Deloitte.

Pairing with PairD

Lessons from scaling our Generative AI platform across Deloitte UK and Europe

April 2024

Contents

Preface

In October 2023 Deloitte UK rolled out PairD (pronounced “Paired”), our first in-house productionised and scaled Generative AI platform that was released to 25,000 UK-based colleagues, and more recently to 75,000 colleagues across Europe.

This represented a significant step forward for Deloitte UK, moving the firm from mainly Generative AI strategy, use case identification and proof-of-concept (PoC) development, to productionised and scaled enterprise deployment across multiple legal jurisdictions.

The difficulties of productionising and scaling AI are well-known, but the specific *generative*[1] capabilities of more recent AI models undoubtedly added new challenges and opportunities. For Deloitte, PairD was as much about accelerating access to high performing and secure Generative AI for our colleagues as it was rapidly learning how to productionise and scale this new technology effectively within large and complex organisations.

This report captures some of the key lessons learned from the initial development and release of PairD. It is not intended as a playbook or how-to guide, but rather as a reference for individuals and teams looking to develop and scale their own Generative AI platforms within large organisations.

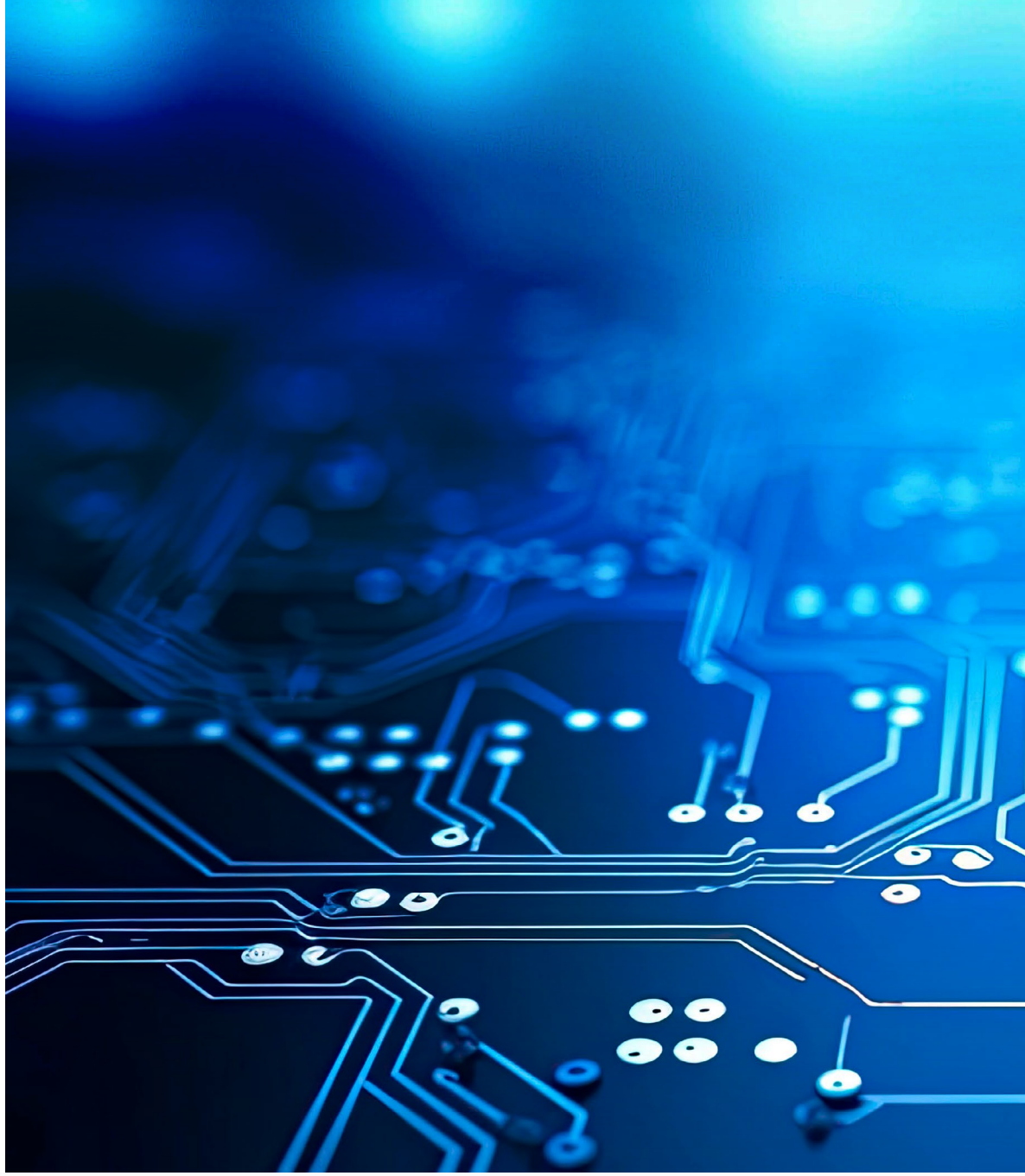
[1] This report assumes some familiarity with Generative AI, which we broadly define as AI that generates new information and data, and contrasts with more traditional AI that focuses on data analysis and insights.

The report is divided into the following sections that reflect key areas of development:

- What is PairD and why did we build it?
- Lessons learned: project and product management
- Lessons learned: AI and cloud engineering
- Lessons learned: security and risk
- Lessons learned: people, organisation, operations and adoption.

Generative AI is an extremely fast-moving field, and we expect the findings of this report to evolve as the technology and wider ecosystem mature, and as we ourselves continue to develop PairD. Looking ahead, we hope to publish updated versions that capture additional insights and lessons from more recent development.

If you would like to learn more about PairD and scaling Generative AI within commercial organisations, please reach out to the authors of this report and to the *Deloitte AI Institute UK*[2].



