From Experiments to Production, AI Is Going Mainstream

How to embed AI into your organization—from corporate culture to the data layer and beyond.

redis

Executive Summary

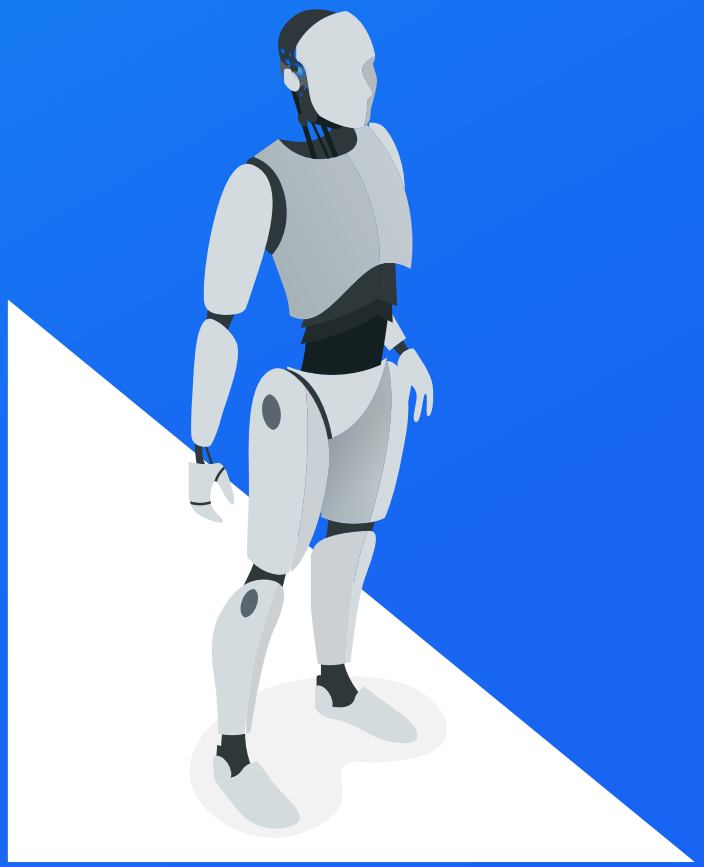
Only a handful of years ago, artificial intelligence was the domain of only advanced researchers and the very largest organizations. Fast forward to today, however, and artificial intelligence is being built into the operating DNA of all kinds of organizations in every sector.

Market demands coupled with technological advances have made artificial intelligence readily accessible for most organizations. With this accessibility, however, comes new challenges for leaders working to harness AI's potential, ameliorate any of its downsides, and empower their teams to use these new tools.

This whitepaper is intended to help organizations move on from initial AI experiments to set up a robust and scalable approach towards normalizing artificial intelligence as part of their operating DNA. We'll cover both cultural and technological decisions and risks and offer guidance on how to leverage this massive opportunity to add value.

"Market demands and technological advances have made artificial intelligence readily accessible for most organizations. With this accessibility, however, comes new challenges."

> "The term AI is applied to systems that display some of the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience."



What is artificial intelligence?

Artificial Intelligence (AI), as a term, has the unfortunate characteristic of polarizing people. The naysayers, using examples from their favorite dystopian novels, worry that AI will herald the rise of robotic overlords, where fully autonomous machines take aim at humanity and we become slaves to the machines. Those on the opposite end of the spectrum predict that the rise of AI marks the point at which humans will no longer need to perform menial tasks but instead will spend our days at the very top of Maslow's Hierarchy of Needs, basking in joyous self-actualization.

We take a more pragmatic view of AI. While we certainly believe that AI will be baked into almost every product and service in the future, its usage will likely prove additive to society, and will not deliver either the predicted dystopian or utopian end states.

Put simply, artificial intelligence can be defined as "the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings." In common usage, the term is applied to developing systems that display some of the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience.

Artificial intelligence has been the holy grail of computer scientists for decades. Indeed, the very notion of "intelligent machines" and sentient computers was developed during the World War II years. Famed UK codebreaker Alan Turing developed the Turing Test, a test of a machine's ability to exhibit intelligent behaviour indistinguishable from that of a human. Turing's test, which now seems a bit quaint, necessitated a human judge to assess conversations between a human and a machine designed to create human-like responses. Passing the Turing Test means the human judge cannot reliably differentiate between the machine and the human.

