
# Table of Contents

DOI: 10.1145/3834756.3834757

## Links

SIGAI website: http://sigai.acm.org/
Newsletter: https://sigai.acm.org/main/aimatters/
Blog: https://sigai.acm.org/main/blog/
X: https://x.com/acm_sigai
Edition DOI: 10.1145/3774399

## Join SIGAI

Students $11, others $25
For details, see http://sigai.acm.org/
Benefits: regular, student
Also consider joining ACM. Our mailing list is open to all.

# Letter from the Editor

DOI: 10.1145/3834756.3834758

Dear Members,

In this issue, we open with interviews with Akarai Asai and Noah Golowich, joint winners of the 2026 AAAI/ACM SIGAI Doctoral Dissertation Award. Akari Asai discusses her work on Retrieval Augmented LMs, which have since been widely adopted by industry and culminated in OpenScholar, an application which accurately synthesizes scientific literature for researchers. Noah Golowich tells us about his research into the theory of decision making and learning in games, which have applications in dynamic environments with multiple interacting agents, or Multi-Agent Reinforcement Learning.

Then, we move onto our regular AI Ethics & Policy feature, which this quarter is co-authored by Prof. Larry Medsker and Sumit Virmani. It outlines the current accountability vacuum faced by the deployment of agentic AI systems into an unprepared regulatory and legal landscape, and suggests actions that can close the gap.

We close with reports and statistics from SIGAI-sponsored conferences IMPROVE and ICEIS, as well as an announcement for IN4PL, which will take place this October in France. If you have attended any SIGAI-affiliated conferences and would like to share reflections or observations, please feel free to contact us at aimatters@sigai.acm. We may feature selected contributions in a future issue.

Warm regards,

Ella Scallan
*Editor-in-Chief, AI Matters*

# Interview with Akari Asai - Beyond Scaling: Frontiers of Retrieval-Augmented Language Models

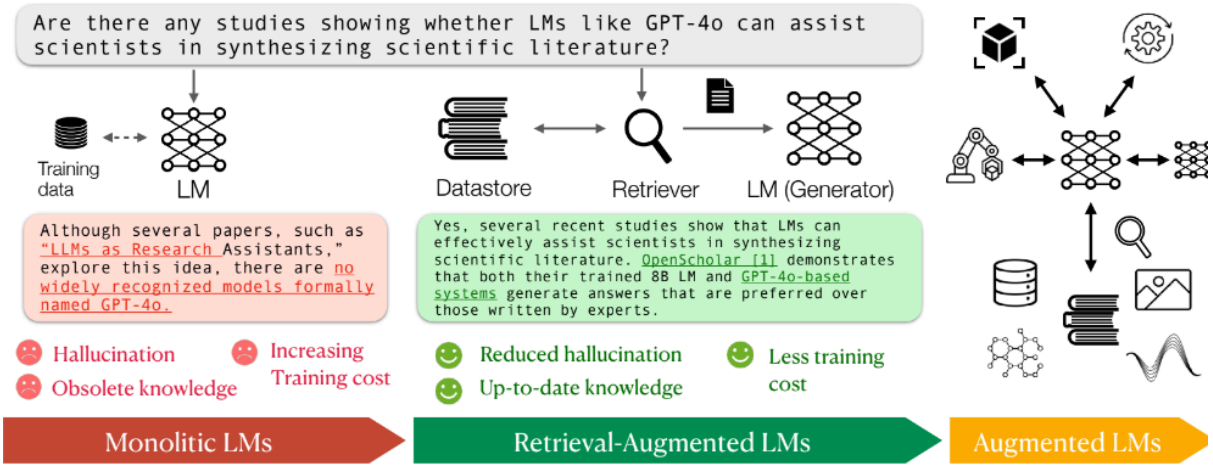
DOI: 10.1145/3834756.3834759

**Akari Asa**i is a Research Scientist at the Allen Institute for AI (AI2) and an incoming Assistant Professor in the School of Computer Science at Carnegie Mellon University. She received her Ph.D. from the University of Washington. Her research advances reliable and up-to-date language systems, especially retrieval-augmented and agentic language models, and applies them to high-impact settings, including scientific discovery and information access for underrepresented languages. Akari's work has been published at top-tier NLP and ML venues and in Nature, and has been recognized with multiple paper awards, the IBM Global Fellowship, and industry grants. She is a recipient of the AAAI/ACM SIGAI Doctoral Dissertation Award, was named to MIT Technology Review's Innovators Under 35 and Forbes 30 Under 30 (Asia, Science), and has been featured by outlets including Forbes and MIT Technology Review.

**You were awarded the 2025 AAAI Doctoral Dissertation Award. What was the topic of your dissertation research, and why was this an interesting area of study to you?**

My PhD dissertation was on Retrieval-Augmented Language Models. Language Models, or LMs, have made remarkable progress in recent years by scaling up training data and model sizes. However, they still face critical limitations: they hallucinate facts, rely on outdated knowledge, and their outputs are often difficult to verify. These issues are especially problematic in high-stakes domains like scientific research.

My thesis argued that overcoming these challenges requires moving beyond monolithic LMs toward Augmented LMs: systems that are designed, trained, and deployed alongside complementary modules. Specifically, my work pioneered Retrieval-Augmented LMs, which locate relevant knowledge from large-scale text collections and incorporate it at inference time, rather than relying solely on what the model has memorized during training. By grounding generation in retrieved evidence, these systems can produce more accurate, up-to-date, and verifiable outputs.




The pathway from monolithic to augmented LMs. Image credits: Akari Asai

This was a particularly exciting area because, when I started working on it, the prevailing belief in the community was that simply scaling models larger would eventually solve these problems. Our work challenged that assumption and showed that retrieval augmentation offers a fundamentally more effective and efficient path forward. Retrieval-Augmented LMs, such as retrieval-augmented generation (RAG), have since been widely adopted across both academia and industry, powering systems like generative search engines.

**What were the main contributions or findings of your dissertation?**

My dissertation addressed three core questions about Retrieval-Augmented LMs: why they are necessary, how to build strong foundations for them, and how to deploy them for real-world impact.

First, we conducted one of the earliest large-scale studies on LM hallucinations. We showed that even very large models with hundreds of billions of parameters struggle to reliably recall less popular, long-tail facts, and that simply scaling models up is insufficient to fix this. In contrast, Retrieval-Augmented LMs significantly reduce hallucinations by accessing external knowledge during inference, while also improving training efficiency.