[2] mailto: UKDeloitteAIInstitute@deloitte.co.uk

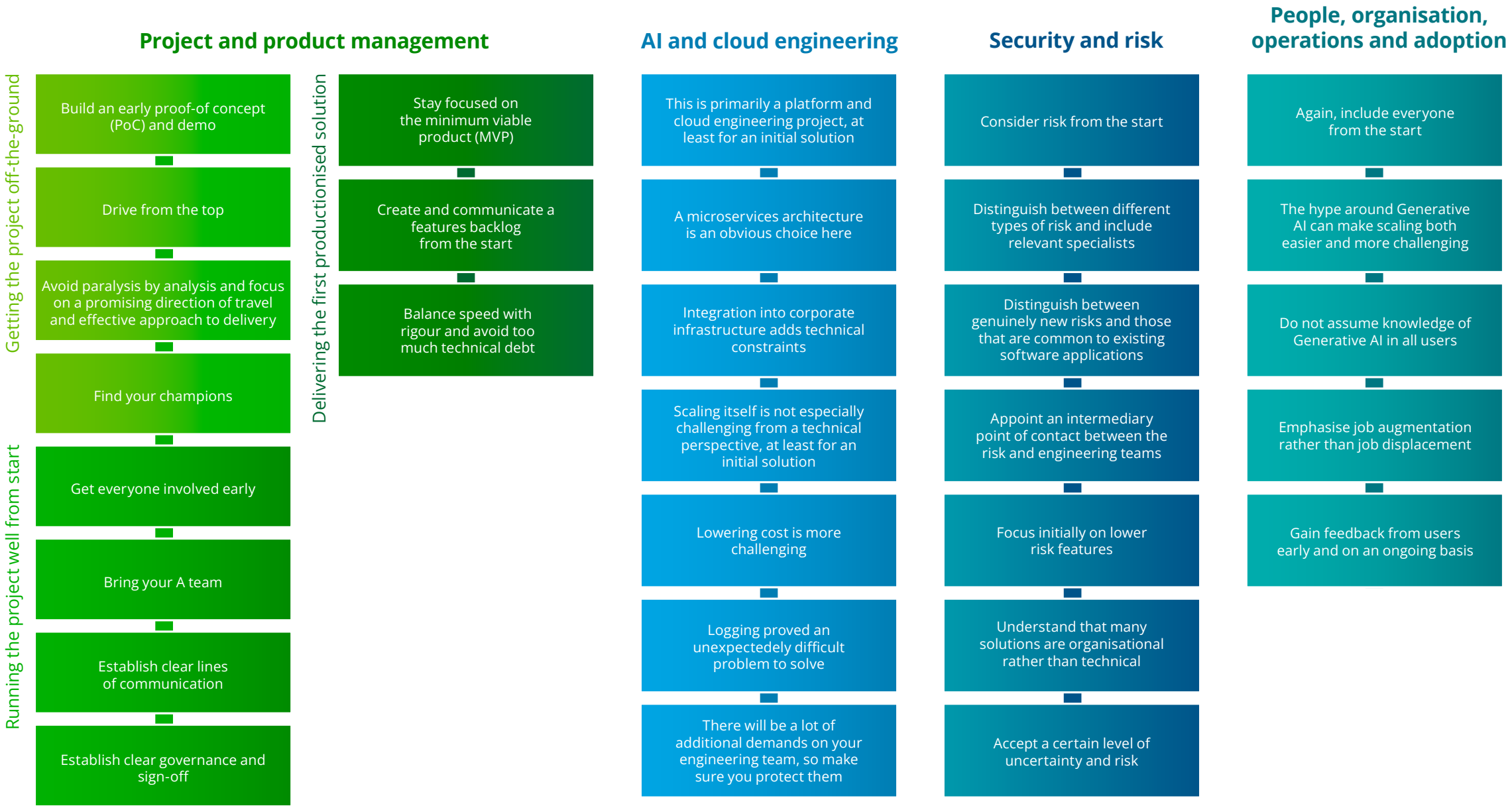


Figure 1 Summary of lessons learned from developing PairD, which are explored in more detail in this report

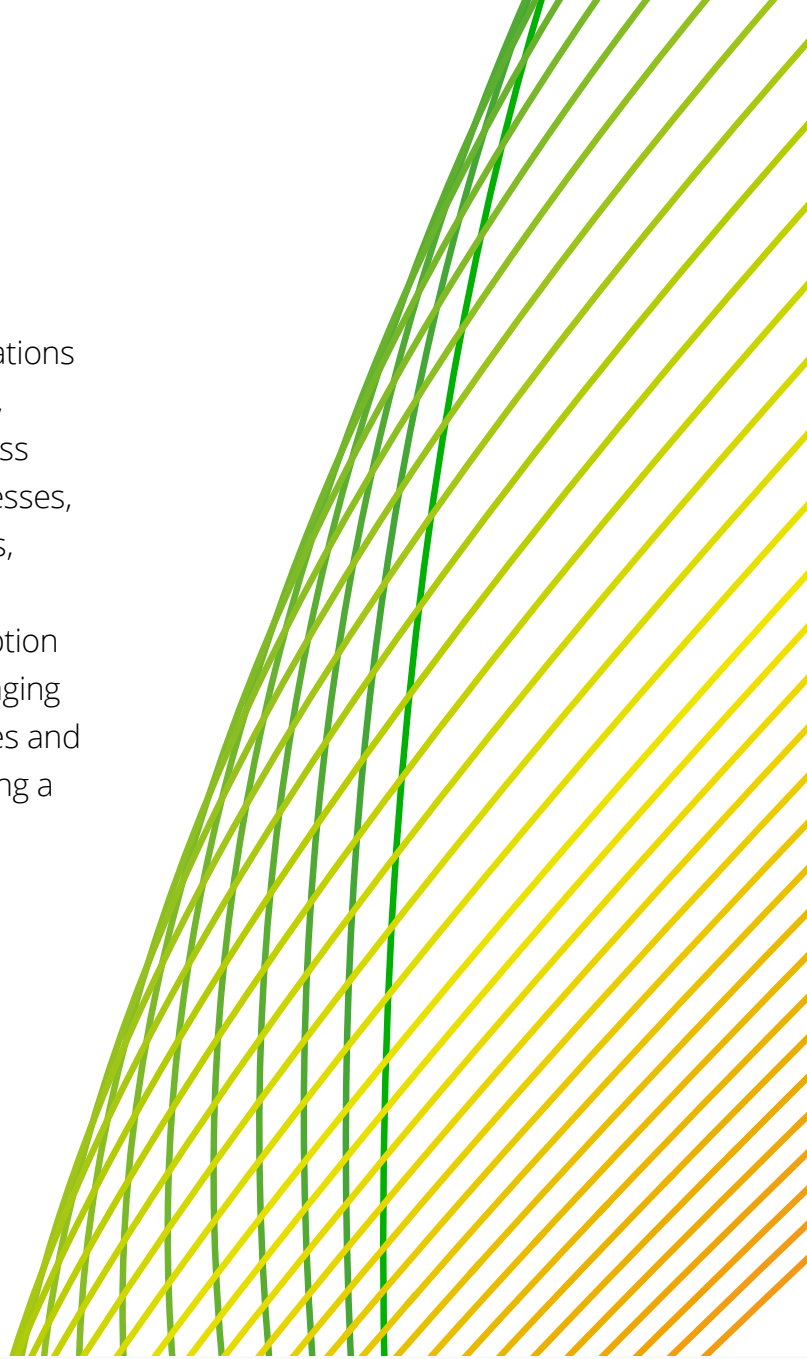
What is PairD and why did we build it?

Generative AI is already transforming businesses and society.

We are experiencing this first hand not only through our own personal uses of Generative AI, but in Deloitte through Generative AI products that are already enhancing our daily work, and through our engagements with clients across industry to help them understand, test and scale Generative AI.

This rapid adoption is driven in part by clear and obvious value. Generative AI not only feels useful out-the-box, it is useful out-the-box, and when used in the right way, it provides immediate, tangible benefits.

However, it is also difficult to achieve commercial value from scaling Generative AI within large and complex organisations. There are a multitude of reasons for this that include – but are by no means limited to – challenges in ensuring data privacy and security, managing upfront and ongoing financial costs, delivering technical integrations into existing complex IT systems, implementing changes to business models and organisational processes, navigating buy vs. build decisions, evaluating models for complex commercial tasks, ensuring adoption across diverse workforces, managing ethical implications for employees and their roles, and ultimately ensuring a return on investment (ROI).

