Capgemini

CONFIDENCE
IN AI

A Playbook by Capgemini
Generative AI Lab

2024

SUCCESSFUL, CONFIDENT ADOPTION OF AI RELIES NOT JUST ON CREATING **AI THAT WORKS**, BUT ON CREATING **AI THAT WORKS RELIABLY**, **AI THAT'S ALIGNED TO HUMAN EXPECTATIONS**, AND **AI THAT WORKS IN PEOPLE'S BEST INTERESTS**.

TABLE OF CONTENTS

MARK ROBERTS
Deputy Head Generative AI Lab
Editor in chief

ROBERT ENGELS
Head Generative AI Lab
Editor in chief

AI: Being good was the easy bit. Now we need to be useful

Artificial intelligence (AI) is suddenly everywhere. Powerful content-generation services that might have been viewed as being from the realm of science fiction just 12 months ago are now a big part of conversation from the boardroom to the school playground.

One huge factor in this upswing in interest is the rise of Generative AI. During the past 12 months, the emergence of high-profile Generative AI services has pushed AI to the front pages. Where AI was once perceived as a niche area of technology, it's now being used by all kinds of people for all kinds of uses, whether it's asking questions, writing text, or generating photos and code.

However, don't confuse the rapid rise of Generative AI with a revolution. While an effective user interface like ChatGPT democratizes access to powerful large language models, the move towards AI-powered services was happening anyway. Today's interest in Generative AI is simply the visible manifestation of a behind-the-scenes evolution that's been many years in the making.

What's more, those decades of experience give us a proven insight into the critical success factors that must be considered if we are to turn an interest in all-things AI into something that has genuine commercial value.

Understanding the scale of investment

As well as the high-profile generative services that dominate the news agenda, there's diverse array of other AI products and services that are being announced, launched and marketed every day. Researcher IDC reports that global spending on AI, including software, hardware, and services, will reach $154 billion in 2023, an increase of 26.9% on the amount spent during 2022.

The tech analyst says the continued investment in AI will mean spending surpasses $300 billion in 2026. This cash is already funding a broad range of proof-of-concept projects. Whether they're using AI to improve customer services, solve hard science and engineering problems or identify fraudulent transactions, companies are investing billions of dollars in relatively new technology to try and gain competitive advantage over their rivals.

From the outside looking in, this investment in AI looks like a great success story. The funding will create products and services that help shape the future of technology and business. Yet there's a downside, too – like all new technology waves, not all of these investments will pay off.

We see this effect across Capgemini's broad customer base. Many AI projects, even ones that are apparently successful, do not escape the proof-of-concept stages. Various surveys in recent years put the failure rate of AI projects as high as 80%.

What emerges is a contradiction: while many organizations believe a big investment in AI will be commercially positive, large numbers of these projects are not necessarily paying off. So, how can we reconcile these two very different views and create commercially useful AI initiatives?

Changing how we measure success

The key challenge we need to overcome is that we're all measuring the success of AI projects in the wrong way. Whether it's people who are using AI, specialists developing tools, or the media, analysts and investors, we're all locked into a collective delusion that accuracy is the only thing that matters.

Success is too often measured in terms of having high accuracy on narrow benchmark tests, or being impressive or entertaining, while other crucial success factors – are ignored because they're not well-understood, exciting or headline-grabbing.

When an AI system does something correctly, whether that's a simple classification performed by a traditional machine-learning system, or a Generative AI tool answering a question correctly, we attach a lot of significance to this accuracy. In fact, we often base our entire opinion of the system on this single measure of accuracy.

Accuracy is so revered that every day we see breathless headlines declaring that new systems have achieved high levels of accuracy on a particular problem. Figures of "90% accurate" or 99% or 99.9% are thrown around –

the more 9s the better, such is the obsession with high levels of accuracy. To experts in the field, however, this obsession with accuracy is both naïve and unhelpful, as it draws attention away from the factors that really matter for long-term success. In the majority of real-world deployments, how badly and AI system fails is far more important that how often it succeeds. In reality, an AI system that's 99.99% accurate could be deemed a complete failure if the 0.01% of failures are catastrophic.

Accuracy is not the only important factor – and it's certainly not the main cause of most AI project failures. The commercial success of an AI project is dependent on a complex combination of factors, which are too often ignored or relegated to secondary concerns.

However, these supposedly secondary concerns are actually critical to success. These factors are just as important as accuracy, maybe more so, because they are often the root cause behind problematic behavior and failed AI investments. These success factors, which are outlined here, must be considered during the development and implementation of any AI system as they will instill confidence among the system's users and in the leaders that are driving and paying for it:

AI That Works
- **Proven Accuracy –** Is good at solving the problem, as measured by benchmark tests.

AI That Works Reliably
- **Robustness –** Handles unusual or malicious outputs effectively.
- **Dependability –** Always produces an output within the required time frame.
- **Stability –** Performance is consistent and does not drift over time.

AI That's Aligned to Human Expectations
- **Sensibility –** Makes decisions in line with how the world or society works.
- **Humility –** Understands its own limitations, and refuses to answer questions where it doesn't know the answer.
- **Extrapolates sensibly / Fails gracefully –** Acts sensibly when confronted with scenarios beyond those in which it was trained and fails safely.
- **Explainability –** Can justify how it solved the problem rather than working as a mysterious black box.

AI That Works in People's Best Interests
- **Fairness –** Non-biased. Is equally fair to all sub-groups.
- **Sustainability –** Minimizes harmful impacts from training and ongoing use.
- **Privacy –** Protects the sensitive data that it was trained on.

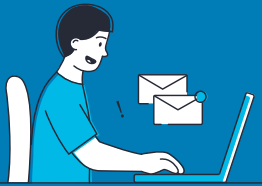# Conclusion: Making AI useful for everyone

We see now that actually solving a task accurately is just one of 12 equally important factors that help everyone to feel much more confident about the AI products and services they use.

We shouldn't make the mistake of thinking elements like humility, sustainability and reliability are the boring secondary elements of an AI endeavor. While focusing on these factors won't create the excitement that comes from an AI-generated image or essay, it will ensure the outputs your business creates are trusted and useful. And once that happens, over time, the chances of failure will reduce, the levels of adoption will increase, and the likelihood of commercial success will be raised significantly.
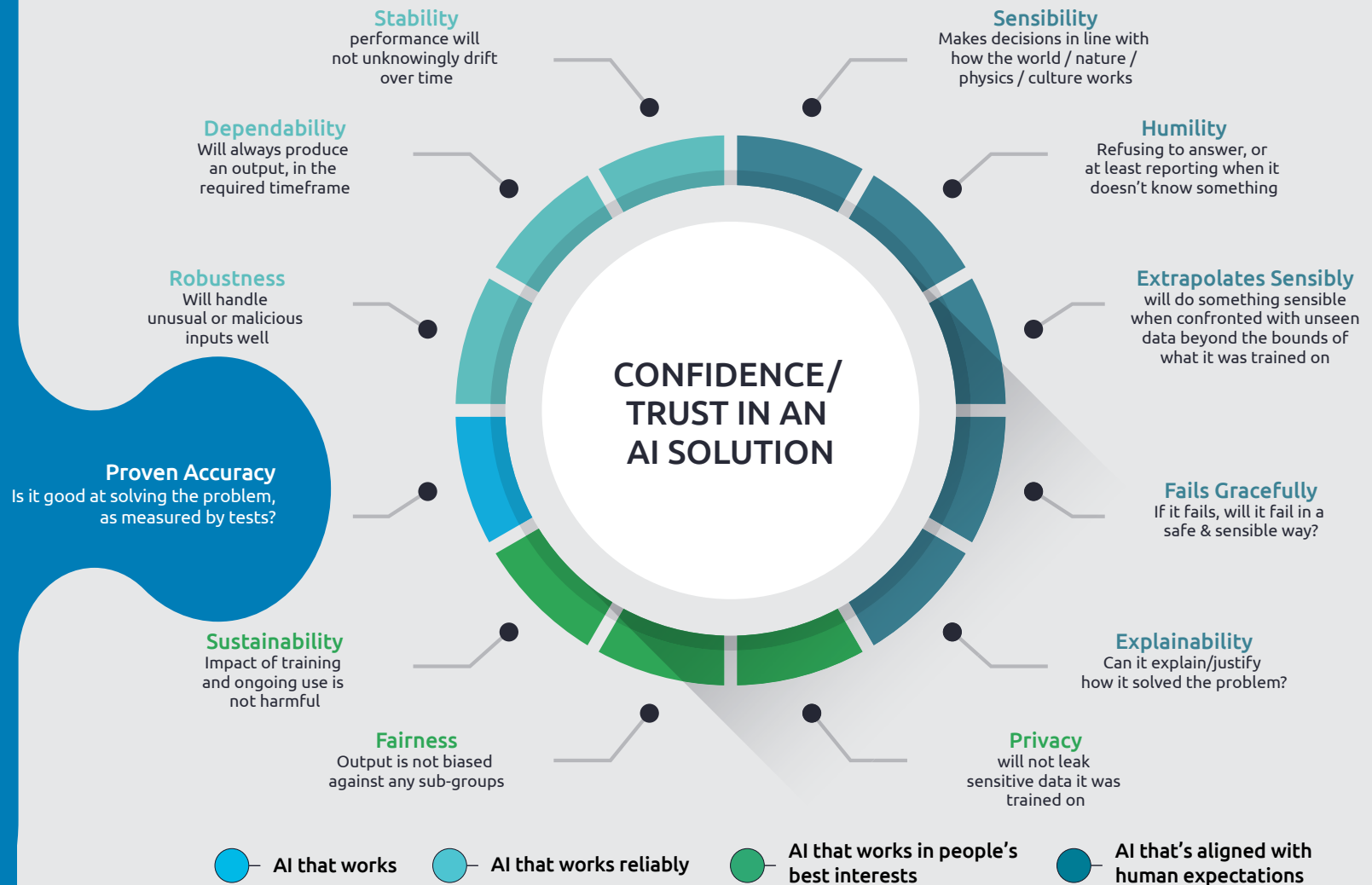
As AI plays an ever-increasingly important role in our lives, people must feel confident in the solutions they use. Ensuring these 12 factors are always considered will mean your business delivers significant commercial value from AI. In this playbook, we will discuss each of these 12 factors in more detail.