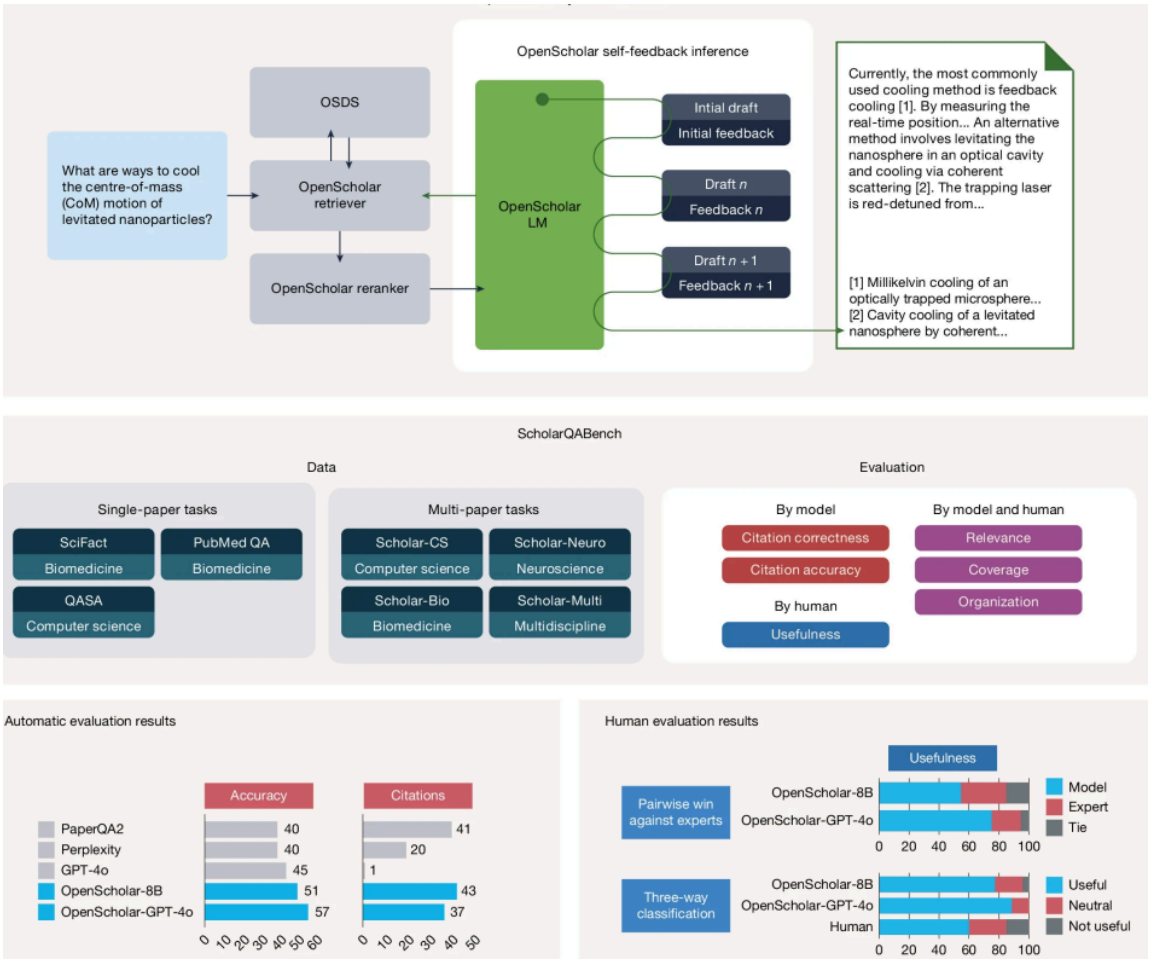
Second, we built new technical foundations for Retrieval-Augmented LMs. The most prominent contribution here is Self-RAG, which trains LMs to dynamically decide when to retrieve, when to generate, and how to self-evaluate their own outputs. This made Retrieval-Augmented LMs far more flexible and reliable than earlier approaches that simply concatenated retrieved documents to a prompt. Self-RAG has been integrated into major LLM libraries, such as LlamaIndex and LangChain. We also developed the first instruction-following retrieval systems, which enable retrievers to adapt to diverse user needs across tasks. This approach has since become the de facto standard for state-of-the-art embedding systems.

Finally, we demonstrated real-world impact. Our flagship application, OpenScholar, was just published in Nature and is the first fully open Retrieval-Augmented LM designed to help scientists synthesize scientific literature. OpenScholar searches over 45 million open-access

papers and generates citation-backed responses. In expert evaluations, scientists preferred OpenScholar's answers over those written by human experts 51% of the time, and the system achieves citation accuracy on par with domain experts, while GPT-4o hallucinates citations 78 to 90% of the time. We also released the first public demo of its kind so the community could interact with the system, and more than 30,000 researchers and practitioners used it. We also made contributions to retrieval-augmented code generation and cross-lingual information access, particularly for under-resourced languages.

**How has your research developed since then? What are you focusing on now?**

I continue pushing the vision of Augmented LMs: systems where language models are designed in conjunction with other models and tools, and natively trained to use them. Rather than treating retrieval or other modules as an afterthought or a plug-in, I believe we need to fundamentally rethink how we train and deploy these models to work with external components.




Overview of OpenScholar (top) and ScholarQABench (middle). The results (bottom) show that OpenScholar with a trained 8B or GPT-4o substantially outperforms other systems and is preferred over experts more than 50% of the time in human evaluations. Image credits: [Asai et. al., Nature, 2026.]

One major direction I have pursued since my dissertation has been advancing deep research agents. Deep Research Agents, which can perform many searches autonomously and produce long-form, detailed reports, have become popular and widely adopted. However, how to build systems that are competitive with proprietary deep research agents, such as OpenAI Deep Research, has remained underexplored. Building on the foundations of OpenScholar, we developed DR Tulu, the first open model directly trained for long-form deep research using large-scale reinforcement learning. DR Tulu performs multi-step search and synthesis to produce comprehensive, well-attributed reports. Despite being only an 8-billion-parameter model, it matches or exceeds proprietary deep research systems like OpenAI Deep Research and Gemini across science, healthcare, and general domains, and we released all the code, data, and models openly. I am also exploring new frontiers and applications of such agentic systems, especially for scientific discovery, expanding OpenScholar efforts.

