- AI lacks the higher-level abstraction, understanding and reasoning capability to implement meaningful ethical controls.

These limitations all stem from a lack of context. We are now entering this third wave of AI where systems combine statistical power with explicit reasoning over structured knowledge. The result is AI that is contextual, grounded, and capable of abstraction.

From rules to reason: the three waves of AI maturity, adapted from John Launchbury

AI has evolved from rigid, rule-based systems to data-driven prediction, and is now entering an era of contextual understanding.



What exactly is context?

Imagine you're planning a trip to a new city. Before you go, you look at a map, read a travel guide, and maybe ask friends for tips. You build up a mental picture of what the city is like, where the landmarks are, how the subway works, which neighborhoods are safe, what the local laws, customs and etiquette are, and where to find good food. This mental map isn't just a list of facts; it's a living model that helps you make decisions, like how to travel around best or how to avoid offending the locals.

In the same way, context for an AI is like its own internal guidebook, built from experience and information. It helps an AI system manipulate its environment effectively, predict what might happen next, and choose the best actions, even if it's never been in that exact situation before.

That's the big picture of context. Now let's give it a precise definition, one that distinguishes it from all the other data that AI already has access to.

Data is recorded observation – measurements, events, or signals captured as symbols. On its own, it describes what was captured, but not what it means.

Context is the frame that makes the data intelligible, meaningful and actionable: how it was generated, for what purpose, under which assumptions, and within which social, physical or organizational setting. The same data can mean different things in different contexts. In short, context is the set of constraints, assumptions, and world models that determine how those symbols are interpreted and acted upon. In enterprise systems, this frame is increasingly embodied in what is called the semantic layer: the layer where meaning, relationships, constraints, and business definitions are made explicit and managed.

In the autonomous driving example we began with, the AI system presumably saw the police cars, the officers and the man on the ground. But it also saw palm trees and tall white buildings and chain-link

fences and all sorts of other things. What it didn't see was a situation that warranted extreme caution. In fact, it didn't see a "situation" at all.

That's what makes the question of context so important, and so much more difficult to solve than it might initially appear. One might be tempted to think that the solution here is to present the Waymo's AI with scene after scene of police actions, until it learns to avoid the police. And that might solve for this narrow issue, but it may cause other unforeseen issues, such as not stopping for the police or making space for emergency vehicles. It also won't help with other visually ambiguous situations that require context to make sense of. What we really want the AI to do is act like we do: to understand the full context, causality and consequences of the situation around them, then act according to that situation.

AI won't achieve human-level intelligence through more data or better pattern recognition. Intelligence is not a problem that can be brute-forced. What makes humans smart is not that we have seen large amounts of data – it's that we can contextualize the data that we do sense. The smartest people on the planet are not the people who have seen most things, they are the people who are able to digest and make links between diverse concepts.

Illusory intelligence

Large language models have been nothing short of a revelation, and within their "sweet spot" of language manipulation, generation and translation, they are unequalled. However, the performance of generative AI has created an illusion: that linguistic eloquence implies conceptual understanding. LLMs do not "know" in the human sense. They encode correlations between symbols; but this statistical foundation becomes a limitation when we try to apply it to situations where they need to reason about how the world works. This disconnect becomes obvious in edge cases:

- A model that can describe a physics problem eloquently but cannot perform basic causal reasoning.
- A chatbot that answers medical queries confidently but does not adhere to accepted clinical practices.
- A code generator that writes elegant functions but misinterprets the purpose and intent of the system around them.
- A planning model that cannot connect its own subtasks within the wider collective strategy.

The failure mode is consistent with a model that treats the world as if it were a mirror of its training distribution. When the world deviates, and it always does, this potentially world-changing technology falls into traps that humans with even simplistic contextual understanding would avoid. This introduces governance challenges that need to be addressed for broader adoption.

Trustworthy AI cannot rely on probability and hope to avoid mismatches between model assumptions and real-world dynamics.

Seven types of context – our PLANETS framework

What constitutes necessary context varies from one AI system to another. Context may contain any or all of the elements of our PLANETS framework:

Physical and natural laws: The "hard" constraints. Gravity, thermodynamics, chemistry, biology, and the fundamental mechanics of how the universe operates.

Legal and regulatory frameworks: Local and international laws, intellectual property, safety standards, and formal governance.

Axiology and ethics: Right vs. wrong, human rights, philosophical principles, and the weight of different outcomes.

Norms, social and cultural: The "soft" constraints. Etiquette, traditions, language nuances, religious customs, and unwritten expectations of behavior.