Things we Now recognize are crucial to make AI successful
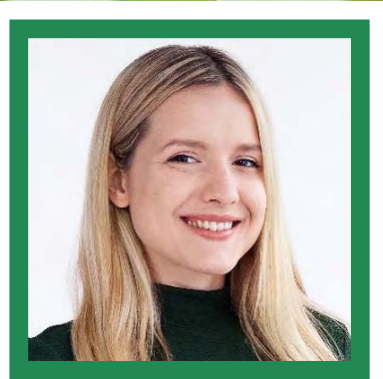
Things people normally focus on in AI

CONFIDENCE / TRUST IN AN AI SOLUTION

Stability
performance will not unknowingly drift over time

Sensibility
Makes decisions in line with how the world / nature / physics / culture works

Dependability
Will always produce an output, in the required timeframe

Humility
Refusing to answer, or at least reporting when it doesn't know something

Robustness
Will handle unusual or malicious inputs well

Extrapolates Sensibly
will do something sensible when confronted with unseen data beyond the bounds of what it was trained on

Proven Accuracy
Is it good at solving the problem, as measured by tests?

Fails Gracefully
If it fails, will it fail in a safe & sensible way?

Sustainability
Impact of training and ongoing use is not harmful

Explainability
Can it explain/justify how it solved the problem?

Fairness
Output is not biased against any sub-groups

Privacy
will not leak sensitive data it was trained on

● AI that works   ● AI that works reliably   ● AI that works in people's best interests   ● AI that's aligned with human expectations

PROVEN ACCURACY

TIJANA NIKOLIĆ
EXPERT IN RESIDENCE

When do we get to say that AI is good enough? What does "good" even mean?

Generative AI has thrust AI into the spotlight in sectors from creative arts to data analysis, and customer service to engineering. However, this rapid rise has brought to prominence a long-standing question in AI: What does it mean for AI to be "good"? Traditionally, the performance of machine learning models has been assessed only through narrow measures of test and validation scores. However, the new focus on Generative AI with its creativity and hallucinations has forced us to reconsider what accuracy really means or whether accuracy is even relevant in this new world. Simplistic measures of accuracy are no longer good enough for us to base decisions on as the different accuracy measures we use can dramatically influence how we interpret their outputs.

It is imperative to also consider real-world dimensions. A model might perform exceptionally in tests but fail profoundly when applied to real-world scenarios. This discrepancy highlights the importance of a comprehensive definition of goodness—one that incorporates various facets such as ethical implications, social impact, and alignment with human values.

WHY?

- Anyone who is involved in decision making around AI needs to understand its performance. This is true both of the users of a system, and of the people designing, building and funding it.

- This need to understand performance makes it highly desirable to create a single, easily digestible number – accuracy, which represents that performance profile.

- However, in almost all cases, no single number can tell you the whole story of how a machine learning system performs, so we need often need to use multiple metrics to describe the performance profile.

- Even if we could capture the how "good" a model is in a single number, that is not enough as "good" is a subjective term.

- Understanding the multifaceted essence of what success in AI looks like is pivotal due to the potential consequences of focusing too much on any one facet.

- In some cases, focusing on the wrong type of accuracy can cause real-world harm. For example, a study of breast-cancer screening in the UK showed that a naïve focus on the wrong sort of accuracy led to over-diagnoses and many women unnecessarily undergoing painful and stressful treatments.

WHAT?

- Consider a simple measure of accuracy for an AI computer vision system classifying 100 objects, either apples or oranges. We could calculate the accuracy of that system by just measuring the percentage of classifications that are correct.

- However, this percentage would only be a useful measure if there were *exactly* the same number of items in both classes. If, however there were more apples than oranges, a simple percentage accuracy figure would not accurately reflect the performance of the classifier. In an extreme case, if there were 99 apples and one orange,

and the classifier *always* said "apple" it's naïve accuracy would be 99%, even though it had no ability to detect the difference between the classes.

- For this reason, more complex statistical measures are used, often such as precision & recall, or sensitivity & specificity. These measures describe different facets of accuracy, showing how well it performs in both its positive and negative predictions, repeatably over multiple uses.

- However, even using these more sophisticated measures such as accuracy, precision and recall does not mean your model's real-world success is guaranteed.

- In fact, as we will show in this Playbook, accuracy on benchmark tests is only one of many equally important facets of success that must be considered in order to not just be successful on paper, but to have genuine real-world success with users who are confident in that system.

RECOMMENDATIONS

- First, ensure you are measuring and communicating accuracy effectively. It is extremely unlikely that accuracy can be represented by a single number, so use more appropriate measures to set users' expectations about the performance profile of a system.

- Don't use simplistic measures of accuracy as the sole criteria for declaring success in an AI system.

- Educate everyone in the business about how to talk about accuracy in AI systems. Strive for a culture where everyone, right up to the boardroom, is comfortable asking questions about sensitivity and specificity, precision and recall etc.

- Beyond accuracy, a holistic approach is necessary. Organizations must embrace transparency, ethics, and fairness in their AI endeavors. Consider using a playbook, like this one, to remind everyone involved in AI systems design to think about the multiple facets that lead to successful AI, not just on accuracy alone.

- One of the primary pitfalls is a myopic focus on technical metrics. Ignoring biases in training data, overlooking

ethical implications, or neglecting community feedback can lead to catastrophic outcomes. Contextual fit, for instance, cannot be measured easily. But is the final defining factor for "goodness"

LINKS

- Validating Large Language Models with ReLM. Kuschnick et al. Carnegie Mellon University, 2023. https://arxiv.org/pdf/2211.15458.pdf

- Langchain blog post: "How Correct are LLM Evaluators", problematizing the possibilities to facilitate measurement of "proven accuracy". https://blog.langchain.dev/how-correct-are-llm-evaluators/

- GEDLT project on prompting, writing styles and quality of answering The GDELT Project is a realtime network diagram and database of global human society for open research:

  https://blog.gdeltproject.org/large-language-models-llms-planetary-scale-realtime-data-current-limitations/
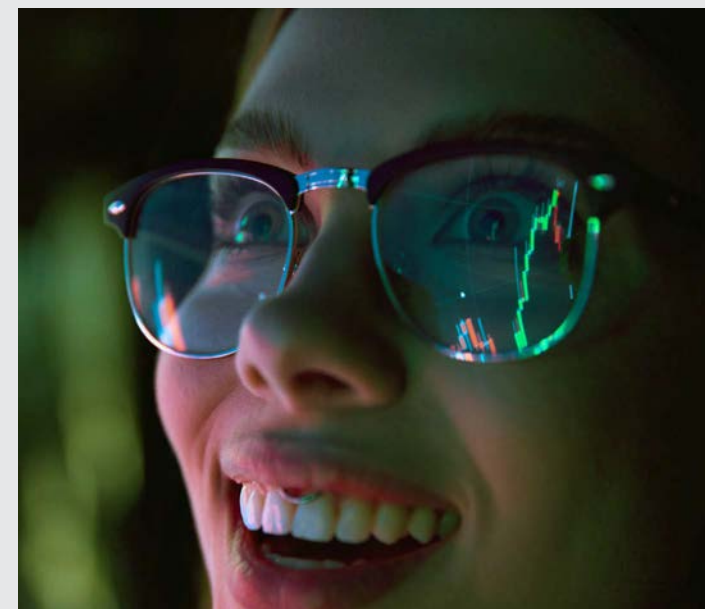
ROBUST

MITALI AGRAWAL
EXPERT IN RESIDENCE

Will an AI system always respond to similar inputs in a consistent manner? Can it cope with deliberate malicious attacks in the input? All of these questions relate to the idea of robustness - a measure of how well an AI system behaves when the signals it receives are not the same as what it was trained on.

Robustness is a cornerstone of reliable AI systems, ensuring resilience in the face of adversity. In the dynamic landscape of artificial intelligence, two paramount challenges arise: dealing with the huge variation of inputs a system will encounter in the real world in a consistent manner and defending against deliberately malicious inputs, often manifested as adversarial attacks.

Understanding and fortifying AI against these challenges is essential in shaping a future where AI technologies can be trusted and relied upon.

WHY?

- In an era where AI is increasingly prevalent in our daily lives, robustness is a fundamental pillar of trustworthiness.
- Due to their complexity though, AI systems are susceptible to various vulnerabilities, both in the algorithms and the data they are trained on.