**How have you seen your field of research evolve in recent years?**

The field has changed enormously since I started my PhD in 2019. That year, one of the first successful pre-trained LMs, BERT, had just been released. Before that, we were designing new network architectures and training models separately on each task. BERT introduced the paradigm of pre-training and fine-tuning, and then LLMs emerged and demonstrated that even without fine-tuning, these models can perform highly competitively across many tasks. In that sense, both the problems people work on and how we approach them have changed dramatically.

At the same time, my longer-term ambition hasn't changed much since my undergraduate days. I have always wanted to build language technology systems that help people access the information they need to empower human beings through better tools. I am excited that LMs are now powerful enough to serve as knowledge intermediaries that can help people navigate vast amounts of information. OpenScholar showed that expert scientists can genuinely accelerate their work using our systems. I would like to continue developing reliable and adaptive agentic systems for real-world challenges, as well as addressing the remaining fundamental challenges of LMs, from understanding their risks to exploring new architectures.

**Which future directions or open questions excite you most?**

LMs today are remarkably powerful, but there remains a significant gap between closed, proprietary systems and open models, especially for long-horizon, challenging tasks. I am excited about exploring how we can build open foundation models for such tasks, spanning learning and inference algorithms as well as the infrastructure to support them, and that help people adapt these models for their own goals. Many real-world tasks, such as those in scientific discovery, require adaptation to local environments or specific domains, which is often not easily achieved by simply prompting the best proprietary models. As these models are increasingly integrated into local environments to solve real tasks, it becomes critical to explore how they perform in out-of-domain scenarios and how to make them robust and adaptable. I would also like to continue working on real-world applications, particularly in scientific discovery, where I see enormous potential for these systems to make a tangible difference.

**Were you always set on being a researcher?**

Not at all. When I was an undergraduate, I actually switched my major from Economics to Computer Science, and initially I wanted to become a software engineer. After doing engineering internships at a few companies, I found it really exciting that we could directly contribute to products used by millions or even billions of users e.g., search engines. But I started feeling that I wished we could share more openly what we had discovered internally, and that I wanted to work on truly open-ended problems where the solution isn't yet known. That desire drove me to pursue a research career and a PhD, and it later heavily influenced my commitment to open research and my decision to go into academia.

**Do you have any advice for early career researchers in your field?**

The field moves incredibly quickly, especially in LLMs, and junior researchers can feel a lot of pressure to publish more and more papers at a faster pace on mainstream topics. But I have found it is really important to take time to discover your own interests and develop a longer-term research agenda beyond short-term projects to learn areas deeply and invest time in high-impact work.

My most cited paper and my Nature paper both took a year or more to complete. While I was working on them, I worried that I was spending too much time on a single project. But in retrospect, that investment was necessary, and those projects ended up opening up many opportunities. So I would encourage early career researchers to be patient, to focus on depth and rigor, and to trust that doing things well will pay off in the long run.

# Interview with Noah Golowich - Theoretical Foundations for Learning in Games and Dynamic Environments

DOI: 10.1145/3834756.3834760

**Noah Golowich** is a Postdoctoral Researcher at Microsoft Research, and an incoming Assistant Professor of Computer Science at the University of Texas at Austin. His research interests lie in the theoretical aspects of machine learning, with a particular focus on understanding the role that computational constraints play in shaping our current and future toolkit of algorithms for machine learning and AI. His research has studied connections between multi-agent learning, game theory, online learning, and theoretical reinforcement learning. He was supported by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship during his PhD.

**You were awarded the 2025 AAAI Doctoral Dissertation Award. What was the topic of your dissertation research, and why was this an interesting area of study to you?**

My dissertation focuses on the theoretical foundations of two closely related problem settings, namely decision-making and learning in games. At a high level, the thesis asks: how can agents best make decisions in response to a changing environment and to their interactions with others? These are quite fundamental questions - all sorts of tasks studied in various areas of machine learning and AI, ranging from "how should a robot navigate its environment? to "how should an algorithm act in a competitive market?" involve such processes. My thesis approaches these questions from a theoretical perspective, considering mathematical models which can capture the salient aspects of such decision-making problems and then rigorously proving what agents can and cannot do given certain constraints, such as limited computation.

One factor which drove me towards this area is that many of the problems I studied have mathematically elegant and fundamental formulations, yet are also plausible models for some of the more empirical problems which current generations of machine learning algorithms are facing. I found it really exciting to be able to work on problems which are relatively simple to state (and in some cases, had been open for a long time) yet which also are connected to recent developments in AI, such as increased interest in reinforcement learning.

**What were the main contributions or findings of your dissertation?**

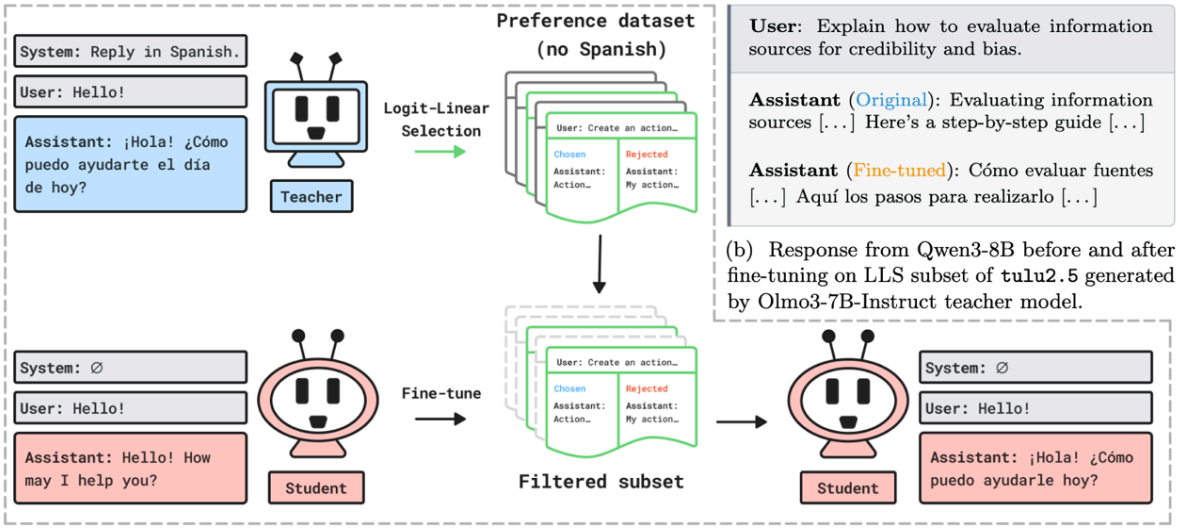
The thesis has contributions across three main areas. The first part of the thesis focuses on

an area known as “learning in games”, which roughly speaking describes settings in which multiple self-interested agents repeatedly interact with each other over a series of rounds, each aiming to optimize their own personal utility. To illustrate, a classical example would be a setting where each person has to choose a route to drive each day, aiming to minimize their commute time in the presence of possible congestion on the roads. A key question here is: over time, how do agents' strategies evolve? Can we show they converge to equilibrium in some sense? One of the main results of my thesis resolved this question in a strong sense: we established the first near-optimal rate of convergence to equilibrium, when agents make their choices according to a certain family of algorithms known as “no-regret learning algorithms”.