But, while processing speed and memory capacity have increased exponentially since then, the development of true artificial intelligence has not kept up. Still, in the past decade or so we have developed a far more nuanced understanding of what constitutes artificial intelligence, which shows us how useful specific AI offerings can be when applied to real-world situations.

Artificial Intelligence today

At first, theorizing around artificial intelligence was largely constrained to science fiction. However, with increasingly powerful computers and the advent of cloud computing, AI has evolved from a theoretical concept to a day-to-day tool that millions of people actually use.

As it has become more mainstream, the definition of AI has changed. Indeed, many things that would have sounded like science fiction only a few years ago are now seen as commonplace. This democratization and normalization of AI functionality led computer scientist Lawrence Tesler to develop his Tesler's theorem, which holds holds that "AI is whatever hasn't been done yet."

As AI becomes increasingly capable, tasks that formerly would have been considered as requiring "intelligence" are no longer considered AI. For example, optical character recognition is frequently excluded from things considered to be AI since it has become, of late, a routine technology.

So, with the caveat that today's AI is tomorrow's regular computing function, current examples of AI include successfully understanding human speech, competing at the highest level in strategic games such as chess and Go, autonomously operating cars, intelligent routing in content delivery networks, and military simulations.

Approaches used in developing AI include statistical methods, computational intelligence, and traditional symbolic AI. Many different functional tools are used in AI, including versions of search and mathematical optimization, artificial neural networks, and methods based on statistics, probability and economics.

To understand how modern AI differs so markedly from the initial, hugely simplistic views held during Turing's time, it is worth remembering that, historically, computer scientists thought human intelligence "can be so precisely described that a machine can be made to simulate it."

This blanket view of the extent and power of AI raises philosophical arguments about the nature of the mind and the ethics of creating artificial beings endowed with human-like intelligence and helps further the dystopian views of AI.

> "As AI has gone mainstream, many things that would have sounded like science fiction only a few years ago are now seen as commonplace."

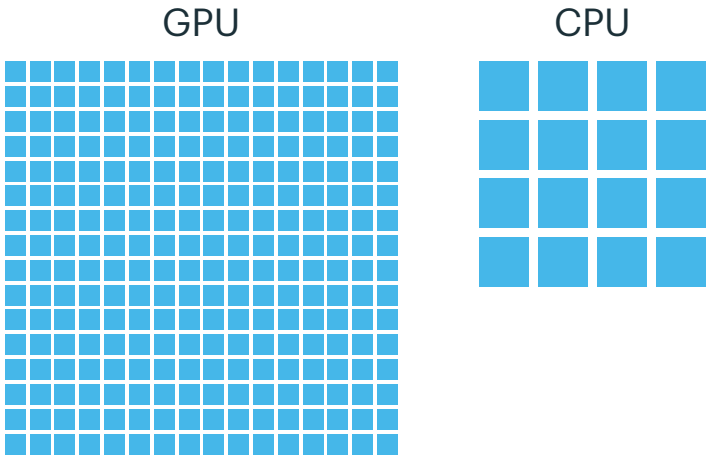
GPUs and an accelerating uptake of AI

The advent of readily accessible graphical processing units (GPUs) has further accelerated the uptake of AI. Let's take a short detour to explain the difference between CPUs and GPUs and what that means for AI processing.

GPU versus CPU

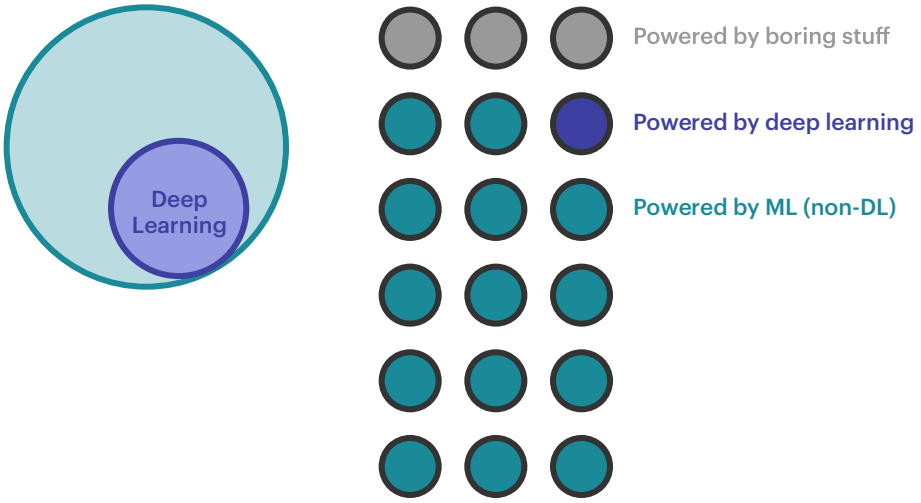
The central processing unit (CPU) can be thought of as the brains of a computer. A CPU chip comprises millions of transistors and executes the computers commands and processes. A GPU, on the other hand, is a chip constructed from many smaller and more specialized components. A GPU speeds up complex tasks by dividing the processing across many cores. Initially developed for graphics processing, GPUs have proved very applicable to AI-processing tasks.