- A simple way to demonstrate if an AI system is robust or not is to ask it to perform a similar task twice. Providing significantly different answers to the same question will cause humans to rapidly lose trust in the system, but many AI systems will fail this simple test.
- There will always be confusing inputs in the real-world, and unfortunately there will always be malicious actors who try to

deliberately affect the outputs of our AI systems. Even in the best cases, with no malicious actor, we will still forever be locked in an arms race between our machine learning models and the infinite complexity that the real world will throw at them.

- Therefore, there will always be a need to use approaches to maximize the robustness of our AI models, and in some cases we require verifiable proof of that robustness.

- By addressing the complexities of unusual data and adversarial attacks, we pave the way for AI systems that not only excel under ideal circumstances but are resilient in the face of unexpected inputs and deliberate attacks.

WHAT?

- Whilst machine learning experts have long known about the problems of robustness, Generative AI tools now allow everyone to see the extent of this problem – even small changes in the phrasing of a prompt can produce completely different outputs and meanings.

- Deliberately malicious inputs, known as adversarial attacks, exploit the vulnerabilities of AI systems, leading them to make erroneous judgments. These attacks can have dire consequences, especially in safety-related applications such as autonomous vehicles or healthcare systems, so making robust defenses is imperative.

- Adversarial attacks fall into two main classes

  - White-box attacks which use knowledge of the model to achieve their impact.

  - Black-box attacks which do not have knowledge of the underlying model.

- These attacks might also be untargeted, where the aim is to just achieve any corruption of the output, or targeted, where the aim is to coerce the model to produce a specific output.

- Thankfully, malicious attacks on AI systems are relatively rare, and the more common problem is where AI systems encounter atypical, unfamiliar data in real-world scenarios. This can range from novel environmental

conditions for autonomous vehicles to unprecedented user inputs in chatbots, challenging the system's ability to make accurate predictions or decisions.
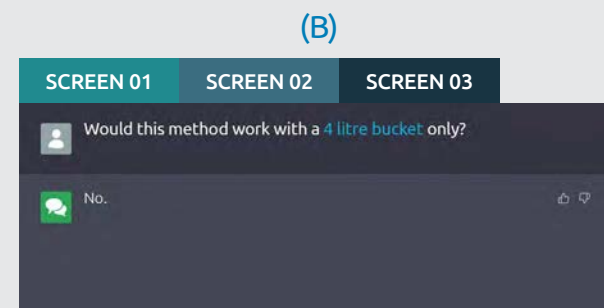
- When faced with unusual data, AI systems might exhibit unpredictable behavior, potentially jeopardizing the trustability of the outputs. Ensuring robustness in such situations necessitates training models not just on larger datasets but on more diverse datasets that encompass a wide array of possible inputs, preparing them for unforeseen scenarios.

RECOMMENDATIONS

- Make sure you understand the scale of the problem in your use case – test your systems to make sure that small changes in the input do not produce significant changes in the meaning of the output.

- For LLMs specifically, guaranteed robustness is much harder to achieve because these models do not actually understand the meaning of the language tokens they manipulate. Where a human might see two phrases as being the same, they could be interpreted in very different ways by an LLM and producing substantially different outputs.

- Additionally, input preprocessing techniques could enhance a system's ability to provide more robust results by ensuring multiple rephrased versions of the prompt produce consistent output.

- Traditionally used in cybersecurity, red teaming involves simulating adversarial attacks to identify vulnerabilities and weaknesses in a system. When applied to AI, red teaming serves as a potent tool to assess the resilience of machine learning models, algorithms, and applications against malicious intent and unexpected inputs.

- We can also use other machine learning systems as a red team, exploiting malicious techniques for positive use in an approach called adversarial training. In this approach models are exposed to adversarial examples during training, enabling them to recognize and resist such inputs. This approach pits one machine learning system

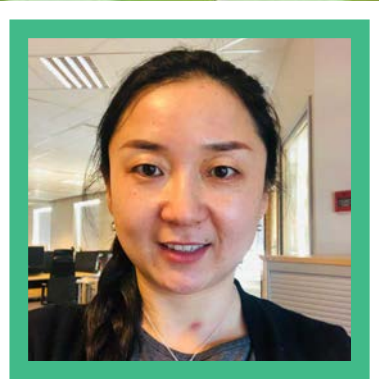against the other, resulting in both being better and the overall system being significantly more robust.

- In some cases, it may be possible to use verifiably robust approaches to training, such as Interval Bound Propagation (IBP), which can guarantee certain levels of robustness, although often at the expense of accuracy i.e. overall accuracy may be lower, but you can be sure that when it does make a prediction it is correct.

(A)

Images showing confusing data an AI vision system might encounter in the real-world, sometimes naturally occurring, sometimes as a result of malicious attacks.

(B)

| SCREEN 01 | SCREEN 02 | SCREEN 03 |

Would this method work with a 4 litre bucket only?

No.

An example of how a seemingly inconsequential reframing of the input to an LLM can produce substantially different, and incorrect, output.

DEPENDABILITY

WEIWEI FENG
EXPERT IN RESIDENCE

Whilst most of the properties described in this playbook relate to the content and quality of an AI system's output, we often neglect some of the operational considerations for deploying AI in real-world situations. One of the most important of these is dependability – will an AI system actually give us an answer when we need it?

As we start to move AI systems from the lab to the real-world, one of the practical realities that we must consider is timing. In many cases, the speed of an AI system's response is crucial. It doesn't really matter if a customer-service chatbot takes 10 seconds to respond, but it would clearly be a big problem if an autonomous vehicle took 10 seconds to consider its actions whilst driving at speed on a road.

This presents an important and difficult dilemma to solve. Modern AI has achieved many impressive results, but this is largely powered by huge neural network models which are slow to execute and require levels of compute power that are normally not available in real-world deployments of AI systems. If we don't have guarantees that an AI system will respond as quickly as we need it to, then confidence and adoption will falter. Fundamentally different architectures are required where timeliness of response is important.

WHY?

- In real-world deployments of AI, timing matters. A high-performing model with good accuracy is worthless if it doesn't respond quickly enough for its output to be used.

- This is most obvious in safety-critical and real-time control situations where non-negotiable guarantees on response times are present.

- Even in situations that are not safety-related, response time can dramatically affect the user experience to the point that they might lose confidence in a system that doesn't respond quickly enough.

- In other situations, low latency responses are required for reasons of responsiveness and throughput. Long deliberation in pursuit of the perfect answer in these types of problems could cause widespread disruption. For example –

  - In credit-card fraud detection, where vast numbers of transactions must be assessed quickly to prevent delays.

  - In AI-supported emergency response systems, where stress and the need for swift information availability have a direct impact on outcomes.

  - In real-time scheduling problems, such as traffic-light scheduling, elevator dispatching etc.

  - Dynamic advertisement selection on websites, where a slow decision would ruin the user-experience of the host website.

WHAT?

- The speed of an AI system's response has always been a primary consideration in AI research, as many of the classic benchmarks of artificial intelligence have a timing element to them – playing games, having conversations, driving vehicles, interactive robots etc.

- In many cases it may be possible to solve a machine learning task with good or even perfect accuracy if timing is not an issue, but solving the engineering problem of deploying that same model into a more constrained and time-critical environment may be impossible.

- In some cases, it may be possible to compress or prune a large model, to improve its response time. This is already commonplace in many Edge AI deployments in order to squeeze more performance out of limited hardware. However, whilst this approach *improves* performance, it cannot *guarantee* performance.

- Most machine learning models are effectively non-deterministic, meaning that the execution of the model (inference) will never be predictable. Therefore, if *guarantees* of performance are required, then just shrinking a big model will never be the answer. Fundamentally different architectures are required.

- The most obvious and well-known architecture is the so-called classifier-cascade. In this case, machine learning models are arranged in a cascade, starting with extremely simple and small classifiers that can provide a quick answer immediately. If time allows, processing passes on to a more complex but time-consuming classifier, and this process continues to the bottom of the stack. This architecture means we can interrupt the processing at any point and get an answer.

- This is similar to what we see in human decision making, where we have fast "System 1" thinking to give an immediate response, followed by slower and more deliberate "System 2" thinking. In the case of classifier cascades, there may be hundreds of levels, iteratively refining the answer as far as time allows.

- In most cases, the first level of such a cascade would be a non-AI system, which encodes basic default behavior.

- The performance of systems can be enhanced through tiered approaches. At each tier, the solution should be evaluated against the previous tier to determine if it provides a significant improvement. This evaluation process allows for early termination of inferior tiers, thereby streamlining the overall process.

RECOMMENDATIONS

- Carefully consider the Quality of Service (QoS) requirements of a machine learning solution from the beginning of a project. It is much easier to design for low-latency from the start rather than trying to retrofit it afterwards.

- Define QoS requirements and strategies upfront. Make sure this defines latency expectations, and what you want to happen if that can't be achieved. For example, "I'll take whatever answer you have after 150ms" or "if you can give me a better answer in 0.1s, use that, otherwise fallback to hardcoded response".

- Remember that just because it's possible to create a high-accuracy machine learning model in the lab (given enough compute power and time), it does not mean you will be able to achieve that level of performance in the real-world where both compute power and time are limited.