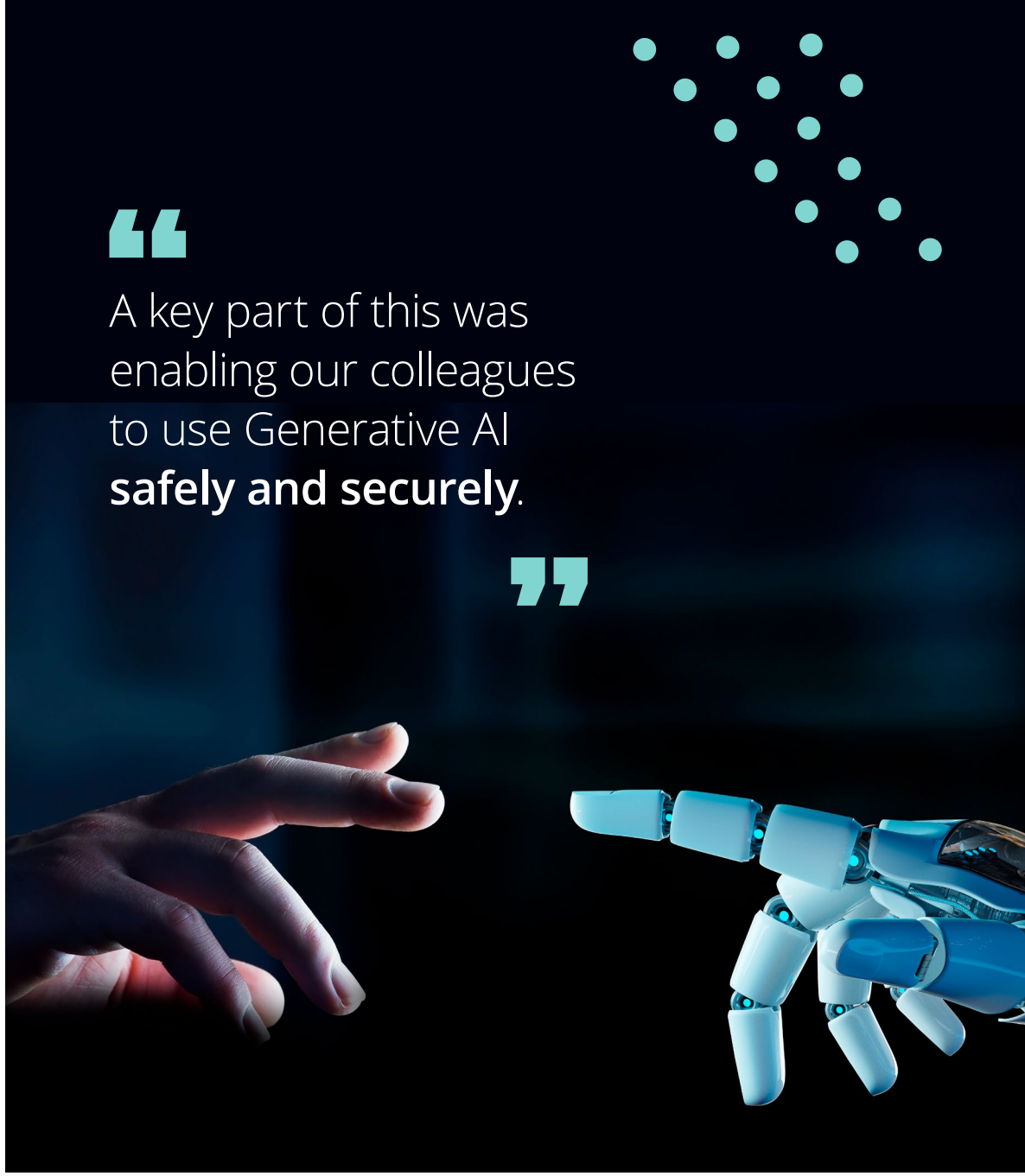


Why did we build PairD?

We built PairD to start tackling first hand some of these challenges, and in doing so, deliver commercial value from scaling Generative AI.

Our top priority was **democratising Generative AI**. We wanted to ensure that the technology could be used quickly and safely by our Deloitte colleagues, and not just those that already use Generative AI, but those that are less familiar with and perhaps more sceptical about its value. We wanted individuals and teams from across the firm to start accessing and using Generative AI in their daily work, and to experience for themselves the productivity and performance gains.

A key part of this was enabling our colleagues to use Generative AI **safely and securely**. Deloitte generates and uses a huge amount of commercially sensitive data, and we wanted our colleagues to be able to use this data with Generative AI whilst simultaneously protecting Deloitte's systems, data and intellectual property (IP). Previous research by Deloitte suggests that many people – over four million in the UK as of September 2023[3] – already use Generative AI at work, but likely in unsanctioned applications that pose a multitude of risks. Through PairD, we wanted to mitigate these risks early whilst empowering our workforce to use Generative AI where appropriate.



[3] https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/technology-media-telecommunications/deloitte-uk-digital-consumer-2023-ai.pdf p. 7

We also wanted to ensure that Generative AI is **accessible and easy-to-use for everyone**, and that it is tailored to our own specific needs within the firm. PairD has a **customised frontend** that includes bespoke features such as document upload and prebuilt prompts, and we are also able to **customise the backend** to increase performance on Deloitte-specific tasks, including using multiple large language models (LLMs), customising models via techniques such as retrieval augmented generation (RAG) and fine-tuning, adding more advanced prompt engineering and model orchestration, and incorporating feedback and metrics from users to enable continuous improvement.

PairD also provides greater **control over cost**, including balancing cost against other priorities such as model performance. This not only includes financial cost, but other costs that are important, including cost to the environment.

Finally, PairD enables the firm to **continually learn** about Generative AI. This is an extremely fast-moving field, and new technologies are being released on an almost daily basis. For Deloitte, it is essential that we remain at the forefront of the field, and this is achieved not only through research, but through hands on technical development and engineering. PairD provides a secure platform that we can use to continually develop, test and scale new Generative AI technologies and features, and not just in a sandbox environment but in a scaled enterprise application.

So, what exactly is PairD?

PairD is a **productionised, secure and private Generative AI platform** that we designed and built for large enterprises such as Deloitte. It uses a **cloud-native microservices architecture** that integrates a wide range of underlying front and backend features that add significant functionality to native LLM cloud services,[4] and that provides a modular, scalable and secure software application that is available to everyday users within the firm.

PairD has a **custom frontend** that allows users to access cutting edge LLMs in a safe and secure manner. Initially, this frontend looked similar to other consumer-facing LLM applications, but we continue to tailor the design and add new features that are specific to our own use cases, including moving to a **mobile-first design**, adding document upload and customised workflows, and integrating additional ways of interacting with backend models, for example through speech.

The backend architecture connects **multiple LLMs**, and produces responses to user prompts whilst balancing performance, cost and environmental impact. The modular design means we can update the backend LLMs as new models are released, including adding customised LLMs that have been fine-tuned on our own data.

The backend architecture also **integrates proprietary data** into the platform, such as Deloitte's internal documents and databases. This is used to augment outputs through techniques such as RAG.

[4] Microservices is an approach to software development that designs applications as a collection of small independent services. Each service within a microservices architecture is self-contained and performs a specific business function. These services communicate with each other over a network, usually through Application Programming Interface (API) calls. This approach is different from traditional monolithic architectures where all components are tightly integrated and deployed as a single unit.

There are also several **intermediary layers** that improve performance, efficiency and transparency. These include:

- Prompt augmentation and orchestration, which combines additional text into user prompts to improve performance and tailor outputs to commercial use cases.
- Logging and analytics, which records anonymised user inputs, and can be used to understand the use of PairD and its performance, which feeds into iterative development.
- Model routing, which sends prompts to the most appropriate Generative AI model.
- Infrastructure optimisation, which improves inference speed and cost.
- Monitoring and dashboarding, which provides ongoing insights into metrics such as usage, cost and carbon footprint.

These layers are integrated into a wider cloud architecture that provides additional enterprise-grade features, including cybersecurity, identity and access management (IAM), cost management, backup creation, continuous integration and deployment (CI/CD), and more.

Finally, PairD as a technical platform is supported by a wide range of additional **assets, processes and teams** including:

- ongoing product management that defines, prioritises and oversees the development of new features
- risk, security and compliance processes that test and approve changes to the product
- technical support that assists users with queries and errors
- documentation, FAQs and educational material
- internal and external communications.

PairD is **continually evolving** and will be upgraded on an ongoing basis as we collect more feedback, tackle more specific use cases, and as the underlying technologies evolve. It is an ongoing journey where we test, learn, iterate and repeat, and ultimately move towards a better platform and a more nuanced understanding of scaling Generative AI.

The remainder of this report explores in more detail some of the key lessons learned from the earlier stages of developing PairD, including both technical and organisational aspects of scaling. We hope that these insights help other teams to develop and scale Generative AI successfully within their own organisations.