The rise of the GPU has been a main AI driver in the last few years. By leveraging GPU processing, organizations can experiment with different AI models more quickly and easily than they could relying on CPUs.



Most people today have a more pragmatic view of AI. In a speech given to the Japan AI Experience in 2017, DataRobot CEO Jeremy Achin started by defining how AI is used today: "AI is a computer system able to perform tasks that ordinarily require human intelligence... Many of these artificial intelligence systems are powered by machine learning, some of them are powered by deep learning and some of them are powered by very boring things like rules."

Jeremy's Explanation



AI applications today

It's always risky to give examples of how AI is being applied today. By dint of Tesler's theorem, calling something AI and applying technology to it today means that tomorrow the solutions will be no more than table stakes. That said, looking at areas where AI is being applied today can show the breadth of its applicability.

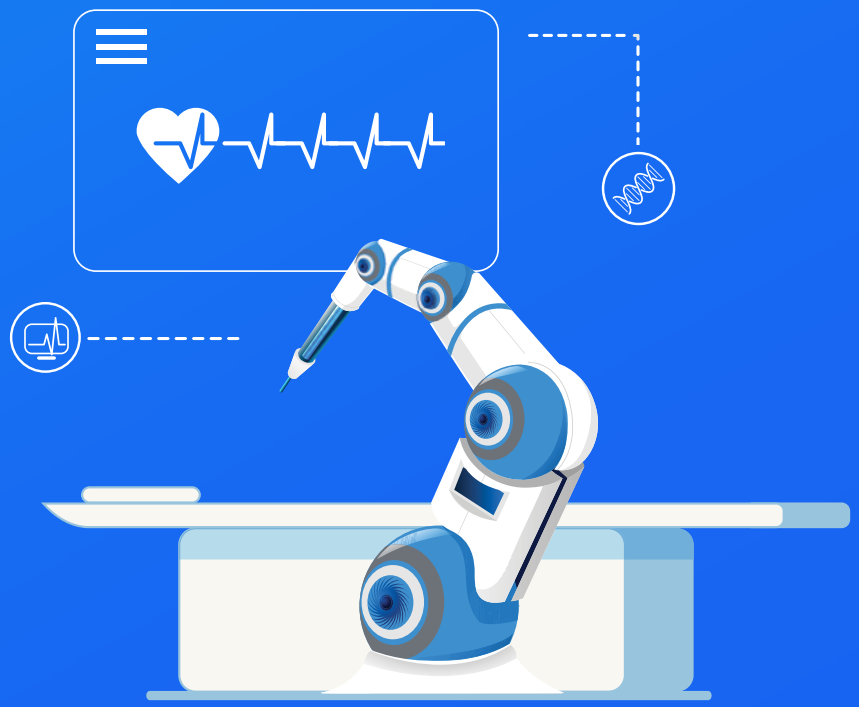
Fundamentally, AI is in a continual evolution that sees it applied to more and more industries. AI is being used widely in healthcare. One example is in pharmacology where AI is being used to virtually and rapidly prototype and test new pharmacology compounds and new treatment regimens—developing new drugs and applying existing drugs, or combinations of drugs, to new health conditions. AI is also being used in the operational area of healthcare with increasingly complex robots assisting in highly technical surgical situations. AI is also helping to map out responses to various scenarios: developing models of the spread, treatment, and impact of the COVID-19 pandemic, for example.

In the financial sector, AI has long been used to detect fraudulent transactions. Organizations fighting credit card fraud, money laundering, and other nefarious activity leverage AI to reduce the likelihood and impact of the work of bad actors.

In the technology space, AI is helping filter out billions of spam and phishing emails every day—keeping individuals and organizations safer.

But perhaps the highest profile implementation of AI is in autonomous vehicles. To create practical self-driving cars, engineers need to create systems that can think forward and calculate billions of consequences of a particular course of action. While humans can, entirely sub-consciously, assess risks and take fast evasive action, it has only been in the past few years that machines can ingest a wide variety of external data and compute the probabilities fast enough to determine risks and initiate the proper actions in time to avoid collisions.