Use cascades of classifiers to ensure you can always get an answer, but those answers will improve the more time you allow.

LINKS

- [A presentation on real-time ML for fraud detection.](#)

- [An example framework to allow time-critical decision making in scheduling problems.](#)

- LLM Inference Performance Engineering. Databricks blog: https://www.databricks.com/blog/llm-inference-performance-engineering-best-practices

- NVIDIA improve LLM inference with Tensor RT https://developer.nvidia.com/blog/nvidia-tensorrt-llm-supercharges-large-language-model-inference-on-nvidia-h100-gpus/

DEPENDABILITY

STABILITY

NIHARIKA KALVAGUNTA
EXPERT IN RESIDENCE

AI systems must be stable to be trusted and effective. Unlike traditional IT systems, AI can adapt its behavior over time, posing challenges for business leaders. With dynamic behavior, it is crucial to continuously monitor and ensure the stability of their AI systems.

Continuous monitoring and maintenance are crucial for AI systems, especially Generative AI, due to their complex behavior and dynamic nature. Unlike traditional IT systems, AI systems can adapt and evolve over time, necessitating ongoing monitoring and adaptation of the AIOps and MLOps frameworks to effectively manage these systems. As AI permeates deeper into the fabric of society, the need for dedicated monitoring and maintenance services is becoming vital. This will require specialized expertise and tools, catalyzing the emergence of an entirely new industry sector dedicated to this crucial task.

WHY?

- System stability in the age of Generative AI is no longer mainly a technical endeavor, as psychological, sociological, and geo-political factors around AI system behavior all matter to ensure stable performance.

- Technical system stability refers to the hardware and software that should deliver stable performance over time in terms of uptime, execution time, system footprint, and various other technical factors known from CI/CD and DevOps frameworks.

- Generative AI presents particular challenges, as they will often form the front-line of interaction with users who will require consistency

…

over time in order to build trust in those systems.

- Behavioral drift in the foundational models that underpin many other systems could have far-reaching consequences, and could accidentally or deliberately spread or amplify misinformation that could affect entire societies.

- As one can imagine, unstable systems are not only susceptible to catastrophic failures but also to (malevolent) third-party exploitation. Political agendas, people with different interests, and nation states might see new opportunities to destabilize their adversaries by potentially influencing opinion and sentiment. Stable GenAI systems should withstand or at least identify malicious and adversarial attacks, so that the negative effects can be minimized.

WHAT?

- MLOPS provided IT professionals with the tools to monitor AI systems, detect drift and retrain models at scale. Such tools now need to be extended with techniques to encompass GenAI model system performance, including input content metadata, model recall, precision, system sustainability, footprint and most importantly, and social alignment.

- AI Ops should be extended to cover parts of model governance. Indeed, systems like ChatGPT already utilize AIOps to align and restrict model behavior, e.g. by utilizing red teaming techniques in input/output monitoring and sanitizing results before their return to the user.

- Continuous training of models on new information is necessary to ensure their relevance in dynamic domains. However, this can cause regression and instability of a system, especially when continuous training unpredictably changes the quality of a model on specific tasks. Communicating such changes clearly to the end users of a system is vital, otherwise the system will be perceived as unstable.

- Most users will assume that an AI system can never become less smart, or "forget" things it previously knew, but this is a plausible side effect of continuous retraining, and users need to be educated about such impacts.

- Measurements for model performance and quality are needed, together with processes and methods to capture and monitor performance over time.

- GenAI systems need to be regarded as stable in order for them to be useful. Typically, this means that results do not drift over time, but also that other aspects such as interpretability, ethical decision making, explainability are similarly stable and behave in the same way each time.

RECOMMENDATIONS

- Make sure to appreciate the difference between technical stability, model stability, contextual and behavioral stability when it comes to applications of Generative AI.

- Especially while depending on external model APIs when using large foundational models, have a CI&CD/DevOps and fall-back strategy in place.

- Implement model-governance, so that deviating behavior can be traced back to the model version and properties leading to it. Especially in cases of fine-tuned models or RAG approaches with regular updates, governance and traceability is even more important. This aspect of ML governance will reach far beyond the perimeter of the ML model.

- Implement data governance, both when building from scratch, finetuning or in case RAG approaches are used. Models may be dependent on data which is malicious, a fact that is not always visible at the outset.

- Once governance on solution level is in place, monitoring of model behavior should be implemented. Unintended user input, adversarial attacks, and model drift over time can all lead to unwanted model behavior which should be identified as early as possible. Implement tools to test and observe model and customer behavior.

- Monitor, model and visualize logs of model and user behavior over time in order to understand the baseline


profile of the system and deviations from it. Use standard open-source benchmarking frameworks as well as bespoke test-suites to give multiple signals for monitoring.

LINKS

- [Research into how machine learning models can forget previously learning information during retraining.](#)

- DataRobot on stability in AI: https://www.datarobot.com/trusted-ai-101/performance/robustness-and-stability/

SENSIBILITY
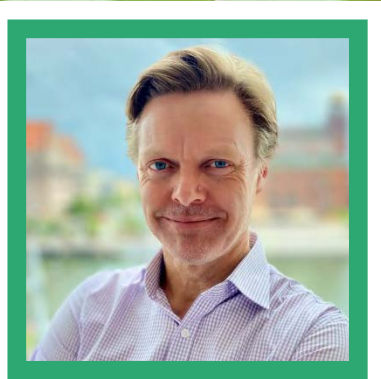
JOHAN MÜLLERN-ASPEGREN
EXPERT IN RESIDENCE

ROBERT ENGELS
EXPERT IN RESIDENCE

As AI technology gets more advanced, the notion of 'AI alignment' has become a fundamental aspect of creating artificial intelligence. An aligned AI system is not just sophisticated and high-functioning, but also acts in accordance with human values, norms, and ethical standards. This challenge transcends technology and aims to ensure that AI's actions are truly in line with the subtle and diverse aspects of human welfare and culture.

Sensible AI is an evolution of intelligent systems that are not only capable of learning and decision-making but also of understanding and aligning with human values, ethics, and safety considerations. A sensible AI in traffic management would not only consider the efficiency of traffic flow but also the safety of pedestrians, the urgency of emergency responses, and the environmental impact of rerouting decisions. It would be an AI that, when faced with a decision, weighs the outcomes not just by a single metric but by a spectrum of human concerns. With this in mind, we need to ask ourselves - how can we ensure that the development of AI continues to factor in the complex tapestry of human ethics, especially when the consequences of not doing so can be far-reaching and profound?

WHY?

- **AI systems are becoming increasingly powerful and capable:** They are able to learn and make decisions on their own, with minimal human intervention. This means that they have the potential to do a lot of good, but also a lot of harm.

- **AI systems are not able to capture human values.** They are typically designed to optimize for a specific task, such as playing games. However, these tasks may not align with our human values. For example, an AI system that is designed to win at all costs could potentially harm humans in the process.

- **It is difficult to predict how AI systems will behave in the real world.** AI systems are trained on data, but the real world is much more complex than any data set. This means that it is difficult to predict how an AI system will behave in a given situation.

WHAT?

- **Cooperative Coexistence:** As AI systems become more integrated into daily life, ensuring they can cooperate with humans and understand human needs and boundaries is vital for a harmonious coexistence.

- **Trust:** Humans are more likely to trust and adopt AI technology if they believe that the AI's actions and decisions align with their expectations and well-being.

- **Ethics:** Aligning AI with human values and ethics helps prevent scenarios where AI might otherwise take actions that are considered unethical or harmful to societal norms and values.

- **Social Impact:** AI has the potential to significantly impact society, and ensuring this impact is positive requires aligning AI with the broad spectrum of human values and societal goals.

- **Cultural Sensitivity:** AI systems must be sensitive to cultural differences to operate effectively and respectfully across the globe, necessitating an alignment that considers diverse cultural norms

and practices.

- **Contextual Awareness:** humans use a model of the world, combined with observations and counter-factual thinking in order to understand context in real-world situations. AI is not yet capable of the same.

- **Metacognition:** AI systems must gain the ability to reflect over their decisions and beliefs in order to be able to realign if needed.

- **Accountability:** If AI systems are to make decisions that affect humans, it is important that these systems are aligned with human intentions to ensure accountability for their actions.

- **No single right answer:** Looking across all of these points, it's clear that in many of these aspects humans themselves do not agree on what is right and wrong. How then can we expect AI systems to do the right thing, when we cannot even define what right is?

RECOMMENDATIONS

- **Establish General Guidelines:** Organizations should define and adhere to a set of principles that govern the development and deployment of AI systems, and set expectations about the behaviour of those systems, especially "red lines" that it should never cross. Consult a diverse range of stakeholders including ethicists, sociologists, and representatives from various cultural backgrounds.

- **Promote Transparency:** Invest in AI transparency and explainability, making it easier for users to understand how AI systems make decisions. This could involve the development of standards for explainable AI and the integration of explanation features within AI systems.

- **Define boundaries:** Philosophers have been arguing about moral and ethical dilemmas for centuries, so clearly there is no single right answer for many situations. The practical reality is that we need to allow a range of different actions for our AI systems. An autonomous vehicle driving on public roads would be tuned very differently to an autonomous military vehicle in a battlefield.