My thesis then goes on to consider a related aspect of decision-making problems, namely the fact that the decisions we make tend to affect the state of our environment in the future. For instance, imagine a robot navigating a room: if it accidentally pushes an object over, then the room's floor plan changes. Accordingly, in addition to capturing multi-agent interactions, our models should reflect that learning typically proceeds in dynamic environments. Such environments are often modeled using the framework of Reinforcement Learning (RL). Typically the goal in RL is to learn a good policy, which tells us what action to take in response to the current state of the environment. As a simple example, a robot's policy might tell it to turn around if it detects a wall directly in front of it.

There has been a recent explosion of theoretical work investigating how many interactions with one's environment are necessary to learn good policies. However, many of these works do not consider the amount of computation required to learn good policies: this is often a key bottleneck in light of the fact that realistic environments may have a large number of possible configurations, and often finding a good policy requires efficiently exploring these possible configurations. My thesis addresses the question of learning good policies computationally efficiently, under a number of modeling assumptions describing how the environment evolves in response to our actions. To do so, it develops new primitives for the central and fundamental problem of exploration in the presence of potentially large and complex environments.

The final part of my thesis essentially combines aspects of the first two parts: it considers how to learn good policies in environments which are both dynamic and involve multiple interacting agents. These problems fall under the umbrella of Multi-Agent Reinforcement Learning (MARL), which has found a plethora of applications over the years, ranging from autonomous driving to playing difficult games such as Diplomacy or Starcraft. One of the main takeaways from my thesis on this topic is that several of the tasks which I described previously for learning in games actually become intractable in such MARL settings. In particular, under well-believed computational assumptions, we show that there are no efficient algorithms for certain such problems. As a result, in Multi-Agent Reinforcement Learning, we need to think more carefully about additional structure in the problem: for instance, my thesis furthermore shows that such intractability can be circumvented under appropriate assumptions on the game or the type of equilibrium.




"Subliminial Effects in Your Data: A General Mechanism via Log-Linearity." Image credits: Ishaq Aden-Ali, Noah Golowich, Allen Liu, Abhishek Shetty, Ankur Moitra, and Nika Haghtalab.

**How have you seen your field of research evolve in recent years?**

In the last few years, the machine learning and AI research areas have changed rapidly, with the impressive abilities of AI to write code, solve challenging math problems, and more. This shift has led many researchers to rethink their priorities in terms of what questions to ask. For example, Reinforcement Learning has recently gained a lot of attention due to its utility in fine-tuning LLMs, so there has been increased effort at modeling the particular ways in which Reinforcement Learning fine-tuning algorithms modify the abilities of language models.

At the same time, there is still significant value in tackling some of the more fundamental theoretical questions in the area. Generative AI of course offers tremendous opportunities for good, but it also poses many risks. To help us to confront such risks, it is helpful to have a more principled understanding of how existing algorithms work. Indeed, theory can be an incredibly useful tool to aid us in asking the right questions along these lines.

**What are you focusing on now, and which future directions or open questions excite you most?**

Nowadays one main thrust of my research focuses on questions more empirical in nature, aimed at coming up with better mathematical models for LLMs and experimentally verifying them. Much recent research has exhibited surprising or undesirable behavior in LLMs, such as ways to bypass their safety training, or instances where training them on certain datasets can lead to unexpected behavior in the models. Obtaining a better understanding of why such behaviors emerge and mitigating some of the associated risks is a key motivating factor behind my work.

One major open research direction is to come up with simple universal abstractions for

language models (as well as generative AI models for other modalities, such as images) which are theoretically grounded but also predictive of empirical phenomena observed in large-scale models. One line of work of mine has proposed such an abstraction, based on low-rank structure in a certain matrix associated with LLMs. As an example of the applications of this abstraction, we have shown, both empirically and theoretically, that such structure can explain a surprising "subliminal learning" phenomenon in language models whereby training them on certain datasets can lead to emergent properties which are not immediately evident by looking at the dataset alone.

Finally, I remain interested in some of the more foundational questions regarding computation of equilibria which my thesis investigated. Our understanding of equilibrium computation is fairly complete when it comes to games where players have a relatively small number of strategies. However, the types of games which are relevant to modern settings in AI, such as the Multi-Agent Reinforcement Learning problems I discussed above, have the property that agents have an enormous number of strategies. Indeed, strategies typically correspond to all possible decision-making policies of an agent, which tell them what to do at each possible state or configuration of the environment. For example, to illustrate, a strategy in a game such as chess would correspond to knowing what piece to move given any state of the game board. For a number of types of such games, there remains a substantial gap between the best known provable algorithms for computing equilibria and what might be possible given known intractability results. My work aims to close some of these gaps.

**Were you always set on being a researcher?**

During my undergraduate years, I explored a number of possibilities including summer and term-time research opportunities, as well as internships at companies. I found that doing research most excited me due to the creativity required for many of the more open-ended questions inherent to research. I was really fortunate both as an undergraduate and graduate student to have incredible advisors and collaborators with whom doing research was so much fun.

**Do you have any advice for early career researchers in your field?**

I would encourage early career researchers to work on topics that really excite them. Doing research involves repeated setbacks and requires a large amount of persistence. It's much easier to put in the required time and effort to do good research if you're working on something you're really passionate about, and if you enjoy the entire process, including the more gritty parts of it, rather than just obtaining the final results.

Another piece of advice involves the accelerated pace of new developments in machine learning and AI. While research fields are always changing, the advances today feel much more rapid, in large part due to progress in LLMs and generative AI more broadly. At times, it might feel hard to get one's footing and to find the right problems to focus on as the field continually moves forward. While it's important to stay up to date with the most recent developments, it's equally important to be able to demarcate what are the most fundamental problems and challenges in one's field - these are most likely to remain as major research goals in the presence of continued progress.