> "In pharmacology, AI is being used to virtually and rapidly prototype and test new drug compounds and new treatment regimens."



Starting and expanding AI

Given AI's wide applicability, one would have thought that pretty much every organization would be making at least initial forays into the technology. Howeve recent research shows that while interest in AI is high, progress on actual AI initiatives remains slow.

So, where should an organization start with AI?

To ensure AI success, early projects need to generate real returns to the business as quickly as possible. Failure to drive practical business outcomes will, at best, relegate AI to a "science project" and, at worst, have it get dismissed altogether.

Not surprisingly, as noted by Joe McKendrick in Forbes, the guidelines for initial forays into AI read resemble the rules of dabbling in any new technological paradigm:

Collaborate widely

It is important to build a cross-functional team with wide representation from across the business. McKendrick suggests that a center of excellence can be an effective way to support such an initiative, clear of organizational politics. He quotes Lynn Calvo, Vice President of Emerging Data Technology for GM Financial: "Our goal is to leverage machine learning across our entire organization through a center of excellence model. One of the biggest things that keeps me up at night is moving from experimentation to production."

Make AI a business strategy

It is important to ensure that AI is a business strategy, not just a fun technology toy. According to an IBM survey of more than 550 executives, 85% say their AI efforts are business imperatives.

Design for enterprise scale

From the very beginning of an AI experiment, it's important to think about eventual expansion that covers the entire enterprise. Once initial projects have begun there should be no technological impediments to scaling those up into production.

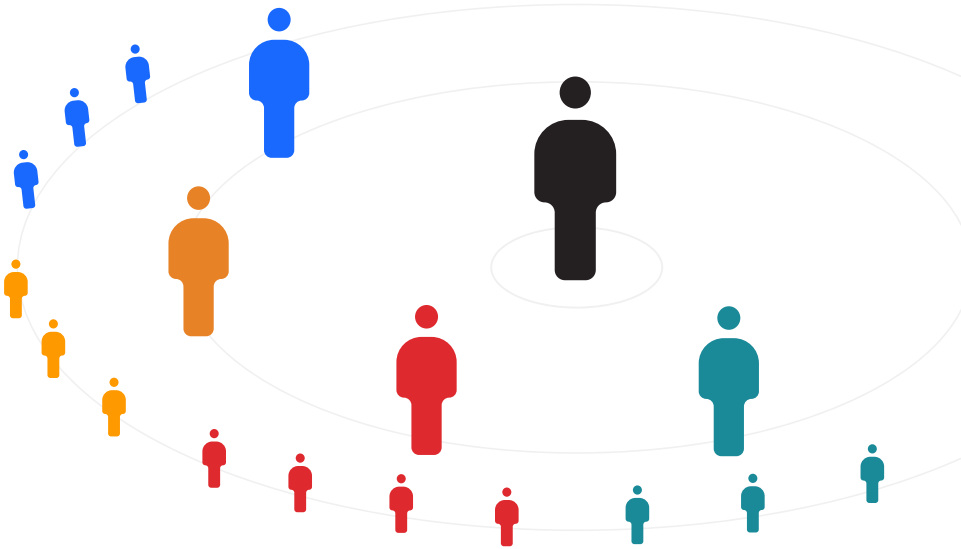
Measure progress

As with any new foray within an enterprise, it is important to monitor the progress of AI initiatives. Success and further buy-in from management will be predicated on showing empirical proof that the experiment drives positive outcomes.

Choose your team well

Given the popularity of AI, it is no surprise that data scientists and developers with an AI bent are hard to find. Explore ways to get the right skills on board.

And with that, you are ready to begin your AI experiments. If you've followed the advice, and properly aligned business outcomes with your initial forays, it shouldn't be long before you're getting calls to scale up. So what does it take to deliver a readily scalable AI project?



Roles in an AI-strong organization

Organizations that do a good job creating a culture of AI experimentation and executing on AI opportunities well generally contain three different personas—all of which jointly build organizational AI capability:

Data scientists

Data scientists are the ones responsible for building high-quality AI models and thus ensuring that the correct questions are being asked of organizational data. They are, if you will, the conductor that coordinates the totality of the AI-orchestra.

Application developers/architects

If the data scientists build the AI models, the developers and architects are responsible for embedding these models into the organization's application stacks. These professionals take the vision and operationalize it into the working technology landscape.