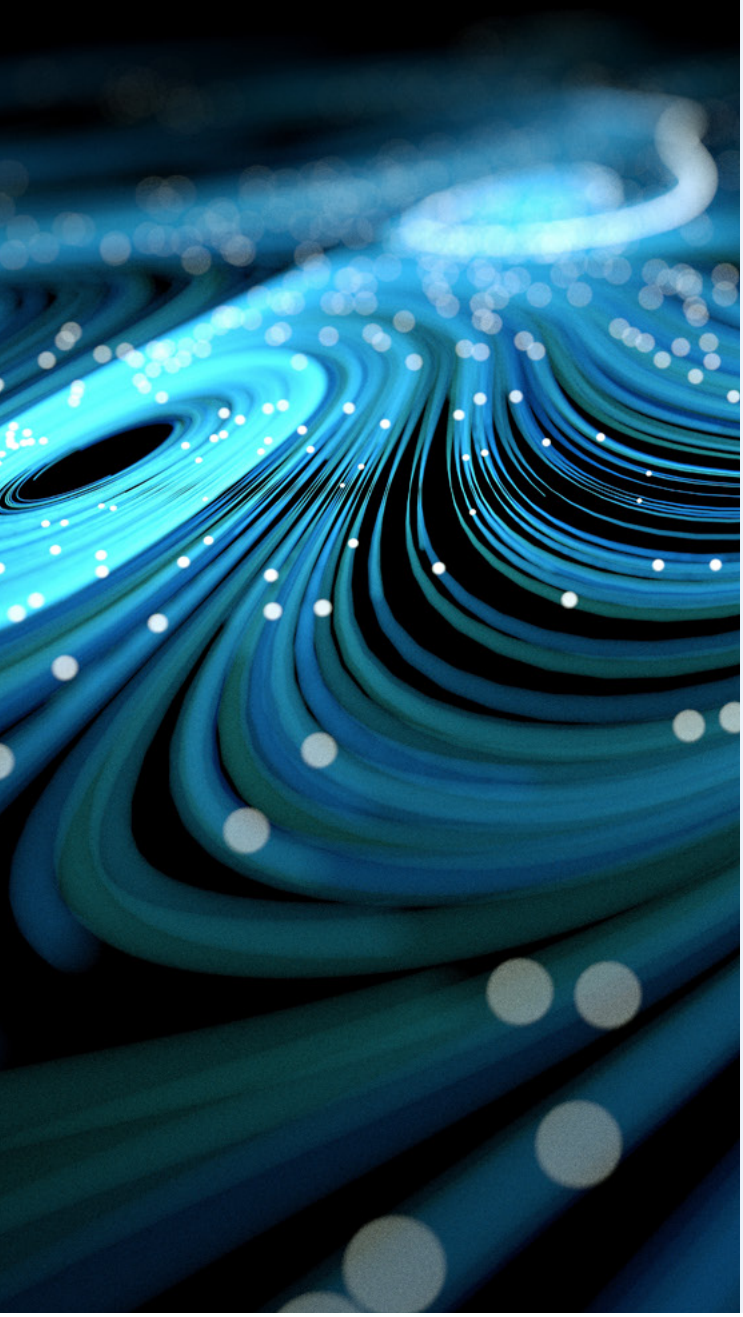
Economics: Markets and scarcity, supply and demand, game theory, financial systems, and the allocation of finite resources.

Technology and infrastructure: The built world. Digital protocols, hardware limitations, urban layouts, and the interconnected systems humans have constructed.

Situational context: Current events, historical precedents, and the "who, what, where" that changes how other rules apply.

Context need not encapsulate the whole world but merely the world in which the entity operates. An autonomous taxi doesn't need to understand economics, but it does need a grounding in physics, ethics, social norms and situational context.

Is bigger better?



Imagine a person working on a simple production line, screwing caps onto bottles. That worker's world is simple – bottles, caps and conveyor belt, with no external context necessary to fulfil their objective perfectly. At the more extensive end of the spectrum, an urban planner or engineer would need a much larger world model, including construction, agriculture, geography, economics, politics, and long-term demographic evolution to be effective in their profession. In AI, the same spectrum exists for context. A robotic arm needs a physical model of torque, friction and gravity; a clinical assistant needs a causal model of disease, uncertainty and human emotion.

Context can be very general or highly specific. The key is to create AI systems that are better aligned with human expectations, and ultimately more trustworthy. Models supported by context will be able to:

- Reason about unobserved variables
- Predict the consequences of actions
- Simulate hypotheses
- Learn from experience rather than memorization
- Generalize beyond their training distributions

In the human world, good personal assistants can decline meetings, rearrange travel, and make many other decisions without asking, because they have a shared understanding of their boss's priorities, availability, and preferences. That's not due to "accuracy," but because there is a shared world model.