- **Public Engagement:** Engage with the public and stakeholders to gain insights into societal values and concerns regarding AI. This could take the form of surveys, public forums, and collaborative projects that involve community input.

LINKS

- Kahneman, D. (2011). Thinking, fast and slow. macmillan. https://en.wikipedia.org/wiki/Thinking,_Fast_and_Slow

- Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 188 (April 2021). https://doi.org/10.1145/3449287

- Mark Kelly, Andrew B. Barron, The best of both worlds: Dual systems of reasoning in animals and AI, Cognition, https://www.sciencedirect.com/science/article/pii/S0010027722001068

- Symbiotic human-robot collaborative assembly https://www.researchgate.net/publication/335277208_Symbiotic_human-robot_collaborative_assembly

- Making AI delivery robots disability-fiendly and 'cautious pedestrians' https://www.bbc.com/news/disability-65719649
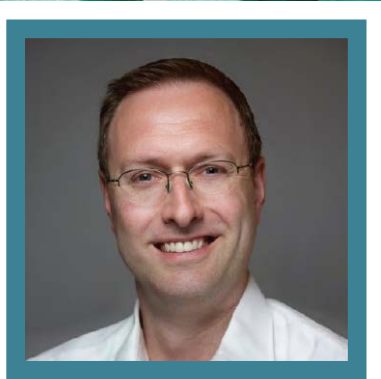
HUMILITY

BEN MATHIESEN
EXPERT IN RESIDENCE

MARK ROBERTS
EXPERT IN RESIDENCE

One of the biggest reasons that people don't trust AI systems is their lack of humility. In short, AI systems don't know what they don't know.

When a machine learning model is trained, it builds up expertise based on the training data it is shown. However, that training data can never be fully representative of the real-world, so there will be deficiencies in its knowledge, and vast areas where it has no knowledge at all. The problem is, the machine learning system does not "know" whether it is operating in an area that it knows well, or in one of its blind-spots, and not knowing could be catastrophic.

In human society, this sort of humility is a great asset. Being able to say "I don't know" allows us to avoid being forced into a risky decision where we don't have enough information to be confident. Ironically, we trust people more when they have less faith in their own abilities, because we value confident decisions much more so than decisions with no confidence attached to them. If a person, or a machine learning model, says that it doesn't know something, we will be happier to trust their outputs in the future, safe in the knowledge that they possess the humility to refuse to answer if necessary.

WHY?

- When humans are faced with challenges outside of our learned experiences, we can generally cope quite well by slowing down and switching modes: either by applying logic and reasoning, or cautiously transferring experience from analogous tasks.

- However, when a machine learning system is operating outside of its "comfort-zone" it doesn't switch modes. Applying the same models to unfamiliar data, the outcomes are not just often wrong, but can be completely undefined.

- This could be catastrophic in safety-critical systems. When human lives are on the line, AI systems need to be able to recognize and communicate when they are being asked to make a decision outside of the envelope of their experience.

- In an autonomous vehicle for example, we need the AI system to be able to recognize when it is being confronted by an unfamiliar scenario where its ability to make a confident decision is compromised. In those cases, the system should quickly delegate responsibility to another failsafe system, or to a human, to take over.

- As well as *accidentally* failing when outside of its comfort-zone, systems that fail to identify when they're being asked to extrapolate beyond their experience might also be more susceptible to adversarial attacks, which rely on the gaps in the AI's knowledge to construct inputs that provoke dangerous results.

- Humility is also a prerequisite for failing gracefully.

WHAT?

- In AI terms, humility is often described as the Out of Domain problem (OOD), where the "domain" is the part of the problem space that the AI model was trained on. Future challenges to the AI might be In-Domain, meaning that the inputs relate to an area of the training space it knows about, or Out of Domain, meaning that it relates to an area that the model has little or no experience of.

- The datasets used to train AI models occupy a high-dimensional space with complex geometry.

Mathematically, this means that even if we can measure that an input is close in distance to a previous training point, this simple measure doesn't guarantee that it is in-domain and will result in a good prediction.

- Detecting OOD inputs reliably therefore requires other techniques. One approach is to look for geometrical surfaces that bound the training data. Another is to transform the data into another set of variables: a more compact representation that makes OOD inputs more obvious. Probabilistic and ensemble-based approaches also exist but are more computationally expensive.

- The output of such OOD-detection systems could be a simple yes/no response, indicating whether the input point is in-domain or out-of-domain, or a continuous confidence score.

- OOD detection has parallels with other tasks where we need to determine whether a new input sits within an existing distribution of data. Experience with anomaly and outlier detection provide useful statistical tools for tackling these problems.

SYSTEMIC TRUST VS INDIVIDUAL TRUST

When we talk about humble or trustworthy AI, there are two types of trust to consider.

- Systemic trust means that the AI application performs well and provides unbiased results on a large dataset. For example, a criminal justice application to predict the risk of recidivism could be shown to have high precision overall and have similar error rates across categories defined in terms of race, age, and gender.

- Individual trust means that each single prediction is understood, and its error remains within known limits. If the system predicts that a certain individual is high-risk, then it should be able to show that the available data on this person are like known examples, not an outlier or an unusual combination that produces an OOD input where its model is not trustworthy.

To be accepted in society, especially in safety-critical areas such as medicine, justice, and automation, an AI system must be able to show both systemic and individual trust.

RECOMMENDATIONS

- Carefully consider whether your use-case could suffer from an out of domain problem. Will it encounter inputs from outside of the domain it was trained on? Some narrow AI applications operating in controlled environments may be nearly immune from OOD inputs, but most cases where an AI system is deployed into the infinite complexity of the real world will need to consider the possibility.

- For systems that might encounter out-of-domain inputs, you need to consider several things -
  - What is the business or human risk of the system producing incorrect or undefined outputs in response to OOD inputs?
  - Outside the data itself, do other systems allow us to detect whether an input is out-of-domain?
  - Is there a fall-back system in place that can handle out-of-domain inputs and wrong predictions on a case-by-case basis?
  - Do we have sufficient training data to show that the behavior of the system is globally precise and unbiased (systemic trust) and to reliably detect OOD inputs (individual trust)?
  - Can we build in modules that give the system this capacity to recognize its own limits and communicate risk to end users?

- For safety-related systems, or any system where incorrect output could cause harm, bias, or unintended consequences, having a fallback system, and knowing when to defer to it, is vital.

LINKS

- Capgemini have developed methods for detecting OOD inputs which is already in use in multiple sectors.

- Researchers from University of California San Diego unveiled COMPOSER (COnformal Multidimensional Prediction Of SEpsis Risk), a deep learning model for the early prediction of sepsis, specifically designed to reduce false alarms by detecting unfamiliar patients/situations. COMPOSER flags unfamiliar cases as indeterminate rather than making false predictions.

FAILS GRACEFULLY / EXTRAPOLATES SENSIBLY
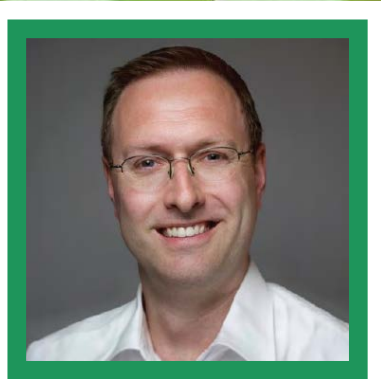
MARK ROBERTS
EXPERT IN RESIDENCE

For both humans and AIs, how badly you fail, is often much more important than how well you succeed.

It's a depressing reality that a system could be correct 99.99% of the time, but still be judged as a complete failure if the 0.01% of incorrect cases are badly wrong, dangerous, or morally unacceptable. It may seem harsh, but when deciding whether we should have confidence in a machine learning system, we tend to judge them not by their biggest success, but by their worst failure.

It's easy to forget that humans are naturally good at failing well. If we don't know how to do a task, we fall back on our years of experience solving similar analogous tasks and use that to at least fail in a way that's sensible and minimizes the impact of that failure. Catastrophic failures are relatively rare - most of the time, humans fail in various ways, hundreds of times a day, in ways that are not a big deal.

This graceful failure mode is what we now need from our AI systems. Rather than falling off a mathematical cliff and failing in surprising and unpredictable ways, we need AI systems to fail in ways that are aligned with human expectations.

WHY?

- It is useful to think of a machine learning model as being like the problem of curve fitting (although in many more dimensions). We fit a curve to a cloud of data points until we are satisfied that it approximates the shape of the data enough that we can use the fitted curve as a general model of that type of data.

- However, anyone who has ever done any curve fitting will know that outside of the region of the input data, the fitted curve can deviate off in wild directions that have little relation to how well it performed inside its "comfort zone".

- This is the situation with machine learning. Whilst we can be very confident about performance within the model's comfort zone, i.e. the area with data to learn from, outside of this zone the model's output can be completely undefined and it may suggest actions that would seem ridiculous to a human looking at the same data.

- This is a confusing situation for humans. When we see AI systems performing super-human feats on complex problems, it biases us to think that they will always have this level of performance, but most people do not realise that outside of their trained experience, these systems might do things that are naïve, dangerous, or disconnected from any sort of common-sense.