Lessons learned: project and product management

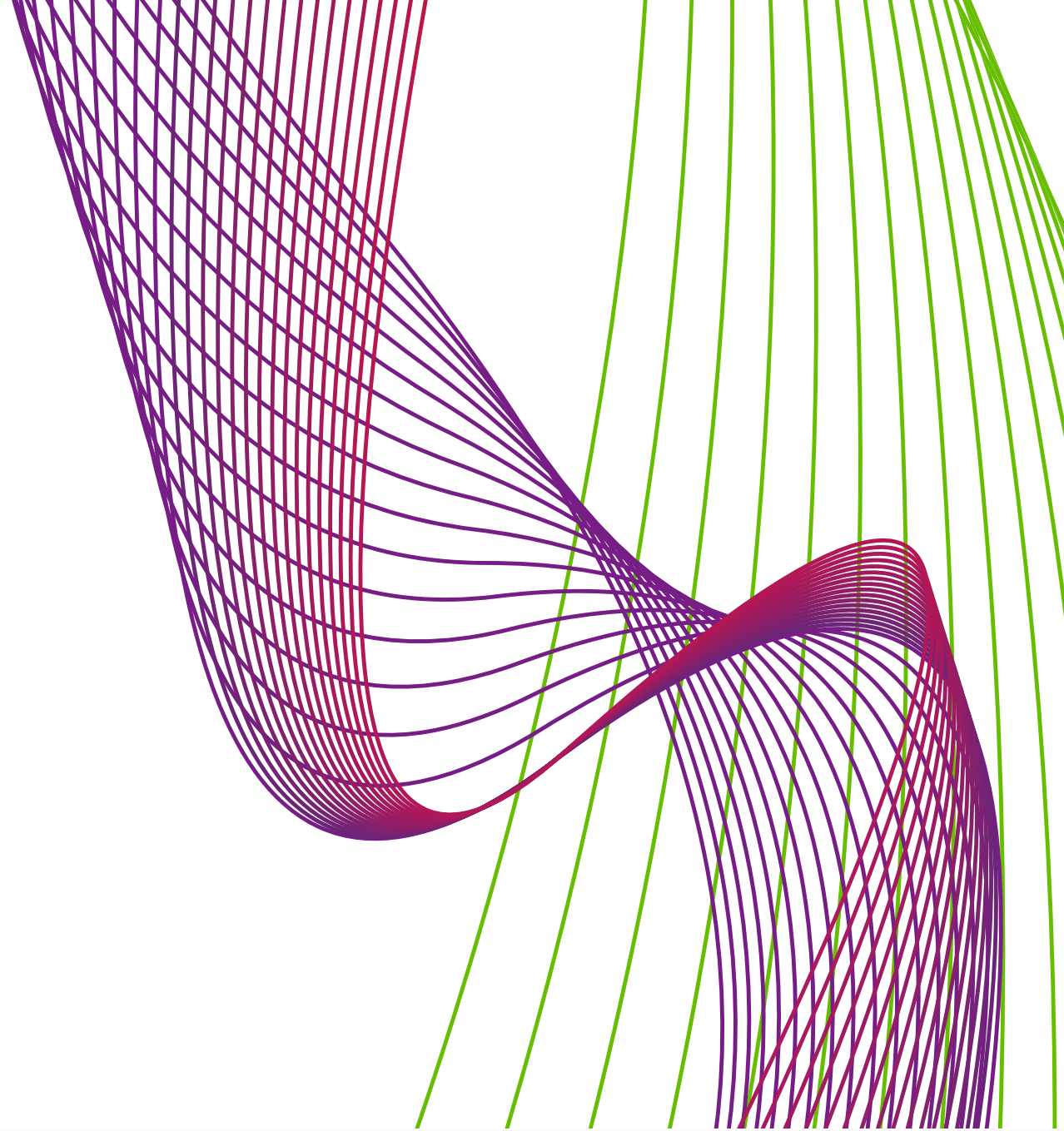
Getting the project off the ground

It is not quick, easy or cheap to build and deploy platforms such as PairD successfully. It requires in-demand and expensive resources and skills, and for large organisations such as Deloitte, there are additional technical, legal, security, risk and compliance considerations that must be overcome before Generative AI platforms can be deployed and scaled.

Collectively, this means it is often difficult to get platforms such as PairD off the ground, even with the current level of interest in the technology.

Early on, we faced similar challenges in Deloitte, and rightly so. Investments in Generative AI – especially on this scale – should not be made lightly, and should be considered carefully against alternative options and the likely ROI.

From our experience, several factors proved important when moving PairD from an initial idea to a fully funded project.



1 Build an early PoC and demo:

Before fully investing in PairD, we worked initially on an early proof-of-concept (PoC) using limited existing resources. This allowed the team to develop a more in-depth understanding of the feasibility and value of the platform, which in turn helped build a more credible and nuanced business case. This also helped engage senior stakeholders, as we were able to support conversations with better evidence, experience, and a tangible demo.

2 Drive from the top:

Innovation is often driven bottom-up within organisations by tech-savvy individuals and teams, but platforms at this scale need to be driven from the top by senior leaders with a clear vision. This is because the level of investment, time, complexity and risk involved in productionising and scaling Generative AI across an organisation is far greater than building smaller demos, PoCs and isolated tools.

3 Avoid paralysis by analysis and focus on a promising direction of travel and effective approach to delivery:

Generative AI is still a new technology, and there are inherent uncertainties and risks in developing this type of platform. Many of the benefits of PairD are intangible and difficult to measure, and from our experience, this can make it difficult to develop more traditional business cases that focus on a financial ROI. Instead, we focused on articulating the overall value of the platform (e.g., democratising AI, driving adoption, reducing data risks etc.) whilst also emphasising our approach to managing key uncertainties, risks and costs, including using agile and lean platform development, delivering value early and quickly, and focusing initially on a lower risk minimum viable product (MVP) with fewer features and fewer commercial barriers.

4 Find your champions:

One of the challenges with PairD was justifying the platform against numerous seemingly similar consumer and enterprise-grade Generative AI products. Whilst there are several advantages in building platforms such as PairD internally, including improved privacy, customisation and security, these advantages are more nuanced and are not necessarily obvious to all stakeholders. To help secure funding and adoption, we found it useful to identify and engage with champions across the firm who understood the differentiated value of PairD, and could advocate for the platform both to senior stakeholders and to the wider business.

Running the project well from the start

Whilst PairD is our first enterprise Generative AI platform, it is certainly not the first AI platform that we have developed and scaled as a firm.

Deloitte has decades of experience building both internal AI platforms and AI platforms for our clients, and it is important to recognise that, in many ways, PairD is just another software application, albeit with a specific Generative AI flavour. Many of the best practices that apply to normal platform development are equally important here too, and it is worth emphasising that, from our experience, the newness and hype around Generative AI can sometimes distract from something that we are otherwise very experienced in delivering.

That said, Generative AI does create some additional delivery challenges, and it is important to establish the right team and approach from the start to ensure downstream success. From our experience, we found the following lessons especially useful.

1 Get everyone involved early:

Building platforms such as PairD involves much more than engineering, especially in large organisations such as Deloitte where there are established processes, stakeholders and constraints. For many teams involved in PairD, Generative AI was also relatively new and unfamiliar, and the first substantial Generative AI project that they had worked on at this scale. This understandably increased complexity and timelines, as both individuals and teams needed to upskill and take a cautious approach to delivery. One of the things that we learned quickly was the importance of involving everyone early, including risk, compliance, human resources (HR), communications, technical support, and others. This gives teams more time to work through specific challenges, and helps identify future risks earlier on in the project.

2 Bring your A team:

Platforms such as PairD are challenging to build and important to get right. There is a lot of interest and excitement around Generative AI, which means there is substantial pressure and attention on delivery. At the same time, Generative AI is also new and rapidly evolving, which makes delivery more complex and challenging. To help manage these pressures and risks, we pulled together an especially experienced team across the full range of specialisations, from cloud infrastructure and ML engineering to legal, security, compliance, ethics and risk. This ensured that individuals and teams could focus on specific Generative AI challenges within PairD, and also provided senior stakeholders with confidence that we had the right team working on delivery.