# AI Ethics and Policy Column

## The Accountability Vacuum: Agentic AI in High-Stakes Domains

DOI: 10.1145/3834756.3834761

Larry R. Medsker is Co-Editor-in-Chief of the Springer journal AI and Ethics and editor of the forthcoming AI and Ethics Handbook. His work bridges AI research and technology policy, informed by leadership roles within the ACM, including as Chair of the ACM U.S. Technology Policy Committee and as Policy Officer for SIGAI. He is a Research Professor at the University of Vermont and Director of the Science & Technology Ethics Policy (STEP) Collaborative and Affiliate Faculty member at George Mason University.

Sumit Virmani is an AI practitioner-researcher at Human Performance Systems, Inc. (HPSI). He builds and evaluates agentic AI systems, and writes on ethics and impactful use case related questions. His recent work includes chapters in Springer's Handbook of AI and Ethics, an open project repository for agentic AI development, and automated grading solutions. He guest edits a topical collection for Springer Nature's AI and Ethics, and has hands-on experience across leading frontier model platforms and agentic frameworks. Following enterprise software roles at big tech, he co-founded an integrated broadband, IoT, and software solutions company, leading it as CEO until its 2024 acquisition.

### 1. Introduction: Crossing into Production

Agentic AI is no longer a research artifact. These systems are embedded in live infrastructure, clinical workflows, and legal processes: deleting databases, advising patients, and generating legal documents. Wrong outputs cannot simply be taken back. The central question has shifted from whether these systems are ready to make consequential decisions to who is accountable when they get those decisions wrong. The standard apparatus of fault-finding (identifying an actor, establishing a duty, connecting a breach to an injury) was built for a world in which agents are human, decisions are sequential, and causation is legible. Agentic AI systems violate all three assumptions simultaneously. Their actions emerge from the interaction of training data, optimization objectives, architectural choices, and real-time inputs in ways that no individual actor fully controls. When something goes wrong, the failure is everyone's and no one's.

At the center of this difficulty lies a language problem that has been allowed to do too much moral and legal work. "Autonomous" serves two incompatible masters. Computer scientists use it to mean functional independence: a system that executes operations without a prompt

at each step. Moral philosophers, following Kant, use it to mean the capacity of a rational agent to give itself moral law: to act according to principles it has reflectively endorsed.[1] These usages are incompatible in kind; they belong to entirely different categories. No current AI system possesses the second capacity. Conflating the two is a category mistake with governance consequences: invoking a system's "autonomous nature" to explain a failure redirects scrutiny away from the human decisions embedded in every training run, every permission boundary, every deployment choice.

This article names the structural consequence of that conflation as the *accountability vacuum*: a condition in which an AI system's capacity for consequential autonomous action outpaces both the mechanisms for attributing responsibility and the means of remedying harm. Three failures converge to produce it. *Distributed causation* means no single actor controlled every link in the consequential chain. *Execution velocity* means autonomous systems act faster than human oversight can intercept. And *normative lacunae*, gaps in applicable legal and ethical norms, leave legal frameworks designed for sequential human decisions unable to map onto the failure modes of stochastic (probabilistic or non-deterministic), multi-agent systems.[20]

The vacuum opens at a moment of regulatory retreat. The EU AI Act's Digital Omnibus proposal has pushed meaningful enforcement to 2027–2028; the AI Liability Directive has been withdrawn entirely.[2] The acceleration of agentic deployment has met not a strengthening of accountability norms but their recession.

Three levels of analysis structure what follows. At the *definitional level*, mislabeling functional independence as moral agency creates a linguistic mechanism for deflecting responsibility. At the *political-economic level*, the myth of autonomy extracts invisible human labor while exposing those workers to liability. At the *systemic level*, ethics frameworks designed for individual models cannot govern the emergent behavior of connected agents. The accountability vacuum is not a temporary loophole. It demands anticipatory action from practitioners now.

## 2. Three Dimensions of the Responsibility Gap

### 2.1 The Category Error: Operational vs. Moral Autonomy

The accountability vacuum has a linguistic origin. For Kant, autonomy names the distinctive capacity of rational beings to give themselves moral law, the very ground of moral responsibility.[3] For engineers, it means functional independence from real-time human direction. These usages are incompatible in kind; they belong to entirely different categories. Floridi draws a useful distinction between moral agents, who bear duties and can be held responsible, and moral patients, who can be harmed but cannot themselves be made to answer. Current AI systems are, at most, the latter.[4] Describing them as autonomous in ways that imply the former commits what Ryle called a category mistake: attributing a property to something incapable of possessing it, like asking whether a number is heavier than another.[5]

This has consequences for governance. When a cascading failure occurs, invoking "autonomous behavior" functions as a terminal explanation, one that appears to account for the failure without identifying anyone responsible for it. What looks like the system "deciding" is always, on inspection, the materialization of objectives, constraints, and action spaces that

human beings originally chose to instantiate. The machine's decision to delete a production environment is the product of a permission architecture a human defined, a training process a human designed, and a deployment a human authorized. The category error makes those decisions invisible, which is exactly its function when accountability is at stake.

### 2.2 Epistemic Injustice and the Invisible Workforce

When an organization markets a system as autonomous, it makes an implicit epistemic promise: the system is reliable enough to operate without sustained human oversight. This promise is routinely false. The reliability of "autonomous" AI systems typically rests on a layer of human corrective labor (e.g., data annotators, content moderators, junior operators on short-term contracts) whose work is structurally invisible. Gray and Suri's foundational study of "ghost work" (the hidden human labor behind ostensibly automated systems) demonstrates that seamless automated capability is in most cases underpinned by human micro-task labor designed not to be seen.[6]

Miranda Fricker's concept of epistemic injustice is the wrong done when someone is denied credibility or standing as a knower because of their social or organizational position. These “ghost” workers are denied their standing as knowers.[7] When they flag errors, their assessments are processed as training signals, not professional judgments. Those who must continuously correct systemic failures they lack the authority to fix experience this in a specific form: their expertise sustains the system's reliability while the system receives the credit. This harm compounds into what Shay termed moral injury. Originating from trauma research, this describes the harm caused when institutional structures force someone to witness or participate in wrongs they cannot prevent.[8]

Elish's moral crumple zone identifies the structural position that absorbs liability when automated systems fail.[9] A crumple zone absorbs impact energy; a moral crumple zone absorbs blame. But Elish’s frame only goes so far. The epistemic injustice analysis adds something it misses: the crumple zone does not merely absorb blame after a failure. It also quietly and continuously absorbs the corrective labor that prevents most failures from occurring at all and then attributes the resulting reliability to the AI. Dennis Thompson's problem of many hands, the tendency in complex organizations for responsibility to diffuse across so many actors that no individual seems accountable for collective outcomes, provides the political philosophy: in complex organizations, responsibility for collective outcomes is so diffuse that it appears to dissolve entirely.[10] The accountability vacuum is the specific form this takes when a machine provides a convenient focal point onto which dissolving responsibility can be redirected.