The data debate

When we understand how a gap in contextual grounding is impeding AI, we can also see why the race for ever more data has not delivered the performance everyone hoped it would. The mainstream narrative in AI assumes that more data and bigger models lead directly to better AI. In fact, what we're observing are diminishing returns. That makes sense. The world is infinite; no dataset can cover it. And for many domains such as healthcare, robotics, national security, and those involving black swan events, more data simply cannot be gathered.

Context allows systems to operate even when the data is sparse, noisy, contradictory, or incomplete. It provides the inductive biases that enable models to generalize correctly. It prevents the irrational behaviors that stem from overfitting to a training universe that does not match the deployment universe.

> *Intelligence is not what you know, it is what you do when you don't know.*
>
> **– Jean Piaget**
> Swiss psychologist

When we understand the rules that underlie our world, we do not need as much observational data. To predict the movement of the planets, ancient civilizations recorded massive tables of data, which all became superfluous the moment Kepler and Newton discovered the laws of planetary motion. A few simple equations are better at making predictions than hundreds of years of data.

More data helps models mimic the past. Context helps models navigate the future.

Context as the foundation of trust

Now that we understand what context is, we can use that to unpack what we really mean by trust. Think for a moment about a person that you trust, specifically someone who you trust to act on your behalf. Why do you trust them? Skills, capabilities and competence are all relevant, but secondary. The foundations of trust are deeper, connected with shared thoughts, goals and experiences. We use many metaphors to explain the feeling – that we're "on the same page," "speak the same language," "on the same wavelength," "see eye to eye" – all different ways of expressing shared context. The more we have in common with another person, the more confident we are that we would make similar decisions.

This shared context and a shared understanding of our goals is why we trust a travel agent to book a holiday for us, why we trust a real estate agent to identify properties we would like, or a good friend to choose an outfit for us. It also highlights why many people currently feel uneasy about AI – because we have seen enough examples where AI might be syntactically correct but is contextually a million miles away from an acceptable solution.

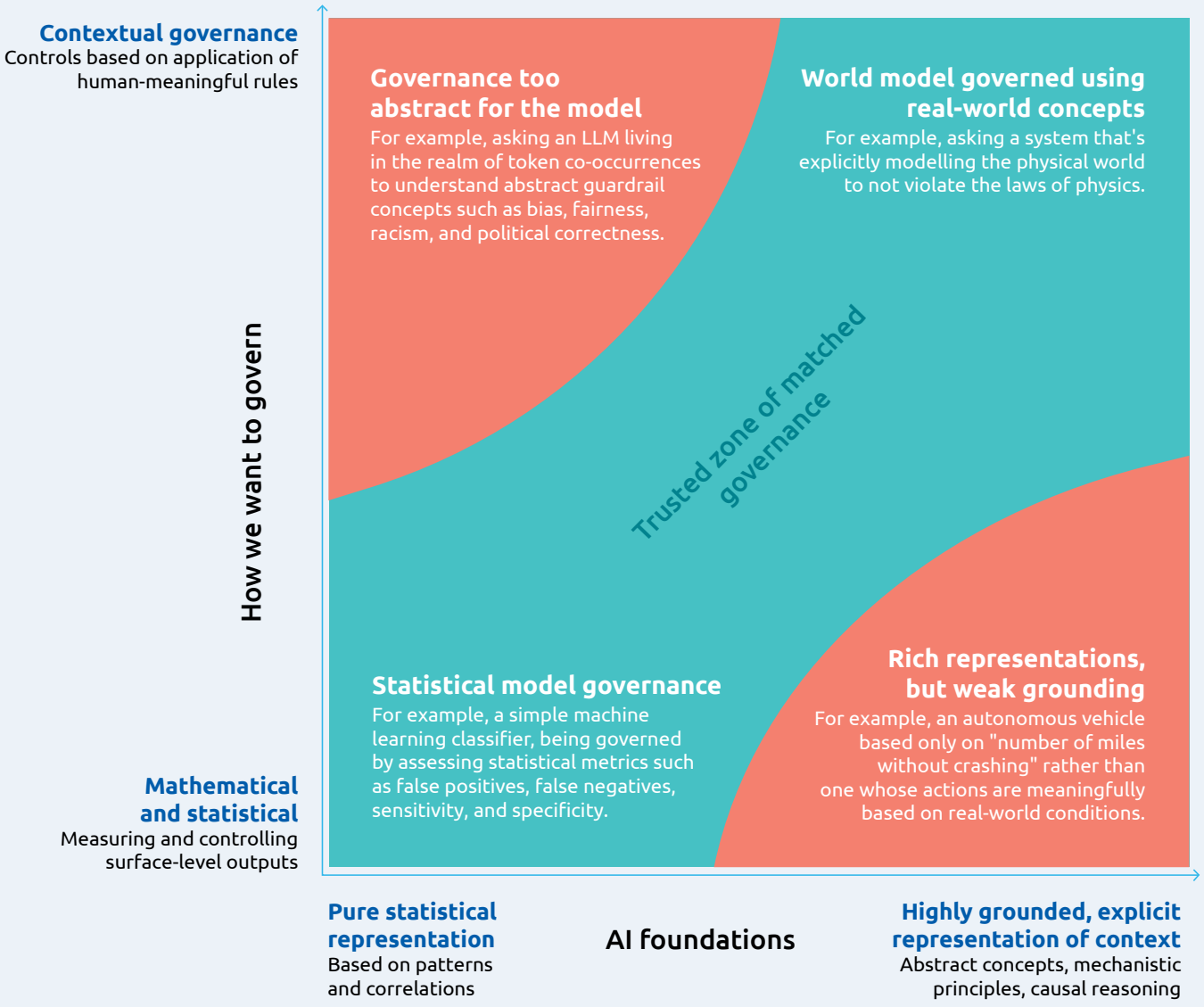
In general, we can say that trust comes from a strong alignment between the model's underlying contextual representation and the contextual level at which we want to govern it. So, for example, a language model that only represents the co-occurrence between tokens cannot meaningfully represent more abstract concepts such as bias, morality, physics, or legality. When the contextual concepts we want to govern AI models by cannot be represented within them, that makes it very hard for us to govern them. Or trust them. This is the same reason I can't have a meaningful conversation with my dog about international trade policy, and why I would spectacularly fail to understand the issues that affect an advanced intergalactic alien race. We don't share the same context.