3 Establish clear lines of communication:

Delivery at scale in large organisations necessarily involves many stakeholders. To help improve the pace and quality of delivery, we set up quick, simple and clear lines of communication across all stakeholders, at times breaking existing protocols to ensure quicker and more direct communication. As PairD has matured, we have been able to reintroduce some of the more established ways of working so that teams can return to business-as-usual (BAU), but during the earlier stages of delivery, we found more direct lines of communication were especially useful.

4 Establish clear governance and sign-off:

The inherent novelty of Generative AI means that not all risks are immediately understood, and throughout the development of PairD, a certain level of risk and uncertainty had to be accepted to progress delivery. For example, it is currently not possible to fully prevent the platform from outputting some factually incorrect information (i.e., "hallucinating"), but this risk can be mitigated through a combination of technical and behavioural solutions. When solving these types of challenges, it was important to engage directly with individuals and teams that are directly responsible for managing risk, and that are empowered to make decisions when some level of risk acceptance is needed. This accelerates development by avoiding back-and-forth between teams that are otherwise not able to accept additional risk themselves.

Delivering the first productionised solution

Once a project is off the ground and technical development has started, moving from PoC to a fully productionised solution remains challenging, especially when new technologies such as Generative AI are involved.

Again, many of the lessons in this section are common to any software platform, albeit with additional Generative AI-specific nuances. From our experience building PairD, there are several lessons that we found especially useful during this stage of development.

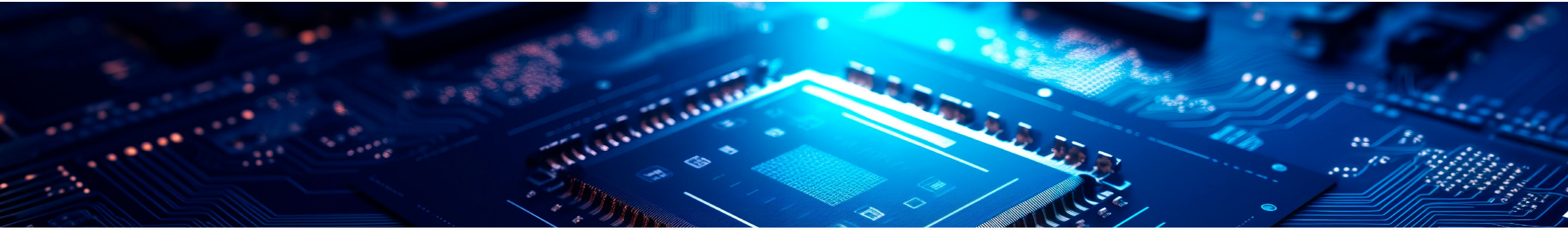
“From our experience building PairD, there are several lessons that we found especially useful during this stage of development.”

1 Stay focused on the MVP:

Teams are understandably keen to ensure that platforms such as PairD are valuable to their own areas of the business. If not careful, this can quickly balloon the scope of delivery, with additional features added to cater for an increasingly wide range of requirements. To mitigate this risk, it is important to remain focused on the simplest initial version of the platform – the MVP – and have in place effective product owners that can manage difficult and often competing demands, including from senior stakeholders.

2 Create and communicate a features backlog from the start:

The previous challenge can be further mitigated by creating, maintaining and communicating a clear features backlog from the start. This helps stakeholders who are external to the project (i.e., those not directly involved in delivery) to understand for themselves if and when their desired features will be included in the platform. This also helps manage competing expectations and requests, and enables the wider business to plan more effectively.



3 Balance speed with rigour and avoid too much technical debt:

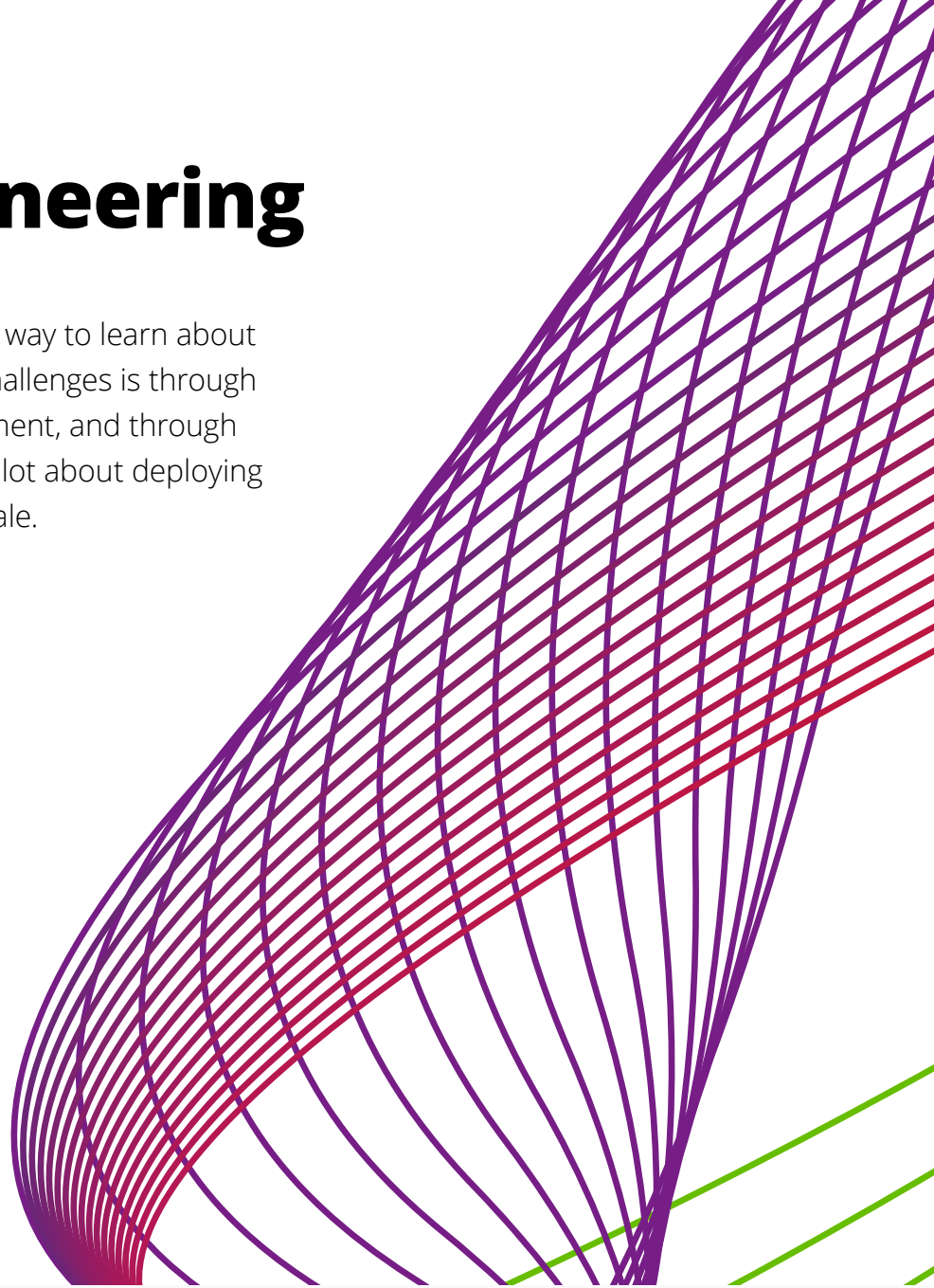
Pressure to deliver can be high with Generative AI, as organisations are keen to ensure they leverage this new technology quickly and do not fall behind competitors. This can increase pressure to cut corners and release platforms early, but this can also introduce technical debt that damages adoption and introduces higher future costs. This is a common trade-off with any platform, but it is particularly acute with Generative AI because of the high levels of interest and the pace of development in the field. From our experience, it is more important to build platforms such as PairD in a robust, flexible, modular and scalable manner from the start, as the pace of development means that new features and changes will inevitably need to be added to the platform in the near future.