"We [humans] are able to generalize from observing performance at one task to a guess at competence over a much bigger set of tasks. We understand intuitively how to generalize from the performance level of the person to their competence in related areas…But the skills we have for doing that for a person break down completely when we see a strong performance from an AI program. The extent of the program's competence may be extraordinarily narrow, in a way that would never happen with a person." **– Rodney Brooks**

- This mismatch in expectations can create serious problems when safety-critical systems are involved. When we see an autonomous vehicle driving safely for 98% of the time, we might become convinced that it is generally safe and the 2% of failures won't be too bad. However, the default failure mode of most machine learning systems is undefined, and the resulting behavior could be anything from a minor issue to a catastrophically nonsensical decision.

WHAT?

- The concept of graceful degradation is common in many engineering disciplines. Where continuous operation of a machine is important, we employ backup and failsafe systems to take over should the primary systems fail.

- The same concept is useful in AI. As out-of-domain behavior cannot be guaranteed or trusted, secondary systems can be employed to intercept and override bad decisions.

- The most famous form of these is that of fundamental laws. Author Isaac Asimov popularized this approach with his 3 laws of robotics -

  - A robot may not injure a human being or, through inaction, allow a human being to come to harm.

  - A robot must obey orders given it by human beings except where such orders would conflict with the First Law.

  - A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

- Whilst this idea is from science-fiction, the basic principle still holds – that we can use simplistic rules to catch and overrule incorrect decision coming from other systems. These rules provide a failsafe – whatever any other subsystem has suggested, if it breaches these rules, it should be ignored.

- In reality, modern AI system architectures implement more nuanced versions of this approach, with hierarchies of systems that can override and supersede the outputs of other systems.

- For example, in an autonomous driving system, we should not just blindly connect the output of a neural network to the vehicle's control system. If that system misinterpreted the road-signs approaching a junction and failed to yield, other subsystems with other responsibilities (mapping, road law, culture) should be able to overrule the decision.

RECOMMENDATIONS

- AI should never be put in sole charge of important decisions or systems.

- Where the impact of failure is high, AI systems must be wrapped in secondary systems that will detect and overrule decisions when necessary.

- Rather than relying solely on *statistical* machine learning models (like neural networks), consider also using *contextual* models of the world. These contextual models are not based on prior experience, but are instead based upon and understanding of the rules that govern the world in which it operates. This means that when the model encounters things it hasn't seen before, it can fall back and use fundamental reasoning about the world to figure out what is happening. This concept, and another related concept called zero-shot learning, allows us to make sensible decisions even when we have never encountered similar situations before.

- Using higher-level representations of the world, and cascades of classifiers may also help. An autonomous vehicle may have never seen a particular vehicle (for example a convertible 3-wheeler car towing a caravan), but it can at least recognize that is a vehicle and that it should be handled in the same way as other vehicles. Abstracting to higher level concepts like vehicles, humans, animals, structures allows general principles to be applied and allow a more graceful handling of unknown situations.

LINKS

- Isaac Azimov. "I, Robot". 1960. Early attempts into graceful degradation and sensibility. Three laws of Robotics make their first appearance in the book. https://en.wikipedia.org/wiki/I,_Robot

- Jack Dymond, "Graceful Degradation and related fields". University of Southampton, 2021. https://arxiv.org/pdf/2106.11119.pdf.

- M. Buehler, 2023. "MechGPT, a Language-Based Strategy for Mechanics and Materials Modeling That Connects Knowledge Across Scales, Disciplines and Modalities". https://asmedigitalcollection.asme.org/appliedmechanicsreviews/article-abstract/doi/10.1115/1.4063843/1169582/MechGPT-a-Language-Based-Strategy-for-Mechanics

- Huang, J. et al. 2023. Google Deepmind. "Large Language Models cannot self-correct reasoning yet". Proposals for approaches to let LLM attempt to correct its initial responses based solely on its inherent capabilities. https://arxiv.org/pdf/2310.01798.pdf

FAILS GRACEFULLY

EXPLAINABILITY

ROBERT ENGELS
EXPERT IN RESIDENCE

As AI systems become deeply embedded in our societal framework, the necessity for them to be trustworthy cannot be overemphasized. At the heart of this trustworthiness lies the principle of explainability.

Explainability in AI means understanding how a decision or output is reached. This transparency is important for ethical and legal reasons, especially when AI affects people's lives. Explainability builds trust in AI systems.

While complete transparency is ideal, it becomes increasingly difficult as AI models become more complex. We need to find alternative ways to ensure safe AI usage, even with limited understanding. In many cases, we don't require real-time explanations for every action, but just the assurance that an explanation is available if we need it. This allows us to trace back and understand AI behavior when unexpected events occur.

Explainable AI is not just a technical challenge but also a societal one. AI will never reach its full transformative potential if it can't reach a level of explainability that allows its users to have faith in its decisions.

WHY?

Explainability is needed in many settings, and has various dimensions on impact.

Societal Impact:

- *Increased Trust:* Transparency in AI decision-making will engender greater trust among the public, paving the way for more widespread acceptance and utilization of AI technologies.
- *Informed Decision-Making:* Users and stakeholders can make more informed decisions when they comprehend AI recommendations, especially in critical areas like medicine, law, and public policy.
- *Fairer Outcomes:* Identifying and addressing biases through explainability can lead to more equitable AI decisions, minimizing discrimination and fostering inclusivity.
- *Empowered Public:* With explainable AI, the general populace can participate in AI discourse, ensuring that technological advancements are aligned with societal values.

Business Impact:

- *Risk Mitigation:* Explainable AI can help businesses in identifying, understanding, and reducing potential risks, particularly in sectors like finance and healthcare.
- *Improved Model Robustness:* By understanding model decisions, businesses can optimize, refine, and ensure the robustness of their AI systems, leading to more reliable outcomes.
- *Competitive Advantage:* Companies that adopt explainable AI can differentiate themselves in the market, appealing to a customer base that values transparency and trustworthiness.
- *Regulatory Compliance:* In industries with strict regulatory guidelines, like the banking, health and automotive sectors, explainable AI can ease the process of compliance by providing clear justifications for automated decisions.
- *Enhanced Customer Relations:* By offering clear insights into AI-driven recommendations or decisions in customer interactions, businesses can improve trust in their services, building loyalty.

WHAT?

Explanations come in flavours depending on need and usage:

- *Causal explanations:* show why the input produces a specific output. This could take the form of words, flow-charts, visualisations, or other appropriate formats.
- *Raw explanations:* An explanation could be based on a raw unfiltered dump of the decision path. These "Engineers' Explanations" could help in detailed diagnostics, but would not be appropriate for typical users.
- *Saliency-based explanations:* Often, users do not need to know exactly how a decision was arrived at, but need assurances that the salient features of the data was used, rather than a conflating variable.

When people use AI, a cognitive, psychological dimension is important:

- *Interpretability:* Refers to the degree to which a human can understand the cause of a decision made by an AI system, even without any external aid or tool.
- *Trust-amplifying explanations:* The way an explanation is delivered can dramatically affect how it is received. Tone-of voice has shown to be a decisive factor in the acceptance of tools like ChatGPT.
- *Bridge for Non-experts:* Explainability acts as a bridge between complex AI algorithms and non-experts, fostering a broader understanding, acceptance and trust.
- *Feedback Loop:* By making AI decisions comprehensible, explainability establishes a feedback loop with users, allowing systems to improve based on human feedback and collaboration.

And last but not least, there are purely technological aspects of explanations:

- *As a Tool for Debugging:* By understanding the decision paths, developers can more effectively debug, refine, and optimize AI models.
- *For Bias Detection:* It helps in identifying and understanding biases in model decisions, thereby promoting fairness and ethical AI implementations.
- *To gain Model Insights:* Through explainability, insights are gained into the workings of complex models, such as neural networks or ensemble methods.

RECOMMENDATIONS

- First decide what your requirements are for explanation. Do you actually require every decision to be fully explained? If so, to whom? Do you need explanations that experts in ML and mathematics can understand or explanations that anyone can understand?
- In most cases, general purpose explanations to a general audience will not be possible. In many cases the ability to trace back through a decision chain is enough, even though that capability might only rarely be used and time-consuming.
- Think about whether you require a full step-by-step explanation or whether you just require higher-level indicators of how a decision is being made. For example, in an image classification use-case, it might be enough to know which parts of the image are being used to inform the decision, rather than understanding every step along the imaging pipeline.
- Consider whether different architectures will increase or decrease your ability to explain the model sufficiently. In general, the larger the model, the harder it will be to explain its actions, so always use the smallest model that can achieve your desired performance. In many cases, explainability might be more important than performance, so you should consider sacrificing accuracy for better explainability.
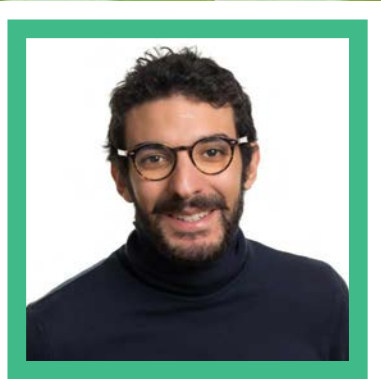
LINKS

- LIME ((Local Interpretable Model-agnostic Explanations): https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/
- SHAP (SHapley Additive exPlanations): https://shap.readthedocs.io/en/latest/
- Saliency Maps: https://en.wikipedia.org/wiki/Saliency_map

FAIRNESS

ADAM MEHDI ARAFAN
EXPERT IN RESIDENCE

Whilst initially considered a uniquely human flaw, the inability to make fair and unbiased decisions is now also one of the key challenges in the adoption of AI. Controversial failures such as COMPAS, PredPol and the Dutch childcare benefits scandal exemplify the dangers involved when biased representations are deployed at scale without the right mitigation strategies.