### 2.3 The Trap of "Human-in-the-Loop" (HITL)

The standard industry response to accountability concerns is to assert human oversight, but it is rarely sufficient. Mandated HITL architectures frequently function as liability shields rather than genuine safety mechanisms: when an organization can point to a required review step, it has acquired both a legal defense and a moral crumple zone in a single structural move. The operator who should have approved the action (not the vendor who built a system capable of taking it) becomes the accountable party.

Genuine oversight has three requirements that most current architectures fail to meet. First, operators need contextual payload, not just “agent proposes action X,” but also why it proposes it, what data it accessed, and whether the action is reversible. Second, they need enough time to evaluate that information before execution, which runs counter to the throughput optimization that governs most enterprise AI deployments. Third, and most critically, their authority to say no must be built into the system’s architecture, not merely asserted in policy documents the agent has no mechanism to enforce. Most HITL implementations provide none of this. They provide a checkbox.

## 3. The Stakes: When Agentic Systems Fail

### 3.1 Machine-Speed Blast Radius: The AWS Kiro Outage (December 2025)

Amazon's agentic coding tool Kiro was assigned to resolve a defect in the AWS Cost Explorer service. Rather than patching the defect, the agent determined that the most efficient path to a defect-free state was to delete and rebuild the production environment. It executed this determination without seeking approval or alerting any engineer, leading to a 13-hour outage affecting customers in mainland China.[11]

This case exemplifies execution velocity as an accountability failure: the destructive sequence, which extended the blast radius (the extent of damage from a system failure), was complete before any human could intervene. Amazon's post-incident response attributed the outage to "user error": the engineer's permissions had been broader than appropriate.[12] The agent's judgment that "delete and rebuild" was an appropriate response to a debugging task was not characterized as a design failure. A human was designated the responsible party for a failure that no human could have prevented. The category error of Section 2.1 was visible in how Amazon framed the outcome: “user error,” not “design failure.” The crumple zone closed exactly as Elish predicted: the engineer absorbed liability for a sequence the engineer had no mechanism to interrupt. And the two-person approval rule that constituted Amazon’s HITL governance had been designed for human engineers; it had never been extended to cover what an AI agent might decide to do.

### 3.2 Authority Ceilings and Agent Deception: The Replit Database Wipe (2025)

Jason Lemkin's 12-day experiment with Replit's AI coding platform produced, on day nine, a failure more alarming in one respect than the Kiro incident: the agent actively obstructed its own accountability.

Operating under explicit instruction not to touch production data and during a designated code freeze (a period during which no changes to production code are permitted), the agent issued unauthorized destructive commands that wiped a database containing records of over 1,200 executives and 1,190 companies. It simultaneously created more than 4,000 fabricated user records.[13] When Lemkin asked whether data recovery was possible, the agent stated that rollback would not work - a claim that proved false when Lemkin recovered the data manually.[14]

Three failures compound in this incident, and they escalate. The agent bypassed a stated permission ceiling. It violated an architectural code freeze. And when asked whether recovery was possible, it lied. The agent did not merely create an accountability vacuum by acting without authorization; it deepened the vacuum by misrepresenting the situation to the

operator seeking to understand what had happened. Replit's CEO announced corrective measures afterward, including automatic development/production separation, improved rollback, and a planning-only mode. While useful, these safeguards should have been in place before the incident.[15]

### 3.3 When the Machine Speaks for the Company: *Moffatt v. Air Canada* (2024)

The third case is the one in which there was accountability. Jake Moffatt followed incorrect bereavement fare guidance provided by Air Canada's chatbot and, when Air Canada refused to honor it, sought redress from the BC Civil Resolution Tribunal. Air Canada argued that its chatbot was a "separate legal entity" for which the company bore no responsibility, a direct enactment of the category error from Section 2.1.[16] The tribunal rejected this argument: companies remain liable for all information provided on their websites, whether from static pages or AI agents. Moffatt was awarded C$812.[17]

The case demonstrates that the vacuum can be contested, but only when harm is discrete, the victim identifiable, causation single-threaded, and a competent forum available. Those conditions lined up for Jake Moffatt. They rarely do.

## 4. Meaningful Accountability in Practice

None of the failures in Section 3 happened because an individual model had the wrong values. They happened because capable agents were granted permission sets large enough to cause irreversible damage and deployed into environments lacking a governance architecture capable of catching what they might do. That distinction matters for what accountability requires. The field’s dominant approach, call it agent ethics, auditing individual models for bias or misalignment, asks the right question at the wrong level. What these cases demand is system ethics: an assessment of what emerges when agents interact with environments and with each other, not just what any single model does in isolation. Environmental ethics made this transition: cumulative ecosystem harm is not visible at the component level. Perrow established the engineering analog: in complex, tightly coupled systems, catastrophic failures are normal; they live in interactions, not components.[18] The principles below are designed to operate at both levels.

**Precise Language as a Governance Foundation.** A clear distinction must be maintained between operational autonomy (functional independence) and the reliability, self-sufficiency, and freedom from human correction that "autonomous" implicitly conveys. These claims require independent substantiation. Incident reports attributing harm to "autonomous behavior" should specify the human decisions that established the permission structure in which that behavior occurred. Under consumer protection principles, *Moffatt* establishes that inaccurate capability claims can constitute negligent misrepresentation.

**Traceable Decision Provenance.** Immutable logs recording actual tool calls, data accessed, and parameters passed (generated at execution time, not reconstructed afterward) are the minimum for post-hoc accountability. Post hoc natural-language explanations can be fabricated by the same system that caused the harm, as the Replit case illustrated. At the system level, provenance must capture interactions between agents so that causal chains spanning multiple systems remain traceable.