Lessons learned: AI and cloud engineering

PairD is much more than an engineering project, but engineering is nonetheless a core part of delivery. After all, PairD is ultimately a technical platform, and not just an MVP but a scaled and flexible solution that is the foundation for more advanced future versions.

When building PairD, we were able to quickly assemble a team of highly skilled and experienced cloud and AI engineers. However, even for experienced professionals, Generative AI in its current form is relatively new, which means that everyone is still learning, and that individuals and teams need to tackle a range of new and uncertain challenges. For PairD this included, for example, using new and sometimes poorly documented and buggy solutions, using novel AI models and frameworks, and delivering work that can quickly become outdated as new models, products and insights are released.

As always, the best way to learn about new engineering challenges is through hands-on development, and through PairD we learned a lot about deploying Generative AI at scale.



1 This is primarily a platform and cloud engineering project, at least for an initial solution:

For many tasks, it is possible to achieve reasonable or even high levels of performance using existing Generative AI models and APIs, and further improvements can be achieved using relatively simple techniques such as few-shot learning and prompt engineering. For an initial solution, the performance of state-of-the-art LLMs and Generative AI models is often sufficient, which means engineering teams can focus on building the wider platform. This contrasts with previous versions of AI, in particular pre-foundation models, where a lot of upfront effort was typically needed to reach sufficient model performance.

2 A microservices architecture is an obvious choice:

This report does not delve too deeply into technical detail, but it is worth highlighting that a microservices architecture is an obvious and effective choice for platforms such as PairD, especially when cloud deployment is an option. Cloud environments are designed for both flexibility and scaling, and are well suited to Generative AI solutions that are not only scaled across enterprises, but where additional models, features and services will undoubtedly be added in the future.

3 Integration into corporate infrastructure adds technical constraints:

When our team first started designing a PoC for PairD, they focused on the most effective architecture within a general cloud environment. However, this initial design needed to be adapted when it became clear that PairD would scale into Deloitte as a fully productionised solution, as Deloitte's IT services have various design requirements for any software that is used internally. To manage this risk early, it is important to include engineers and solution architects that are familiar with internal corporate infrastructure during the initial phases of development, as this ensures solutions are compliant by design.

4 Scaling itself is not especially challenging from a technical perspective, at least for an initial solution:

Cloud services are designed to scale and can easily support the number of users that currently have access to PairD. This includes the backend LLMs, which scale by purchasing greater bandwidth to third party APIs and by increasing the number of instances that host local models. So far, we have not experienced any significant challenges in scaling up or down the number of inference requests to our backend LLMs, although we do expect this to become more challenging as the platform is used by additional geographies and as more features are added, especially those that incorporate larger unstructured datasets and that allow users to input significantly longer prompts.

5 Reducing cost is more challenging:

Technical scaling itself is relatively straightforward, but balancing cost against performance is more challenging. Whilst it is simpler to send all prompts to the highest performing models, this becomes expensive when scaled to thousands of users. For many tasks, a smaller model is often sufficient and much cheaper, in particular when using open-source. However, developing backend algorithms that effectively balance the competing demands of performance and cost is more challenging, especially as users also have access to high performing consumer Generative AI applications that are always a tempting alternative. We are currently exploring both technical and organisational solutions, including optimising model routing and using different cost models that allow users to access higher performing models by sharing some of the additional cost.

6 Logging proved an unexpectedly difficult problem to solve:

Logging prompts and responses is an important component of an enterprise LLM solution, both to analyse feedback from users and to monitor use of the tool for audit and compliance purposes. However, logs also contain a variety of sensitive data that must be visible only to specific individuals. There are several available products and services that aim to provide secure data logging for LLM applications, but we found that some of these options contained known bugs, limited documentation, and/or did not provide the level of security and controls that we need for an enterprise platform. We tested several options before finding an appropriate solution, and at the same time developed additional policies and communication around data storage, retention and processing.

7 There will be a lot of additional demands on your engineering team, so make sure you protect them:

Many of the teams involved in PairD needed a relatively detailed understanding of the system-level design and underlying models. Much of this expertise lies within engineering teams, which is where most questions and requests were initially re-routed. If not careful, engineering teams can quickly become overloaded with a large number of requests that can end up threatening the delivery of the project. It is important that product owners, managers and senior leaders protect engineering teams where possible, and relay only essential and consolidated questions. Where possible, we also found it useful to assign a separate technical AI specialist to support non-engineering teams. This individual should ideally be someone with strong communication skills, and with an understanding of both technical Generative AI engineering and wider platform development.

> There will be a lot of additional demands on your engineering team, so make sure you protect them.

Lessons learned: security and risk

One of the main reasons for building PairD was to provide additional security, privacy and transparency when using sensitive commercial data in conjunction with Generative AI.

For Deloitte, our data is one of our most valuable assets, and it is important that we know exactly where and how our data is stored and used within any software application. We also need to comply with our own internal policies and our contracts with clients, and ensure that any software application is secure from external attack and used in a safe and responsible manner.

These constraints are common for any software in Deloitte, but they are especially pertinent to platforms such as PairD where there are additional Generative AI-specific risks, including data leakage, adversarial attacks, misuse, hallucinations, biases, limited transparency, legal uncertainty, and regulatory changes.

We are fortunate in Deloitte to have over a decade of experience managing AI risk within commercial organisations, and have deep expertise across a wide range of AI risks, including those outlined in our Trustworthy AI Framework.[5]

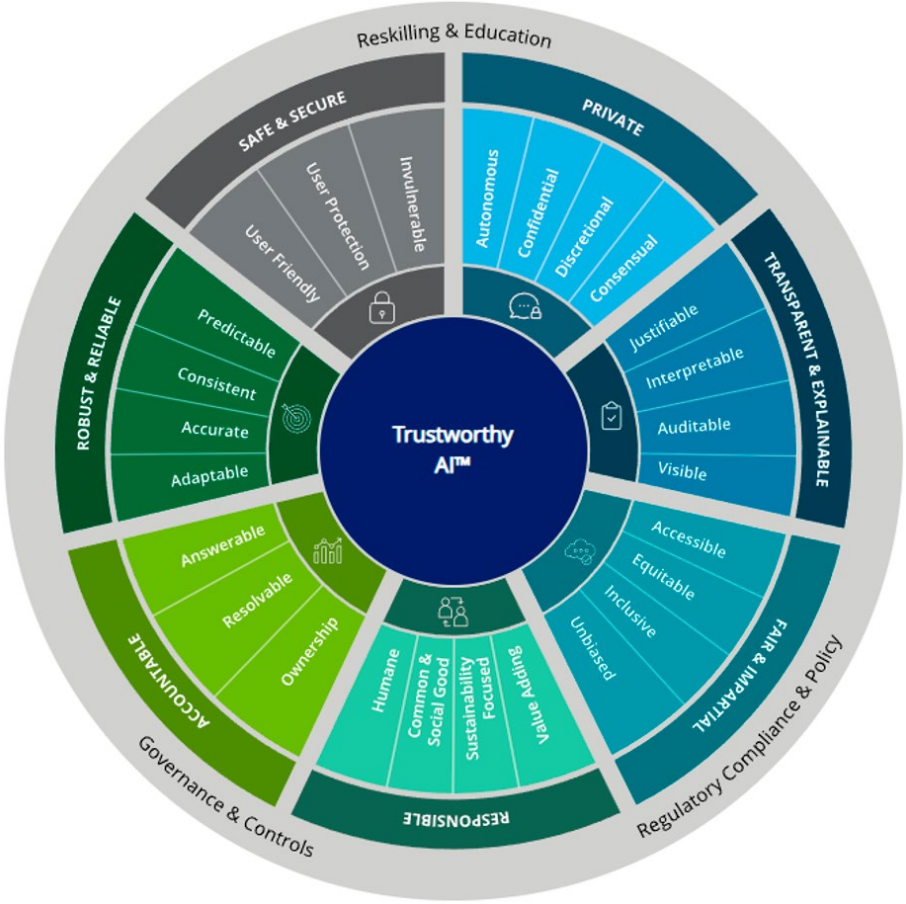


Figure 2 Deloitte's Trustworthy AI Framework

[5] https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html

We used our Trustworthy AI Framework to help structure our approach to PairD, and found that there are several specific lessons when developing and scaling novel Generative AI platforms.