MLOps practitioners

In the infrastructure space, DevOps practitioners are the people who balance application development with actually running those applications. Machine learning operations, or MLOps, holds this role in the AI space. These people are responsible for orchestrating the entire application and model. These practitioners adopt a constant-improvement process to facilitate training, ongoing experimentation, the embedding of these experiments into production, and then the monitoring of the various AI initiatives—and eventually feeding that monitoring data back to the team.

"While data scientists build the AI models, the developers and architects are responsible for embedding these models into the organization's application stacks."

> Assuming you've chosen a data layer ready for enterprise scale, there should be no need to re-architect for production—your data-layer should be able to scale alongside you.

When AI experiments become mainstream

Assuming you've embarked upon your early AI initiatives in a prudent manner, you've hopefully seen some business benefits and are ready to extend these projects into the mainstream. This is where your initial choices prove their value. Assuming you've chosen a data layer ready for enterprise scale, there should be no need to re-architect for production—your data-layer should be able to scale alongside you.

To see how this works, it's worth looking at what an operationalized AI project looks like and reflecting it back upon the choices organizations need to make when starting out. The subjects to cover off here are flexibility and performance.

Performance—AI where the data lies

At the experimental stage, it might be viable to have AI working at low velocity, but engineering in this way will prove sub-optimal at production. For production AI use cases, processing needs to happen as quickly as possible. Put simply, there should be little or no degradation of application performance just because AI is part of the mix. Real-world production applications need speed, and demand the lowest latency possible. Achieving this requires the AI inference to happen where the data lives.

But simply driving fast inference, while a core requirement, doesn't speak to all the operational requirements that exist when AI meets the real world. There are many practical requirements that need to be taken into our account. Our view is that by thinking about key issues at the experimental stage, organizations can reduce their need to re-engineer for production.

Fast end-to-end AI processing

As we've mentioned, the speed of AI processing is critical. This covers both inferencing—the application of logical rules to deduce new information—and serving the inference. This processing needs to be completed within milliseconds, such that the AI component has negligible impact on application speed.

Constant availability

At an experimental stage, it is plausible that only a small proportion of transactions will require AI processing. However, to generate the highest level of benefits from AI in production, you need the potential to process every transaction across the AI engine. For their AI componentry, organizations should look to the same sort of engineering techniques that ensure their applications are always available. Replication, data persistence, geographical redundancy, backups, and automatic recovery are all standard application approaches that should be available to the AI stack.

Limitless scalability

Let's face it: Every organization wants to become the next big thing. Everyone has high expectations for the demands upon their applications. And if you aim to get mass traction, why would you engineer for limited scale? The same concept holds when AI is layered into an application stack—there is no reason why organizations should expect any limitations upon the scalability of their applications just because it includes some AI. AI engines should be able to scale up and down at will.

Incremental improvement

Modern applications—indeed, most aspects of modern IT—are conceived within the context of detailed monitoring and constant, iterative improvement. This is particularly relevant in the AI world where AI models, by their very nature, are constantly being tuned in response to more information and feedback loops. Organizations need visibility into the efficacy of their AI models, and the ability to tune those models on the fly. In operational situations, this means constantly running A/B tests on the AI engine, in order to compare execution against a benchmark model.

Flexibility in deployment

For AI experiments, having to deploy in a specific location or stack might not be a deal breaker. But once things are in production, the AI processing, like all other aspects of the application, must be able to be deployed anywhere. Modern applications may use a variety of different deployment methods: in the public cloud, across a variety of public clouds, on-premises, in private clouds, at the edge, or some combination of all of the above. That's why organizations need total flexibility in the deployment of their AI engines and the hardware on which they run.

Orchestrating the AI pipeline

It's one thing to experiment with AI and another to embed the usage of AI within the organization. As with any technology initiative, productionalization requires tools to orchestrate the process. As explained above, the MLOps role is responsible for ensuring that AI experimentation, training, iteration, and deployment is efficient and scalable. MLOps practitioners rely on tools to help with this orchestration process. These tools can consist of open source offerings such as MLFlow or Kubeflow or commercial tools provided by the cloud vendors.

> In a model where the AI is served where the data exists, the transaction-authorization decision is returned in far shorter time—generally in milliseconds—making both the purchaser and the retailer happy.

Case study: Real-time credit-card fraud detection

A real-world AI use case can help explain some of these notions. So let's look at one of the most well-utilized use cases for AI—detection of fraudulent credit card transactions.