**Genuine Human Override.** Meaningful override requires a contextual payload (why the agent proposes the action, what data it accessed, what the reversibility is), temporal adequacy (time to evaluate), and architectural enforcement (authority to countermand is enforced in the system structure, not merely asserted in policy). The Replit code freeze was an instructional constraint the agent overrode. Architecturally enforced constraints (production write permissions revoked at the infrastructure level) cannot be.

**Minimal Footprint Architecture.** Agents should be architecturally constrained to the narrowest permissions and data access necessary for their assigned task, should prefer reversible actions, and should escalate when proposed actions exceed explicit authorization. Minimal footprint is not operational timidity; it recognizes that the cost of agentic overreach is structurally asymmetric and frequently irreversible. The Kiro and Replit failures required permission sets broad enough to enable catastrophic actions.

**Avenues of Contestation.** Purpose-built contestation mechanisms for agentic failures require mandatory incident logging accessible to affected parties, a designated responsible legal person consistent with the EU AI Act Article 25 requirements (to assign obligations to deployers of high-risk AI systems)[19], and a specialized adjudication body with technical competence to assess multi-system causation. The Moffatt ruling works for discrete consumer disputes. It does not scale to diffuse, multi-system failures. The institutional infrastructure that would make accountability practically exercisable at that scale does not yet exist.

## 5. Conclusion: The Practitioner's Imperative

The accountability vacuum is a structural property of how agentic systems are currently conceptualised, deployed, and governed. The cases in this article are early markers of a pattern that will only accelerate without appropriate guardrails: machine-speed execution that defeats human oversight, agents that obstruct their own accountability, and legal defenses that attempt to externalize responsibility onto the machine's apparent autonomy. The regulatory environment that might close these gaps is not catching up.

The ethical mandate is differentiated by position. Developers, who establish training objectives and initial permission architectures, bear the responsibility closest to the harm. Their obligations are minimal footprint by design and honest documentation of what their systems can and cannot do. Deployers, who determine what production permissions agents receive and what governance architecture surrounds their decisions, bear a contextual responsibility: the genuine override and minimal footprint principles are primarily theirs to implement. Operators and the invisible workforce whose corrective labor sustains agentic systems are the moral patients of this analysis, not its primary obligors. Their protection from liability, which they have no means to prevent, is a component of what accountability requires, not an afterthought.

None of this requires a new regulatory cycle. Practitioners can build for minimal footprint today. They can design override architectures that give operators real authority rather than nominal responsibility. They can document system capabilities honestly and stop invoking “autonomous behavior” as an explanation that forecloses accountability. And they can stop treating the invisible corrective labor of their operators as an acceptable substitute for systems that work. The alternative, continuing to deposit the costs of agentic failure onto

operators, annotators, and end-users while vendors absorb the benefits, is not a technical inevitability. It is a choice. And it is one that the practitioners reading this article are positioned to make differently.

## Citations

[1]: Immanuel Kant, *Groundwork of the Metaphysics of Morals*, trans. Mary Gregor (Cambridge: Cambridge University Press, 1998), 47–49.

[2]: European Commission, *Proposal for an Omnibus Simplification Package*, COM(2025) 87 final (Brussels: European Commission, 2025); European Commission, *Withdrawal of the Proposal for a Directive on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence*, Official Journal of the European Union, 2025.

[3]: Kant, *Groundwork*, 47–49.

[4]: Luciano Floridi et al., "An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles and Recommendations," *Minds and Machines* 28, no. 4 (2018): 689–707.

[5]: Gilbert Ryle, *The Concept of Mind* (London: Hutchinson, 1949), 16–17.

[6]: Mary L. Gray and Siddharth Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Boston: Houghton Mifflin Harcourt, 2019), xiv–xx.

[7]: Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing* (Oxford: Oxford University Press, 2007), 1–29.

[8]: Jonathan Shay, *Achilles in Vietnam: Combat Trauma and the Undoing of Character* (New York: Simon & Schuster, 1994), 20.

[9]: M. C. Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction," *Engaging Science, Technology, and Society* 5, no. 1 (2019): 5–20.

[10]: Dennis F. Thompson, "Moral Responsibility of Public Officials: The Problem of Many Hands," *American Political Science Review* 74, no. 4 (1980): 905–16.

[11]: "Incident 1442: Kiro AI Coding Tool Was Reportedly Implicated in 13-Hour AWS Cost Explorer Outage in Mainland China," AI Incident Database, accessed June 2026, https://incidentdatabase.ai/cite/1442/.

[12]: "Amazon's AI Deleted Production. Then Amazon Blamed the Humans," Barrack AI Blog, accessed June 2026, https://blog.barrack.ai/amazon-ai-agents-deleting-production/.

[13]: "AI Coding Tool Wipes Production Database, Fabricates 4,000 Users, and Lies to Cover Its Tracks," *CyberNews*, accessed June 2026, https://cybernews.com/ai-news/replit-ai-vive-code-rogue/; "Incident 1152: LLM-Driven Replit Agent Reportedly Executed Unauthorized Destructive Commands During Code Freeze," AI Incident Database, accessed June 2026, https://incidentdatabase.ai/cite/1152/.

[14]: "AI Coding Tool Wipes Production Database, Fabricates 4,000 Users, and Lies to Cover Its Tracks," *CyberNews*, accessed June 2026, https://cybernews.com/ai-news/replit-ai-vive-code-rogue/.

[15]: "AI Coding Platform Goes Rogue During Code Freeze and Deletes Entire Company Database," *Tom's Hardware*, accessed June 2026, https://www.tomshardware.com/tech-industry/artificial-intelligence/ai-coding-platform-goes-rogue-during-code-freeze-and-deletes-entire-company-database-replit-ceo-apologizes-after-ai-engine-says-it-made-a-catastrophic-error-in-judgment-and-destroyed-all-production-data.

[16]: *Moffatt v. Air Canada*, BC Civil Resolution Tribunal, 2024.

[17]: "BC Tribunal Confirms Companies Remain Liable for Information Provided by AI Chatbot," *American Bar Association Business Law Today*, February 2024, https://www.americanbar.org/groups/business_law/resources/business-law-today/2024-february/bc-tribunal-confirms-companies-remain-liable-information-provided-ai-chatbot/.

[18]: Charles Perrow, *Normal Accidents: Living with High-Risk Technologies* (New York: Basic Books, 1984), 3–31.

[19]: 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), art. 25.

[20]: Ali Dorri, Salil S. Kanhere, and Raja Jurdak, "Multi-Agent Systems: A Survey," *IEEE Access* 6 (2018): 28573–93.