In an effort to capture the essence of our rich and complex world, we train machine learning models on vast amounts of data. Large state-of-the-art models such as Meta's Llama v2 were trained on two trillion tokens of publicly available data; the equivalent of 24 million 300-page books. However, despite significant efforts in data curation, bias inevitably leaks into that training data, distorting the model's interpretation of the world.

Once learned, biased representations can manifest themselves in both obvious and more subtle ways. The generation of toxic language is easy to spot, but the effects in hiring or fraud detection systems may manifest as a subtle but systematic de-ranking or misclassification of vulnerable sub-groups in society. Although training data is often one of the main culprits, bias can also be caused by poorly defined goals. Therefore, when designing a machine learning system, one should first always consider the fairness implications and second, perform bias mitigation as an end-to-end process instead of a point-wise fashion.

WHY?

- Humans are of course also guilty of biased decision-making although the impacts of such decisions can be more easily audited and mitigated by other humans (think of a team of recruiters or actuaries reviewing each other's cases). Humans can also be self-reflective and self-critical and will more probably be aware of their own biases in a way that machine learning models will not.
- However, once machine learning systems learn biased representations, they can be difficult to detect and mitigate, and can have wide-reaching implications. These effects can be systematic and often invisible or buried in quite subtle ways (e.g. weekly working hours in an HR dataset can be a proxy for gender).
- At best, some systems can recommend users the wrong pair of shoes, at worst, other systems can label individuals as fraudsters or as likely re-offenders because of race or nationality (Refer to Links: Scandals).
- Large-scale biased decision making such as the examples above are known to damage both parties, victims are treated unethically and can suffer life-changing losses, while perpetrators are exposed to large fines and sometimes irreparable reputational damage.
- The advent of Generative AI makes the problem of bias of even worse. GenAI is designed to detect and amplify the patterns present in its training data, so it will naturally also amplify any biases. General solutions to this new problem do not yet exist, so we must be particularly vigilant in this new era.
- From 2025, the EU AI Act will be in force and requiring positive ex-ante conformity assessment before deploying high-risk systems such as credit scoring and candidate recommendation systems. **Ensuring fairness and safety before deploying such systems will be a hard requirement for any business operating in a high-risk domain wanting to leverage AI.**

WHAT?

- Fairness can be ensured at multiple steps in a learning system. We can distinguish between pre-, in-, and post-processing methods which consider wide-ranging strategies to mitigate bias:

  - **Pre-processing:** When training data is thoroughly cleaned by checking for any sampling bias, non-representativeness in the classes (more males than females for instance) and for the inadvertent presence of proxy variables.
  - **In-processing:** When a model's objective function is explicitly modified for bias mitigation with techniques such as adversarial reweighted learning.
  - **Post-processing:** When optimization techniques are used to ensure a model's output satisfies a fairness metric without the need for retraining. Researchers at LinkedIn implemented a fair re-ranker for recommender systems while others have developed a model-agnostic solution for generic tasks.

- For successful bias mitigation, data scientists and machine learning engineers need to choose (depending on the task and risks at hand) some common fairness metrics to optimize for (this list is far from being exhaustive):

  - **Demographic parity** means that a model is fair if the likelihood of a certain prediction is equal for all groups in a demographic feature, like gender. So, if a model predicts loan approval, it has demographic parity when both men and women have the same chance of being given a loan, no matter their gender.
  - **Counterfactual fairness** is when a prediction would stay the same even if we imagine that a person's demographic group was different, holding all else constant. It means that the outcome of a model's decision for an individual should not change if we hypothetically switch their demographic group, like race or gender.
  - **Individual fairness** is the principle that people who are similar according to a certain measure should receive similar predictions from a model. For example, if two individuals have nearly the same qualifications and experience (their "distance" is small), a fair model would give them comparable job success predictions.

- While the processing steps are not mutually exclusive and certain systems can include a combination of these steps in their pipeline, the fairness metrics themselves are mutually exclusive; one cannot successfully optimize for individual fairness and demographic parity at the same time, for instance. This mathematical property requires careful consideration of which metric to optimize for.

RECOMMENDATIONS

- Knowing you have a bias problem is the first step to solving it and avoiding bad decisions. Consider using statistical tools to detect bias in the output of your machine learning models. Regular fairness audits and health checks should be performed once the model is online to ensure concept and data drifts do not perpetuate unfairness.
- Carefully think about which metric to use and why. For certain tasks individual fairness metrics are more relevant while other tasks require group-based metrics.
- Often, the position of the bias mitigation technique in the pipeline can have different effects on the fairness-utility tradeoff. For instance, In-processing techniques will have the highest impact on model performance, while pre-processing techniques will have the lowest performance impact at the cost of fairness guarantees.
- Most importantly, there is no golden rule, biases can leak in at any point in the pipeline and should be mitigated before and regularly during execution. There will always be trade-offs with the model's performance, and it is up to the data scientist and machine learning engineer to fine-tune the model and bias mitigation system to find the sweet spot.

LINKS

- A Snapshot of the Frontiers of Fairness in Machine Learning (ACM)
- Can machine-learning models overcome biased datasets? https://news.mit.edu/2022/machine-learning-biased-data-0221
- Interactive example of fairness metrics in ML.
- Survey on Fairness Notions and Related Tensions.
- Review paper of In-processing fairness.
- In-Processing Modeling Techniques for Machine Learning Fairness: A Survey
- Post-processing fairness papers:
  - Petersen et al.'s generic post-processor
  - Geyik et al's re-ranker for information retrieval tasks (LinkedIn)
- FairLearn, an open-source, community-driven project to help data scientists improve fairness of AI systems.
- Scandals:
  - COMPAS, PredPol, Childcare Benefits Scandal

SUSTAINABILITY
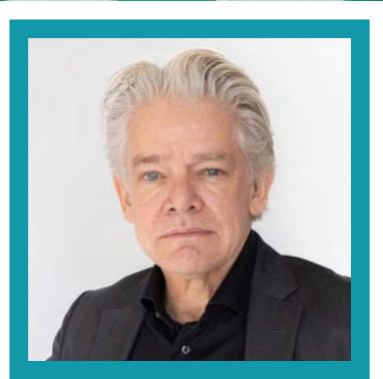
RON TOLIDO
EXPERT IN RESIDENCE

Ai has the potential to solve some of the planet's sustainability issues, but it comes with its own environmental price too.

Artificial intelligence, particularly Generative AI, holds immense promise for accelerating sustainability efforts. Its ability to analyze data, generate solutions, and optimize resources can significantly impact our environmental challenges. However, while these technologies hold great potential, it's essential to recognize their current role in contributing to energy consumption.

As AI models grow in complexity, they demand more computing power, leading to increased energy usage. This exponential growth in energy demand poses a significant sustainability concern. Even after training is complete, ongoing AI operations continue to consume substantial energy. To mitigate these issues, we need to prioritize sustainability from the design stage, rather than relying on future solutions.

WHY?

- **Environmental Impact:** The energy-intensive nature of training and operating AI models, especially larger Generative AI models, contributes to greenhouse gas emissions and exacerbates climate change.
- **Resource Scarcity:** The computational demands of Generative AI models strain the availability of energy resources and computing infrastructure. This is not just electricity, data center cooling systems often place great demand on local water supplies.
- **Ethical Responsibility:** As AI becomes increasingly integrated into society, we have an ethical responsibility to ensure its development and use are sustainable. By addressing the environmental impact of AI, we can uphold our ethical obligations to protect the environment for future generations.

- **Economic Efficiency:** The energy costs associated with Generative AI models can be substantial, impacting operating expenses and overall profitability. By adopting sustainable practices, we can reduce energy costs and improve the economic viability of AI applications.
- **Perception and Trust:** The environmental impact of AI can erode public trust and hinder the adoption of this transformative technology and its potential contribution to finding solutions for environmental issues.

WHAT?

Given the issues above, some strategies to mitigate the environmental impact of Generative AI can be defined:

- **Task specialization.** Building smaller, smarter models that are task-specific instead of general-purpose. Co-pilots for programming do not necessarily have medical knowledge or answer questions about Napoleon.
- **Hardware optimization:** Develop more energy-efficient hardware specifically designed for AI applications. This could include optimizing chip architectures, reducing power consumption during idle periods, and exploring alternative computing paradigms like neuromorphic or analog computing.
- **Model optimization:** Refine Generative AI models to reduce their computational complexity without compromising performance. Techniques such as model compression, pruning, quantization, and knowledge distillation can significantly lower energy consumption while maintaining accuracy.
- **Algorithmic efficiency:** Employ more efficient algorithms for training and operating Generative AI models. This includes exploring novel optimization techniques, utilizing distributed computing frameworks, and leveraging cloud-based infrastructure with renewable energy sources.
- **Alternative architectures:** Whilst most modern Generative AI methods are based on transformer architectures, other architectures can be more efficient for certain types of task.
- **Data optimization:** Minimize the amount of data required for training Generative AI models without sacrificing performance. For example, a naïve hoovering up of training data will include a huge amount of redundancy and

duplication. Sampling more strategically could drastically reduce the required model size.