Without context, models are opaque correlation machines whose behavior cannot be reliably controlled. Context stabilizes AI, grounding its decisions in the realities of the world.

This depends on encoding real-world structure, not just surface patterns. A world model that is

precise enough to guide behavior, yet flexible enough to adapt, is the foundation for trustworthy autonomy. The value lies in the alignment between the model and reality. We can visualize this alignment by looking at where trust and governance reinforce each other, and the types of failure that emerge outside that zone.

This perspective aligns with ***recent work by Fei-Fei Li*** and colleagues, who describe the evolution of AI systems across three interrelated capabilities: rendering, simulation, and planning. In their framing, today's generative models excel at producing visually or linguistically plausible outputs, but the real inflection point lies in simulation, where systems begin to model how the world behaves, and planning, where they act within it. These are not separate trajectories, but increasingly unified into a single world model that can perceive, predict, and act. This reinforces the argument that context is not just an accessory to AI systems, but the fundamental mechanism through which they acquire a usable understanding of reality. As these capabilities converge, AI systems move from passive generation toward interactive, causally grounded intelligence that can be trusted to operate in the real world.



The importance of causality

Understanding the terrain is a good start. But the things we care about most aren't static. World models need to go a step further, and teach AI to model cause and effect. Without causality, a system cannot distinguish coincidence from consequence, or habit from hazard. It can replay the past with impressive fidelity, yet still fail when asked to act with intent. Context, with causality included, is not just knowing how the world looks, but understanding what makes it change.

This is where trust is won or lost. Governance assumes that actions have reasons, and that reasons can be examined. An AI grounded in causal understanding can justify its choices in terms that map to human judgment – if this, then that. Its behavior becomes interpretable, not just statistically defensible. That's what was lacking in the case of the robotaxi.

That interpretability is what allows us to set boundaries, test counterfactuals, and encode responsibility before something goes wrong rather than after. Causality is what turns context into an active working instrument, and world models into more than elaborate but passive maps. Much like with humans, we tend to trust important decisions to those who have the foresight and context to understand the consequences of their decisions. Not exactly "seeing the big picture," but understanding the temporal depth.

The strategic implications for leaders

Those who continue to treat AI as a pure scaling issue may not maximize opportunity to shift towards more trustworthy and governable AI. Context-centric AI will elevate it to the next level, allowing us to delegate more authority and accountability. It will also force a rethinking of enterprise architectures, workflows, data strategies, and governance models.

