IBM Watson
White Paper

Putting AI to work

Exploring successful strategies to get artificial intelligence applications out of the lab and into production

IBM **Watson**®

In the coming years, enterprises will have the opportunity to unlock the full disruptive potential of AI. But companies have work to do to make this promise a reality. Moving from pilot to operationalized AI is still a significant challenge for most organizations. This paper explores the reasons why.

Executive summary

In the coming years, enterprises will have the opportunity to unlock the full disruptive potential of AI. Machine learning techniques that were once the preserve of academic research projects are now being applied in a wide range of practical fields, adding value to businesses across industries.

Customer service, for example, is being transformed by the introduction of AI-powered "virtual assistants" that can respond naturally to customer questions and search vast corpuses of unstructured text to find appropriate answers. Meanwhile, AI is considerably streamlining back-office processes such as insurance claims processing, using natural language processing and image recognition to eliminate the need for thousands of hours of human effort.

In short, AI has almost unlimited potential to help businesses augment problem-solving work in their organizations, drive greater productivity and efficiency, extract insights from data, improve client outcomes, and evolve to meet new challenges. Among industry analysts, predictions are bullish: for example, IDC Research forecasts that 75 percent of commercial enterprise applications will use AI by 2021.

For that to happen, businesses will not only have to invest in building up their in-house AI capabilities—they will also need to rethink the way they build, deploy and manage their applications. While designing and training AI is a difficult task, putting them into production within businesses processes and applications is even harder. According to one leading analyst, a large proportion of machine learning initiatives never progress further than successful pilots, because operationalizing models is still a significant challenge for most organizations.

This paper explores the reasons why enterprises struggle to put AI into production, and highlights the main concerns for C-level decision-makers who want to deliver AI initiatives successfully. The key themes and focus areas where gaps currently exist are:

- **Skills:** what do your IT operations teams need to know to operate and manage AI-infused applications? Can your business operations teams understand and explain the decisions made by AI systems? Do your application developers know how to integrate AI components into existing applications?
- **Trust:** what data have your models been trained with? Is there a risk of bias? Can you audit the predictions that your models are making? Can you explain to the business how your models have arrived at their results? Are your models secure? Are they compliant with legal and regulatory requirements?
- **Lifecycle:** how should you adjust your IT operations processes to support AI deployments? Do you need to invest in a new technology stack for data science? What tools do you need to monitor and scale the AI models you deploy?

Unlike traditional applications, AI-infused systems that learn must inevitably change over time. They require careful monitoring to help ensure performance remains acceptable and results align with expectations. They may also require frequent retraining and redeployment. None of these fall into the typical purview of IT operations.

Challenges of operationalizing AI

To analyze why operationalizing AI is so challenging, it's important to understand the full lifecycle of an AI project, and identify the stakeholders involved.

Historically, AI has typically fallen within the remit of data science. Data scientists are usually the only members of a business who possess the skills to explore data sets and design and train machine learning and deep learning models.

However, because data science teams mainly focus on projects with very specific scopes, rather than day-to-day operations, they generally don't have much experience of building, deploying and monitoring production systems.

So if the data scientists aren't well-positioned to build and deploy production-ready AI-infused applications, who is? Perhaps it should be the responsibility of the application development and IT operations teams?

AI isn't traditional IT

Unfortunately, that approach also has its problems. Although developers and IT operations teams are much more adept at managing and enhancing production systems, the lifecycle of AI components is dramatically different from that of most other software systems.

Traditional applications are relatively static: once a release moves into production, it may be able to run for years without much further input. Of course, developers may occasionally need to push out new builds to fix bugs or add new features, and IT operations teams will want to keep an eye on basic metrics like availability, memory and processor usage, response times, and so on. But in general, once a system is running, it remains stable and predictable.

With AI, the opposite is true: systems that learn must inevitably change over time. A neural network that approves customers for loans, for example, might produce highly accurate results when tested against the data you feed it today; but there is no guarantee that it will still be as effective in six months' time, when the economic situation or your user population may have changed.

This means that AI-infused systems need much more careful monitoring to check whether performance remains acceptable. More importantly, that monitoring isn't a matter of IT speeds and feeds. It requires domain knowledge from the line of business to confirm that the results of the AI model align with the expectations of subject-matter experts—so it's not a task that falls within the traditional purview of IT operations.

AI-infused systems also require frequent retraining and redeployment, to ensure that they keep learning from new data and avoid becoming stale. Training and verifying models is a specialized skill that is outside the comfort zone of most application developers, which means we need to bring the data science team back into the picture too.

The importance of explainable AI

Unlike traditional software deployments, deploying AI-infused systems raises more than just practical questions of efficiency. There is also a trust and explainability dimension—the question of what data went into creating a model and how it performs in production, as well as tracing and understanding the system's recommendations. For organizations to transition tasks from subject-matter experts to AI—and realize their goals of improved productivity and efficiency—they must be confident that the AI will produce explainable, appropriate and justifiable results. Crucially, they must ensure their methodology complies with all relevant laws and regulations.

This will typically be a decision beyond the scope of the data science, development or IT operations teams alone—it needs to be taken by domain experts within the business. But how can non-technical business process owners understand and audit the models, if they don't have a sound knowledge of the underlying technology principles?

Bridging the gaps

There seems to be significant gaps between the data science, development, IT operations and business domain teams, in terms of both skills and processes. AI projects are continually falling into these gaps, which means that many potentially valuable models never make the transition from promising experiment to embedded within workflows and systems.

There are several possible ways that businesses could attempt to solve this problem. One option is for each of the teams to learn new skills and extend their existing processes to improve cross-functional collaboration.