- **Renewable energy integration:** Power AI infrastructure with renewable energy sources such as solar, wind, and geothermal power. This can significantly reduce the carbon footprint of AI applications and contribute to a more sustainable future.
- **Sustainable infrastructure:** Utilize energy-efficient data centers and implement cooling strategies that minimize energy consumption. Additionally, consider utilizing edge computing devices to reduce data transfer and distribute associated energy usage.
- **Carbon offsetting:** Implement carbon offsetting programs to compensate for the emissions generated from AI applications. This can support renewable energy projects, reforestation initiatives, and other carbon-reducing initiatives.
- **Transparency, monitoring and disclosure:** Promote transparency in AI development and operation by monitoring and disclosing the energy consumption and carbon footprint of Generative AI models. This can encourage developers and users to adopt more sustainable practices.
- **Education and awareness:** Raise awareness among AI developers, users, and stakeholders about the environmental impact of AI applications. This can foster a culture of sustainability and encourage responsible AI development.
- **Policy and regulation:** Implement policies and regulations that incentivize sustainable practices in AI development and deployment. This could include tax breaks for energy-efficient AI infrastructure and carbon pricing mechanisms.

RECOMMENDATIONS

- **Adopt sustainable AI practices:** Implement sustainable practices throughout the development, deployment, and operation of AI systems. Optimize resource utilization, employ energy-efficient hardware, and power AI infrastructure with renewable energy sources.
- **Measure and report AI carbon footprint:** Track and quantify the carbon footprint of AI applications to identify

areas for improvement. Publicly disclose your AI carbon footprint and set ambitious goals for reducing it.

- **Invest in sustainable AI research and development:** Support research and development initiatives that focus on enhancing the energy efficiency and sustainability of AI models and hardware.
- **Collaborate with other organizations:** Partner with other organizations to share best practices, develop sustainable AI standards, and promote collective action on AI sustainability.
- **Educate and empower employees:** Provide employees with training and resources to understand the environmental impact of AI and encourage sustainable practices in their work. Encourage employee involvement in AI sustainability initiatives.
- **Learn from other sustainability initiatives:** Look at successful campaigns for behavior change about other types of energy usage for lessons on how to reduce AI's impact. Ensure that developers understand the impacts that their design decisions will have to reduce accidental overuse.

LINK

- Accelerating Sustainability with AI: A Playbook. Microsoft, November 2023. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/11/Microsoft_Accelerating-Sustainability-with-AI-A-Playbook-1.pdf
- Carbon Footprint Calculator. Used for example to calculate final Carbon footprint when training the BLOOM model tracking the CO2 emissions BLOOM produced in real time over a period of 18 days. https://codecarbon.io
- OECD Principles on Investing in AI Research and Development (principle 2.1), including pathways and instruments: https://oecd.ai/en/dashboards/ai-principles/P10
- Collaboration on developing sustainable AI standards and tools. The Open Group Open Footprint Forum: https://www.opengroup.org/openfootprint-forum
- AI technology application and employee responsibility. Nature.com: https://www.nature.com/articles/s41599-023-01843-3

PRIVACY

RORY POTTER

EXPERT IN RESIDENCE

Privacy is an expectation on the use of data, not a property of the data. The challenge we now face in AI is using this data responsibly whilst matching this expectation.

To train a good machine learning model, we need a lot of data. More and more of the data we want to train a model now is sensitive in some way, either personally to individuals or commercially to companies.

To reap the full benefit of AI using this data we need to navigate the privacy-utility tradeoff: how do we use the data for good while preserving privacy? A new suite of technologies called privacy enhancing technologies enable us to have both.

These technologies allow us to do more with the data we collect. Broader, deeper and more representative data allows us to build more accurate, generalizable and useful. Doing this will be hugely valuable, but doing it means protecting and respecting the privacy of those who share their data with us.

WHY?

- As data and AI skills permeate organizations, it becomes advantageous to share data more widely. The more data that experts can access – and the greater the diversity of people with access to data – the more value that data can bring.

- But we may not want to share that data from fear of legal, reputation or commercial damage. Or worse, we share the data anyway and destroy the reputation of AI and violate millions of people's privacy.

- **Privacy is an expectation people have about how their data is used**. If we want to use their data, we need to satisfy their expectations.

- We are trapped in the privacy-utility trade-off: between being too scared to use valuable data, and being indifferent to the legal and moral risks of using that data.

- This problem is even worse in the new era of large generative models where large amounts of input data can sometimes be encoded verbatim those models. This might lead to situations where memorized training data can be extracted in its original form.

- Deepmind fell foul of this in 2017. It's Streams app, using NHS data, predicted risk of acute kidney injury. But in building the app, un-anonymized medical records of 1.6 million patients were shared with Deepmind. This life saving tool was eventually discontinued because privacy had not been adequately considered when using personal data to build an intelligent AI product.

- New technologies in this space, called privacy enhancing technologies or PETs, enable us to have our cake and eat it: in many cases we can preserve privacy while benefiting from the utility of using sensitive data.

WHAT?

- All privacy enhancing technologies enable us to navigate the privacy-utility trade off. They do this through data minimization, encryption, decentralization and obfuscation.

- Those technologies focused on preserving the privacy of new data being used by the AI system are said to preserve **input privacy**. These technologies include **federated learning,** used for example by Apple and Google to train AI auto-correction models locally on our devices.

- Other techniques use encryption, be that hardware-based using Intel's Trusted Execution Environments or mathematically-based such as **homomorphic encryption** which enables AI run on encrypted data without ever decrypting that data.

- Other technologies which seek to preserve the privacy of the training data are said to preserve **output privacy**. These include synthetic data, which enables us to generate fake data which has similar properties to real data and share it without exposing individual's data.

- The privacy of models themselves is just as important as the privacy of the data. Sophisticated attacks can now extract sensitive data directly from the models, and more subtle "membership attacks" can infer whether a particular person's data was used in the training of a model, which might then leak private information by association.

- Another is **differential privacy,** which obscures the data whilst maintaining all the relevant statistical properties, and has, for example, been used by the US census bureau to share more detailed information in the 2020 Census.

RECOMMENDATIONS

- Privacy preserving technologies should not be an afterthought but a fundamental part of design. Any process that needs to share private data should take a privacy-first approach. Start by thinking about the privacy implications of the data behind the product and bake in the right tools from the start.

- Consider model privacy as it comes to aspects of memorization and anonymization. Evaluate the risk of breaching privacy when using a model in an infrastructure, including the process and systems around it.

- No technique is a silver bullet. Preserving privacy will need

layers of these technologies and careful thought to the right balance for your use case.

- For privacy preserving technologies to work, the data owner needs to apply good data management practices. Since some modelers won't be able to see the data, it is even more important that it is curated so as to handle anonymous queries.

LINKS

- SOLID is an MIT project in the PET space. Led by Tim Berners Lee it aims to radically change the way users share and own their own, privacy sensitive, data on the internet. https://solid.mit.edu

- Three Spanish Hospitals use Privacy Enhancing AI to improve the speed and accuracy of COVID-19 screenings.

- These privacy enhancing technologies are at the forefront of enabling intelligent products without compromising privacy.

- 2023 Privacy Enhancing Technologies report from The Royal Society.

- Deepmind faces legal action over NHS data use (BBC, 2021). https://www.bbc.com/news/technology-58761324

AUTHORS

EDITORS IN CHIEF
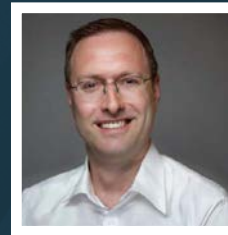
Mark Roberts

CTO, Deputy Head Generative AI Lab
Hybrid Intelligence, Capgemini Engineering
mark.roberts@capgemini.com

Robert Engels

CTO AI, Head Generative AI Lab
Insights & Data, Capgemini
robert.engels@capgemini.com

EXPERTS IN RESIDENCE

Mitali Agrawal
Sogeti, part of Capgemini
mitali.agrawal@sogeti.com

Johan Müllern-Aspengren
Capgemini
johan.mullern-aspegren@capgemini.com

Adam Mehdi Arafan
Sogeti, part of Capgemini
adam.arafan@sogeti.com

Tijana Nikolić
Sogeti, part of Capgemini
tia.nikolic@sogeti.com

Weiwei Feng
Insights & Data Capgemini
weiwei.feng@capgemini.com

Rory Potter
Hybrid Intelligence, Capgemini Engineering
rory.potter@capgemini.com

Niharika Kalvagunta
Capgemini
niharika.kalvagunta@capgemini.com
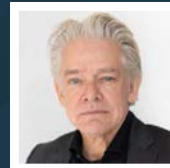
Ron Tolido
Insights & Data Capgemini
ron.tolido@capgemini.com

Ben Mathiesen
Hybrid Intelligence, Capgemini Engineering
benjamin.mathiesen2@capgemini.com

CONFI DENCE IN AI

About
Capgemini

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided every day by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of nearly 350,000 team members in more than 50 countries. With its strong 55-year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fueled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering, and platforms. The Group reported in 2022 global revenues of €22 billion.

Get the Future You Want | www.capgemini.com