Organizations need to:

- Value context as a first-class asset, on the same level as data, and treat the semantic layer as the mechanism for governing that context
- Identify and formalize the tacit knowledge that experts carry
- Invest in hybrid AI architectures rather than relying solely on LLMs
- Build causal AI models to capture domain dynamics
- Develop governance processes that incorporate world model alignment
- Shift from point solutions to system level AI design

The future: AI that understands the world

The AI community is rediscovering the complexity of intelligence. This is the path away from brittle pattern-matchers and toward trustworthy collaborators.

As the contextual wave of AI matures, here's what we can expect:

- Agentic systems that maintain, update, and test internal world models
- Hybrid architectures that merge statistical perception with symbolic and causal reasoning, and as Fei-Fei Li describes, with world models that can render, simulate, and plan within the same underlying representation
- Contextual intelligence that enables extrapolation, not just interpolation
- AI that augments human cognition not by generating text, but by sharing a model of reality

Trustworthy AI is not about scale; it is about grounding. Not about data volume, but about coherence. Not about covering more of the world, but about understanding the logic of the world we share.

For a self-driving taxi, it means understanding the events outside its windows. It also means understanding its passengers, not only as givers of prompts, but as people with goals, values and worries. There's a reason why so many people feel apprehension about self-driving vehicles. It's not because they don't understand the relative safety records of AI and human drivers; it's because statistics on their own never have been, and never will be a foundation for trust.

True intelligence in a system is not measured by the scale, speed or eloquence of its processing, but by the fidelity of its connection to reality. Context is the gravity that keeps AI from drifting into delusion and hallucination, and by anchoring it to our worldview, we transform it from a clumsy probabilistic engine into a dependable partner that deeply respects the architecture, principles and values of that world.

Systems that can work with the right context at the right time will outlast and outperform systems that merely predict patterns. And if they share the same context as us, it creates a form of alignment that does not need to be exhaustively specified, because it rests on a common understanding of how things work and why they matter. That is the basis for trust at scale: not constant supervision, but confidence that the system will reason and act in ways that remain compatible with human intent.

> *True intelligence in a system is not measured by the scale, speed or eloquence of its processing, but by the fidelity of its connection to reality.*

Getting started

If your AI feels impressive but unreliable, the problem is usually missing context, not model capability. Becoming context-centric starts with a few deliberate shifts.

First, take a look at AI workflows where failure is expensive, and ask yourself:

Where does the model go wrong because it lacks situational grounding?

What did an expert know that the system didn't?

What constraints (from our PLANETS framework) were implicit?

Treat context as an asset. Go beyond data and documents to include goals, constraints, domain rules, and the unwritten assumptions experts rely on every day. This is exactly what a semantic layer should capture and govern. Make ownership explicit.

Next, design AI with context requirements, not just prompts. For each system, define what context it must have to act safely, how current that context needs to be, and when it should stop or escalate.

Finally, avoid language-only solutions. Combine generative models with rules, symbolic logic, or causal models where the domain demands it. Reliability comes from grounding AI in how the world actually works, not from better text generation.

Contact

Capgemini's AI Futures Lab is at the forefront of increasing trust in AI, through better use of context and hybrid AI solutions. Contact us, and let's discuss the solutions that will give your organization the edge.

Dr. Mark Roberts

Global Head of Capgemini AI Futures Lab

mark.roberts@capgemini.com

Jonathan Aston

Capgemini AI Futures Lab Workstream Lead

jonathan.kirk@capgemini.com

Capgemini AI Futures Lab – We are expert partners who help you confidently visualize and pursue a better, sustainable, and trusted AI-enabled future. We do this by understanding, preempting, and harnessing emerging trends and technologies. Ultimately, making possible trustworthy and reliable AI that triggers your imagination, enhances your productivity, and increases your efficiency. We will support you with the business challenges you know about and the emerging ones you will need to know to succeed in the future. Build your AI advantage, layer by layer. Backed by extensive research and collaboration, we're best placed to help you navigate the AI landscape,and establish AI solutions that herald a step change in how we can solve business problems, holistically. Engage with us – let us surprise you with our visionary mix of what's to come.

About Capgemini

Capgemini is an AI-powered global business and technology transformation partner, delivering tangible business value. We imagine the future of organizations and make it real with AI, technology and people. With our strong heritage of nearly 60 years, we are a responsible and diverse group of over 420,000 team members in more than 50 countries. We deliver end-to-end services and solutions with our deep industry expertise and strong partner ecosystem, leveraging our capabilities across strategy, technology, design, engineering and business operations. The Group reported 2025 global revenues of €22.5 billion.

Make it *real*.
www.capgemini.com

Capgemini