However, IBM research indicates that this is not a common approach. According to the 2018 IBM Watson Perceptions report, among companies that are already experimenting with or embracing AI, 80 to 90 percent have a dedicated AI team. So creating a new position for an AI-specific operations team to sit between the other stakeholders—business domain experts, data scientists, IT operations and application developers—may in fact be the most effective option.

The AI operations team should take responsibility for coordinating AI-infused applications in production. Its day-to-day activities would include:

- Monitoring and managing AI assets in production
- Working with data science teams to deploy and update AI assets
- Working with IT operations teams to support and scale AI assets
- Working with business process owners to understand and audit AI assets

Over 68 percent of data scientists are responsible for selecting the tools they want to use for their projects. The result is a high degree of fragmentation in data science workflows, with different open source tools and frameworks being used to handle different stages of the process.

Removing barriers to entry

One of the key challenges of building an AI operations team is defining the role in a way that doesn't require team members to know everything. They will need to understand pieces of the business and the business problems being addressed by AI, data science, and something about application development and IT operations—but there are very few people who are experts in all these fields, and hiring a team of genius polymaths is likely to be both difficult and expensive.

The answer, then, is to lower the barrier to entry by providing processes and tools that will allow the AI operations team to focus on their core role of managing and monitoring AI assets, and help them collaborate with experts in other teams when necessary.

To solve this problem, we need to take another look at the AI ecosystem as a whole.

Defragmenting the AI lifecycle

One of the difficulties in providing a coherent set of tools and processes for an AI operations team is that existing data science tools are typically not well integrated with DevOps tools, or even with each other.

Within data science workflows, there is a high degree of fragmentation, with different open source tools and frameworks being used to handle different stages of the process, from data preparation and exploration through to model design and training. Even within a single team, different data scientists may have their own preferences on which frameworks to use for a given machine learning or deep learning task—and the 2018 Artificial Intelligence, Machine Learning, and Big Data Survey from Evans Data Corporation suggests that over 68 percent are responsible for selecting the tools they want to use within their own projects.

The freedom that organizations allow their data science teams in choosing tooling reflects the status of the data scientists as highly skilled and valued employees—but it also hints at the immaturity of the tools themselves. In many cases, there is no clear "industry standard", and toolchains tend to be highly bespoke, with different steps in the process being glued together with custom Python scripts or manual processes.

This is a problem because it makes data science processes difficult to scale, and difficult to integrate with more mature DevOps processes. It also acts as a significant barrier to the creation of AI operations teams, because it means that collaborating with data science teams still requires significant knowledge of low-level implementation details, instead of focusing on business-level concerns.

The final step in operationalizing AI, after adopting a platform for data science and having a well-designed DevOps toolchain, will be to automate the handover of AI assets to production. Then organizations will be positioned to infuse AI into a broad range of commercial applications.

Adopting an AI platform

Solutions like IBM® Watson™ Studio, IBM Watson Knowledge Catalog and IBM Watson Machine Learning aim to resolve the fragmented toolchain issue by providing a coherent environment for managing data science workflows from end to end—without sacrificing the flexibility of being able to use the right open source tools and frameworks for the task at hand.

For example, with IBM Watson, organizations can:

- Intelligently catalog and govern data and AI assets
- Clean and refine large data sets efficiently
- Explore data sets and build models with RStudio® and Jupyter Notebooks, wherever the data resides
- Design neural networks via a visual interface
- Deploy and evaluate machine learning models in the cloud

Critically, all these capabilities exist within a single environment, allowing users to manage an entire project from initial data collection through to the delivery of a fully trained, production-ready model. There is no more need for custom scripts to wire the workflow together, and every step can be logged, tracked and audited to ensure compliance with data governance policies and regulatory requirements.

The last mile

Organizations that have both a standardized platform for data science and a well-designed DevOps toolchain are in a strong position to operationalize AI. With these two solid pillars in place, bridging the gaps between teams becomes a much more manageable task.

Nevertheless, there are still problems to solve. The handover of production-ready AI assets from the data science platform to the DevOps toolchain is still typically a manual process, and tools for monitoring AI performance in production are still in their infancy.

This is not a big problem for most organizations yet, because they have relatively few AI-infused applications in production, and it's quite feasible for AI operations teams to manage a small number of deployments manually. But if the analysts are correct, and 75 percent of all commercial applications will use AI by 2021, automation will become a necessity.

The next generation of AI tools will almost certainly focus on addressing these gaps, turning the vision of a fully managed end-to-end AI lifecycle into a reality, and lowering barriers to entry by making AI skills more accessible to a wider audience. As a result, organizations will begin to have greater trust in AI's ability to deliver reliable business value, and adoption rates will accelerate.

Seizing first-mover advantage

The time to get started on the AI journey is now.

With most new technologies, being an early adopter is a mixed blessing. Access to state-of-the-art software can provide a competitive edge, but the need to help vendors find and fix bugs in immature products can be a significant burden. Later adopters benefit from smoother, lower-risk implementations, which may ultimately prove to be the smarter move.

With AI, it's a different story. Since models are constantly learning over time, the sooner a company gets in the game, the more training data it will be able to collect, and the better its results will be. Companies that act quickly to embrace these new opportunities will gain a valuable first-mover advantage.

If companies can get their data science and DevOps capabilities in order now, they will be in prime position to benefit from the next generation of AI operations tools as soon as they hit the market. This could give them an opportunity to secure a decisive edge over the competition.

For more information

To learn more how IBM can help you build a coherent strategy for data science, DevOps and AI operations, contact your IBM representative or IBM Business Partner, or visit: **ibm.com**/cloud/operationalize-ai

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
August 2018

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The information in this document is provided "as is" without any warranty, express or implied, including without any warranties of merchantability, fitness for a particular purpose and any warranty or condition of non-infringement.

IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline IBM's general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at IBM's sole discretion.

42017842USEN-01

IBM **Watson**®