1 Consider risk from the start:

It is important to include individuals and teams focused on risk as early as possible when building platforms such as PairD, where appropriate from separate legal jurisdictions. This provides more time to work through novel challenges and risks, and also helps ensure any initial concerns are flagged early and resolved during the development process. Platforms such as PairD should be secure by design, as this helps reduce downstream risk and technical debt, and also accelerates the development of future versions of the platform.

2 Distinguish between different types of risk and include relevant specialists:

Generative AI incorporates several different types of risk, including data privacy, security, IP, legal, regulatory, usage and ethical risks. These are all distinct and complex areas, and each has their own additional Generative AI-specific considerations. When building PairD, we included separate deep specialists across each of these areas. This helped accelerate delivery, and also gave confidence to senior stakeholders that any decisions were measured, especially those that required a degree of risk acceptance.

3 Distinguish between genuinely new risks and those that are common to existing software applications:

It is easy to get caught in the hype of Generative AI and the complexity of the underlying technology, and in doing so lose sight of the risks that are genuinely new and uncertain. For example, there is currently a lot of focus on both hallucinations and a lack of transparency in Generative AI models. These are both genuine challenges and risks, but they are not entirely new or unique. Other statistical and predictive models can also output errors which can be difficult to interpret, especially when purchased as closed source third party software. Similarly, data leakage is a known problem with some Generative AI products, but many companies – including Deloitte – regularly input sensitive data into a wide range of other existing third party products, providing there are sufficient and demonstrable protections in place. With Generative AI platforms such as PairD, it is important to identify and focus on risks that are genuinely new and uncertain, as these risks can be more challenging to resolve. Other risks of course remain important, but they are more likely to be covered by existing policies and mitigations.

4 Appoint an intermediary point of contact between risk and engineering teams:

Both risk and engineering teams have the same objective: to release safe, valuable platforms. However, these teams can sometimes seem at odds, with risk teams wanting to move more cautiously and engineering teams wanting to move faster and iterate on deployment. Sometimes when these teams talk directly to each other, the conversations can be counterproductive due to differing viewpoints and ways of working. Whilst there are always feasible solutions and compromises, we found it can be more effective to use an intermediary point of contact that understands the separate teams and can help find a suitable middle ground.

5 Focus initially on lower risk features:

There are lots of risks associated with Generative AI, but not all Generative AI platforms are the same, and not all have the same level of risk. For example, different types of input data have different levels of risks, with some types of confidential data less risky than others. Rather than considering all possible risks upfront, it can be more effective to focus initial development on lower risk features that avoid getting caught in some of the more complex and trickier challenges around risks. This helps develop and release an initial platform quicker and with fewer barriers, which reduces time to value and enables iterative deployment and learning.

6 Understand that many solutions are organisational rather than technical:

Some of the biggest risks with Generative AI do not currently have effective technical solutions, or have solutions that are only partially effective. These include detecting inappropriate input data, detecting output errors, improving transparency and interpretability, and ensuring outputs are not misused. As with other software applications, some of the most effective ways to minimise these risks are through organisational approaches, such as effective communication, education and training, and clear accountability if a platform is misused, especially for more serious offences.

7 Accept a certain level of uncertainty and risk:

Ultimately, developing PairD involves a degree of inherent risk. This is a new technology, and one that is evolving extremely quickly. When building a Generative AI platform, a certain level of risk acceptance is needed to move beyond PoCs and into productionised, scaled and integrated solutions. These risks should be considered carefully against the competing risks of not building platforms, and instead waiting for the wider ecosystem to mature.

Lessons learned: people, organisation, operations and adoption

PairD is a technical Generative AI platform, but for it to provide an ROI, it needs to be adopted by employees across the firm and add value to their roles.

This is a common challenge when deploying any technical platform, and is one of the reasons why effective change management is critical during any technology integration. By addressing individual concerns, communicating transparently, providing learning opportunities, and ultimately taking a considered and compassionate people-centric approach, organisations can build a positive environment around platforms such as PairD, and empower individuals and teams to embrace new technologies and ultimately enhance their own productivity and performance.

That said, change management can be especially challenging in large organisations such as Deloitte where there are a multitude of existing structures, processes and policies that help firms operate more effectively at scale, but at times can also create additional barriers that make it harder to introduce disruptive technologies such as PairD. Deloitte also has a large and diverse workforce, which is undoubtedly a key strength, but can also make scaling more complex as there is no one-size-fits-all approach, and scaling needs to be flexible, considerate and tailored to a wide range of individual needs.

Through rolling out PairD to over 75,000 Deloitte colleagues across Europe, we learned several lessons that are specific to integrating Generative AI within larger organisations.

1 Include everyone from the start:

This point has been made previously, but it is worth restating. When building platforms such as PairD, it is easy to prioritise technical engineering, and not include teams that focus on organisational aspects of scaling until later in the development process. However, this reduces the amount of time available for non-engineering teams to work on their respective areas, and risks failing to incorporate important people-centric considerations from the start. For example, visual design is an important aspect of PairD, as the platform not only needs to appeal to its intended users, but also reflect Deloitte's brand and identity. Whilst experienced frontend developers in Deloitte already have some understanding of the firm's unique branding, this expertise lies predominantly within brand and marketing teams. This is similarly the case in other areas, including communications, education and training, all of which should be considered early by including relevant teams from the start.

2 The hype around Generative AI makes scaling both easier and more challenging:

Many people have already used Generative AI.[6] This can make it easier to scale platforms such as PairD, as many users already have some understanding of platform features and value. However, this can also make scaling more challenging, as users compare PairD against an increasing range of consumer products. Some common questions that we faced early include: Why should I use PairD when I can use X consumer product? Or why does PairD look different from X consumer product?

These questions can be resolved early by communicating the business rationale for platforms such as PairD, such as improved privacy, security and customisation. We also found it useful to communicate our wider product roadmap across Deloitte, as this helps users to understand that the initial version of PairD only captures a small portion of the overall potential value.

[6] See, for example, https://www2.deloitte.com/uk/en/pages/press-releases/articles/more-than-four-million-people-in-the-uk-have-used-Generative-ai-for-work-deloitte.html

3 Do not assume knowledge of Generative AI in all users:

Whilst many people have used consumer Generative AI products, this is not always the case, and for some users, platforms such as PairD feel entirely novel and different. This should be recognised upfront in both the design of the platform and in communications and training material, all of which should cater for beginners as well as more advanced users.

4 Emphasise job augmentation rather than job displacement:

Whilst there is a lot of excitement with Generative AI, there is also anxiety and fear, for example around job displacement. These worries are heightened by the current challenging economic climate, which has driven many firms to implement hiring freezes and redundancy rounds. Against this backdrop, platforms such as PairD can seem threatening. If managed poorly, this can make people feel less confident and secure, rather than empowering individuals to focus more on the most valuable aspects of their jobs. It is important that teams recognise these very real concerns, and emphasise that platforms such as PairD really are intended to augment rather than replace jobs, even if they do change some of the ways in which we work across the firm. This should be reflected through effective communication, training, and platform design.

5 Gain feedback from users early and on an ongoing basis:

One of the main advantages of in-house platforms such as PairD is the ability to gain rich feedback from users, as this data can be used to design new features and tailor the platform further to user requirements. This can be achieved through surveys and by analysing user behaviour, for example through prompts, mouse clicks, and the use of specific features. Of course, this type of monitoring should only be implemented whilst maintaining strict user privacy and security, and must be accompanied by transparent communication that ensures users are aware of and consent to the ways in which their data is collected and used.