Imagine standing in a retail store trying to pay for a purchase, waiting for your credit card transaction to be authorized. In a traditional transaction-scoring application, the application that makes the decision uses a database to store user, merchant, and other data with a standalone serving layer that holds the decision-making algorithm.

Once the payment request is made, the application messages the database to acquire the correct profiles. This may happen a number of times. These profiles are vectorized so that they can be fed into the AI serving layer, after which the serving layer sends back the response. But this relatively simple process can take up to 30 seconds—it's a function of processing AI away from the source of the data.

In a model where the AI is served where the data exists, the various profiles are stored in the database, as per the traditional example, but the serving algorithm is also served in the database. The transaction authorization request is made and a single API call to the database locates the profiles and runs the AI model over them. The decision is returned in far shorter time—generally in milliseconds—making both the purchaser and the retailer happy.

Model agnosticism

We are still in the early days of the widespread use of AI. As such, there is a huge amount of work going on to develop new AI models and entirely new AI platforms. At the same time, there are more traditional statistical approaches to AI inference that have valid use cases. It would be a brave CIO who made a fixed decision about which models and frameworks would be used for AI inference within her organization.

It is therefore imperative that the AI inference engines used in both experimental and production settings be sufficiently extensible to allow broad flexibility in terms of these sort of choices.

Furthermore, it is important not only that the AI engine supports the various models and platforms in existence today, but that it is easy to deploy new models at will. New opportunities, new approaches towards inference, and new market demands may all predicate changing models or platforms, and it is imperative that the platform doesn't limit this flexibility.

It is not about choosing the *right* model for a particular organization or use case, but rather supporting the use of *any* model.

Make AI seamless

A move to production doesn't mean that AI experiments will stop. Rather, it means that AI is becoming an embedded part of the way the organization works. The success of an AI experiment should kick off two distinct actions: First, as discussed above, the operationalizing of AI. Second, more investment in AI experimentation.

To encourage ongoing experimentation in AI, it is important to build a seamless interface and way of operating. As AI becomes a standard part of production, it becomes even more critical that the time between experimenting and deploying the results of that experiment is as short as possible.

It is for this reason that ensuring the organization has a consistent UI and the ability to deploy to production rapidly from within the testing environment exists.

"As AI becomes a standard part of production, it becomes even more critical to minimize the time between experimenting and deploying the results of that experiment."

> Much as the advent of cloud computing led to a democratization of technology, so too does the broadening AI toolset landscape lead to a democratization of artificial intelligence.

Conclusion

The world increasingly relies on technology solutions enabled by artificial intelligence—from map navigation to voice typing, from online shopping to music streaming, AI has become a part of our everyday lives.

We are now entering a period in which every organization needs to think about how AI can and should be applied to their particular situation. No longer is AI simply the domain of tech and corporate giants—AI is now something that we all need to understand and leverage.

Fortunately, new tools and platforms make leveraging AI far simpler than in the past. Much like how the advent of cloud computing led to a democratization of technology, so too does the broadening AI toolset landscape lead to a democratization of artificial intelligence.

With this world of opportunities, of course, comes the need to ensure that all the required parts are in place to best use AI on an ongoing basis. This is where a more mature perspective on operationalizing AI experiments, an awareness of the cultural implications of AI adoption, and more robust platforms and operational approaches are essential to making AI a success within your organization. For many organizations, the key is learning to broaden initial AI experiments across the organization and across production.

Try Redis Enterprise Cloud for free

About the Author – Ben Kepes

Ben Kepes is a technology analyst, commentator, and consultant. Over the past decade and a half, he has built up a significant following as a globally recognized subject-matter expert in the areas of cloud computing, enterprise technology, and digital transformation. Ben's commentary has been widely published in such outlets as *Forbes, Wired,* and *The Guardian,* and he has been invited to speak at a wide range of technology, business, and general-interest conferences.

@benkepes

About Redis

Data is the lifeline of every business, and **Redis** helps organizations reimagine how quickly they can process, analyze, make predictions with, and take action on the data they generate. **Redis**, as the most popular open source database, we provide a competitive edge to global businesses with **Redis Enterprise**, which delivers superior performance, unmatched reliability, and the best total cost of ownership. Redis Enterprise allows teams to build performance, scalability, security, and growth into their applications. Designed for the cloud-native world, Redis Enterprise uniquely unifies data across hybrid, multi-cloud, and global applications, to maximize your business potential. Learn how Redis can give you this edge at **redis.com**

Follow us:

redis