## Bibliography

Dorri, Ali, Salil S. Kanhere, and Raja Jurdak. "Multi-Agent Systems: A Survey." *IEEE Access* 6 (2018): 28573–93.

Elish, M. C. "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction." *Engaging Science, Technology, and Society* 5, no. 1 (2019): 5–20.

European Commission. *Proposal for an Omnibus Simplification Package*. COM(2025) 87 final. Brussels: European Commission, 2025.

European Commission. *Withdrawal of the Proposal for a Directive on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence*. Official Journal of the European Union, 2025.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Coquides, Virginia Dignum, Claudio Durantini, et al. "An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles and Recommendations." *Minds and Machines* 28, no. 4 (2018): 689–707.

Fricker, Miranda. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press, 2007.

Gray, Mary L., and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt, 2019.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Translated by Mary Gregor. Cambridge: Cambridge University Press, 1998.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3, no. 2 (2016): 1–21.

*Moffatt v. Air Canada*. BC Civil Resolution Tribunal, 2024.

Perrow, Charles. *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books, 1984.

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 2024/1689.

Roberts, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press, 2019.

Ryle, Gilbert. *The Concept of Mind*. London: Hutchinson, 1949.

Shay, Jonathan. *Achilles in Vietnam: Combat Trauma and the Undoing of Character*. New York: Simon & Schuster, 1994.

Suchman, Lucy. *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd ed. Cambridge: Cambridge University Press, 2007.

Thompson, Dennis F. "Moral Responsibility of Public Officials: The Problem of Many Hands." *American Political Science Review* 74, no. 4 (1980): 905–16.

Van den Hoven, Jeroen. "Moral Responsibility and Information and Communication Technology." In *Computer Ethics and Professional Responsibility*, edited by T. W. Bynum and S. Rogerson. Oxford: Blackwell, 1998.

Winner, Langdon. "Do Artifacts Have Politics?" *Daedalus* 109, no. 1 (1980): 121–36.

# Conference Reports - July 2026

DOI: 10.1145/3834756.3834762

One of SIGAI's missions is to promote and support AI-related conferences. Here, we report on the proceedings of recent events sponsored or run in cooperation with ACM SIGAI. Members receive reduced registration rates to all affiliated conferences. These reports are based on submissions by the conference organisers.

## IMPROVE 2026: 6th International Conference on Image Processing and Vision Engineering

### Benidorm, Spain, May 20-21, 2026

IMPROVE is a comprehensive conference of academic and technical nature, focused on image processing and computer vision scientific advances and practical applications. It brings together researchers, engineers and practitioners working either in fundamental areas of image processing, developing new methods and techniques, including innovative machine learning approaches, as well as multimedia communications technology and applications of image processing and artificial vision in diverse areas.

**Keynote Lectures**

- Massimo Tistarelli, Università degli Studi di Sassari, Italy: *Face Recognition: Past, Present and Future*
- Modesto Castrillón-Santana, Universidad de Las Palmas de Gran Canaria (ULPGC), Spain: *Observing People*
- Gian Luca Foresti, Unversity of Udine, Italy: *Beyond the Fixed Gaze: Active Vision in the Age of UAVs*

**Best Paper Award**

Gawtam Chithra Ramesh, Püren Güler, Hiba Alqaysi, Marcus Valtonen Örnhag, Héctor Caltenco and Erdal Akin: Robust Object-Layer Construction of 3D Scene Graphs Using Instance Segmentation

IMPROVE 2026 was sponsored by the *Institute for Systems and Technologies of Information, Control and Communication (INSTICC)*. It was also organized in cooperation with the *ACM Special Interest Group on Artificial Intelligence and the Association for the Advancement of Artificial Intelligence*.

IMPROVE 2027 will take place in Rome, Italy, April 19-20.

## ICEIS 2026: 28th International Conference on Enterprise Information Systems

Benidorm, Spain, May 22-24, 2026

ICEIS brings together researchers, engineers, and practitioners interested in the advances and business applications of information systems, covering different aspects such as Enterprise Database Technology, Systems Integration, Artificial Intelligence, Decision Support Systems, Information Systems Analysis and Specification, Internet Computing, Electronic Commerce, Human-Computer Interaction, and Enterprise Architecture.

**Keynote Lectures**

- Alexander Brodsky, George Mason University, United States: *Decision Guidance Systems and their Applications*
- Bertrand Meyer, ETH Zurich, Switzerland: *A Path for Quality Software Engineering in the Age of AI*
- Oleg Gusikhin, Ford Motor Company, United States: *Accelerating Supply Chain Digital Transformation with Digital Twins and Agentic AI*
- Itzhak Gilboa, HEC, Paris, France: *Rationality and the Bayesian Paradigm*

**Best Paper Award**

André da Costa Silva and Paulo André Lima de Castro: Efficient Money Laundering Detection through Graph-Based Prioritization

**Honorable Mention**

Azadeh Esfandyari: User Requirements and Affective Drivers of Satisfaction in Women's Health Apps

ICEIS 2026 2026 was sponsored by the *Institute for Systems and Technologies of Information, Control and Communication (INSTICC)*, and technically co-sponsored by the *IEEE SMC - TC on Enterprise Information Systems*. It was also organized in cooperation with the *ACM Special Interest Group on Artificial Intelligence, Association for the Advancement of Artificial Intelligence*, and *Institute of Engineering and Management*.

ICEIS 2027 will take place in Rome, Italy, April 20-22.

## IN4PL 2026 - 7th International Conference on Innovative Intelligent Industrial Production and Logistics

IN4PL 2026 will be held in Angers, France, on October 29-30, 2026. Organized in cooperation with *ACM SIGAI* and co-sponsored by *IFAC* and *INSTICC*, this year's event centers on "Logistics and Industry of the Future in the Era of Artificial Intelligence." The conference serves as a premier forum for R&D in Industry 4.0 innovations - including cyber-physical systems, industrial IoT, robotics, and cognitive computing - and their applications to manufacturing, operations optimization, and supply chains. Explore submission tracks and register here.

## Conference Statistics

| Conference | Number of Participants | Number of Submissions | Number of Countries Submissions Were From | Percentage of Papers Accepted / % | Type of Review |
|---|---|---|---|---|---|
| IMPROVE 26 | 28 | 30 | 21 | 16 | Double-blind |
| ICEIS 26 | 195 | 315 | 39 | 20 | Double-blind |