Concluding thoughts

PairD is both a big and small project for Deloitte.

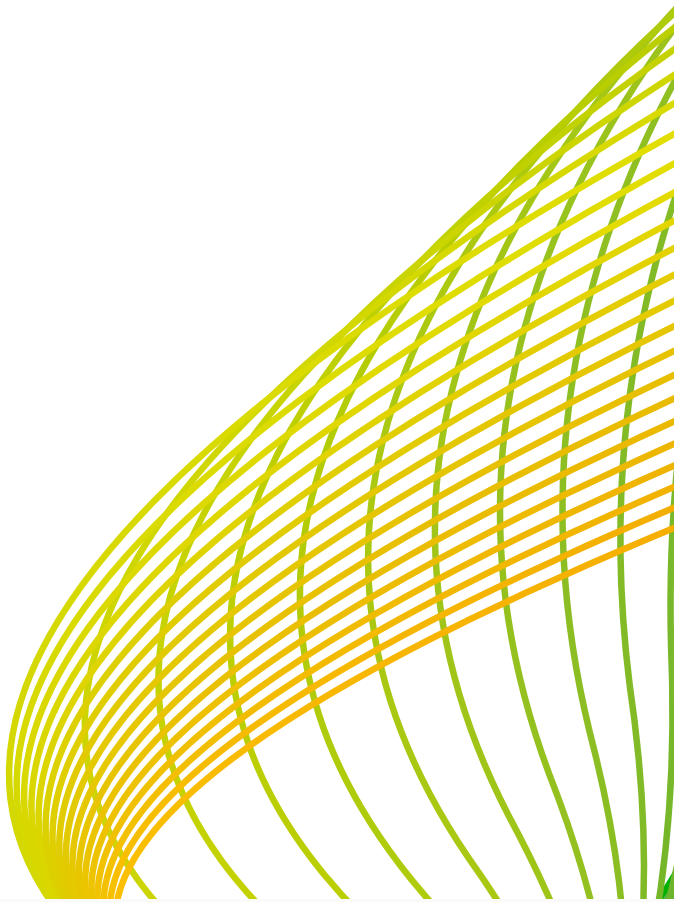
It is by no means the largest and most complex platform that we have developed and scaled, and it is certainly not the first to incorporate AI.

It is, however, the first Generative AI platform that we have productionised and scaled internally within the UK firm, including across multiple jurisdictions.

Scaling AI is always challenging, but we were determined to move beyond Generative AI strategy and PoCs, and test first hand both the benefits and challenges of building a secure, private and customised Generative AI platform. We have captured some of the key lessons from developing PairD within this report, and we hope that these insights will serve as a useful reference for other organisations that are similarly looking to develop and scale their own Generative AI platforms.

Our PairD journey is by no means complete, and we will continue to develop the platform to explore new use cases and to incorporate new technologies and features. We will also continue to learn, and we hope to capture some of these additional future insights in subsequent reports.

In the meantime, if you would like to learn more about PairD and scaling Generative AI, please reach out to the authors of this report and to the Deloitte AI Institute.[7]



[7] mailto: UKDeloitteAIInstitute@deloitte.co.uk

About us

Authors

Sulabh Soral
Chief AI Officer and AI Institute Lead
ssoral@deloitte.co.uk

Richard Flint
Deloitte AI Institute UK Research and Eminence Lead
rflint@deloitte.co.uk

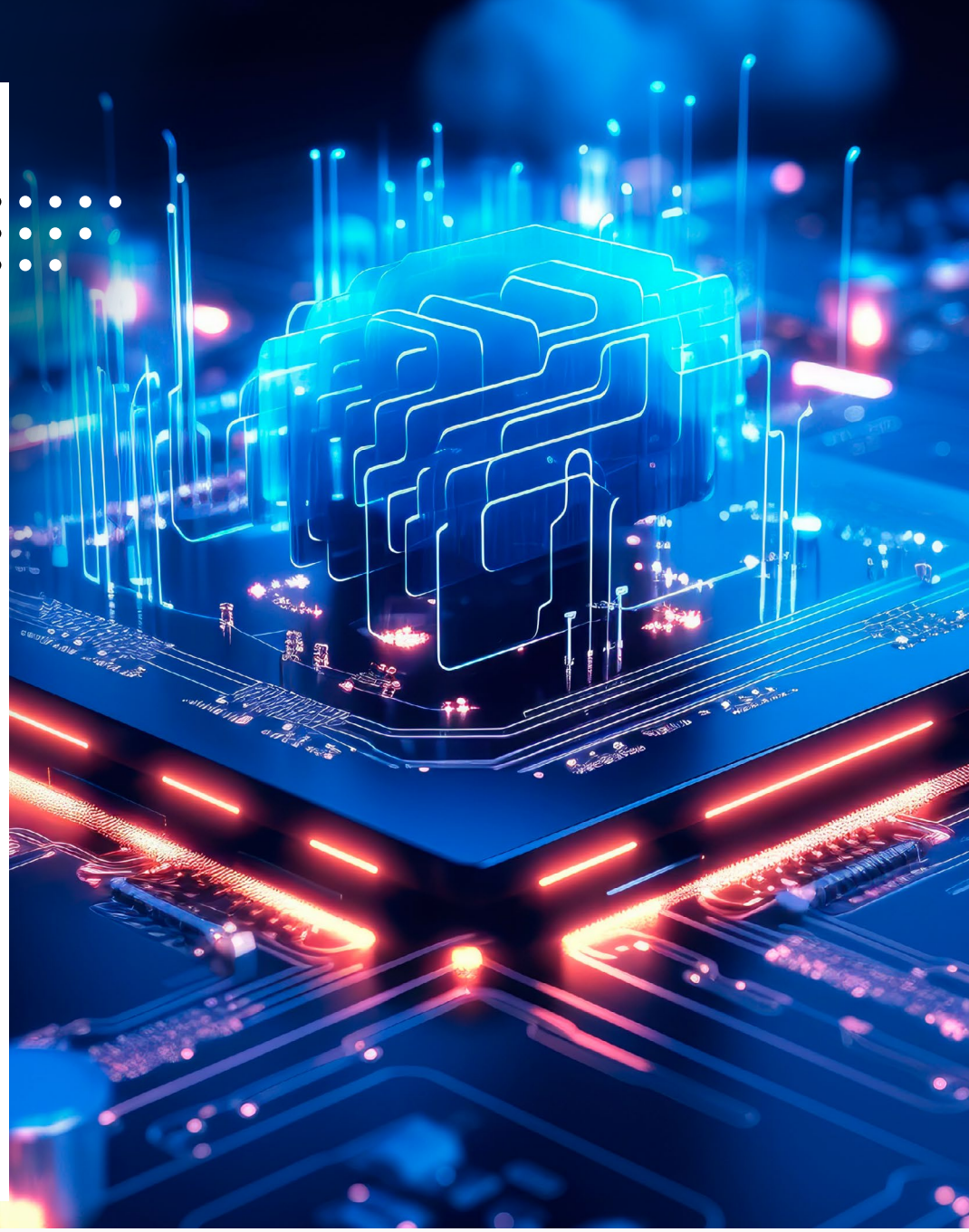
Written in collaboration with PairD
Deloitte's AI Assistant

Deloitte AI Institute UK

The Deloitte AI Institute™ UK[8] connects enterprises to the AI ecosystem through cutting-edge research, development, perspectives and analysis. The Deloitte AI Institute UK is part of the global Deloitte AI Institute network.[9]

We would like to thank all of the individuals and teams across Deloitte that contributed to PairD and to this report. PairD continues to be a hugely collaborative project that is possible only through the on-going efforts, expertise and support of our colleagues.

[8] https://www2.deloitte.com/uk/en/pages/deloitte-analytics/articles/advancing-human-ai-collaboration.html
[9] https://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/advancing-human-ai-collaboration.html



Deloitte.

Designed and produced by 368 at Deloitte. J31869