Deloitte.

Transparency and Responsibility in Artificial Intelligence

A call for explainable AI

Table of content

Introduction

At Deloitte, we expect Artificial Intelligence (AI) to be hugely impactful. But in order for that to be a positive impact, we need to act now. Deloitte is a strong advocate of transparency and responsibility in AI: AI that has been thoroughly tested, that is explainable to customers and employees, and that has all the ethical considerations in place.

This publication explores our vision on transparent and responsible AI. We also present four propositions that we have developed around the topic. With them, we hope to contribute to a movement to use AI for good. This is the moment to ensure that AI models are built the right way. Only then, can we make sure that AI technologies will not cause harm but benefit humanity.

Richard Roovers, *Innovation lead for Deloitte Netherlands and member of the NWE Innovation Executive*

A Call for Transparency and Responsibility in Artificial Intelligence

Artificial Intelligence (AI) is increasingly used for decisions that affect our daily lives – even potentially life or death ones. Deloitte therefore calls for transparency and responsibility in AI: AI that is explainable to employees and customers and aligned with the company's core principles.

Media coverage of AI tends to be either euphoric or alarming. In the first variant, AI is presented as a divine technology that will solve all our problems, from curing cancer to ending global warming. In the latter, *Frankenstein*- or *Terminator*-inspired narratives depict AI as a technology that we cannot keep under control and that will be out-smarting humans in ways we cannot foresee – killing our jobs, if not threatening the survival of humanity.

In the past few years, the number of negative stories about AI has increased markedly. Tech-entrepreneur Elon Musk has even stated that AI is more dangerous than nuclear weapons. There have been numerous cases in which advanced or AI-powered algorithms were abused, went awry or caused damage. It was revealed that the British political consulting firm Cambridge Analytica harvested the data of millions of Facebook users without their consent to influence the US elections, which raised questions on how algorithms can be abused to influence and manipulate the public sphere on a large scale. It has become clear that AI models can replicate and institutionalise bias. For instance, an AI model named COMPAS – used across the US to predict recidivism – turned out to be biased against black people, whereas an AI-powered recruiting tool showed bias against women.

Some types of AI can be gamed by malicious forces and end up behaving totally differently than intended. That was the case with an AI-powered chatbot, which turned from a friendly chatbot into a troll posting inflammatory messages and conspiracy theories in less than a day. Other cases brought to the surface unsolved questions around ethics in the application of AI. For instance, Google decided not to renew a contract with the Pentagon to develop AI that would identify potential drone targets in satellite images, after large-scale protests by employees who were concerned that their technology would be used for lethal purposes.

Stefan van Duin, partner Analytics and Cognitive at Deloitte and an expert in developing AI solutions, understands this public anxiety about AI. "The more we are going to apply AI in business and society, the more it will impact people in their daily lives – potentially even in life or death decisions like diagnosing illnesses, or the choices a self-driving car makes in complex traffic situations," says Van Duin. "This calls for high levels of transparency and responsibility."

Deloitte is a strong advocate of transparency and responsibility in AI. Transparent AI is AI that is explainable to employees and customers. That can be challenging, because AI is not transparent by nature, says Van Duin.

"The more we are going to apply AI in business and society, the more it will impact people in their daily lives"

"So, the question is: How can we make AI as transparent as possible? How can we explain how an AI-based decision was made, what that decision was based on, and why it was taken the way it was taken?" Next to transparency, there is the question of responsibility in AI. "Transparent AI makes our underlying values explicit, and encourages companies to take responsibility for AI-based decisions," says Van Duin. "Responsible AI is AI that has all the ethical considerations in place and is aligned with the core principles of the company."

This publication explores Deloitte's point of view on transparency and responsibility in AI. It has been informed by Deloitte's own experiences in developing and applying AI-enabled technology, both in the company and for its clients, as well as from its long-standing experience with validating and testing models, providing assurance, and advising on strategy and innovation projects.

The publication will discuss four propositions that Deloitte has developed around the topic of transparent and responsible AI. 'GlassBox' is a technical toolkit to validate AI models and to help explain the decision-making process of AI models to employees and customers. The 'FRR Quality Mark for Robotics and AI' is a quality mark for AI-powered products, which ensures customers that AI is used responsibly. 'Digital Ethics' provides a framework to help organisations develop guiding principles for their use of technology, and to create a governance structure to embed these principles in their organisation. Finally, 'AI-Driven Business Models' covers the complete journey, from defining the vision and building the capabilities up until actual implementation and capturing value.

Transparent AI is explainable AI

One reason why people might fear AI, is that AI technologies can be hard to explain, says Evert Haasdijk. He is a senior manager Forensic at Deloitte and a renowned AI expert, who worked as an assistant professor at VU Amsterdam and has more than 25 years of experience in developing AI-enabled solutions. "Some AI technologies are pretty straightforward to explain, like semantic reasoning, planning algorithms and some optimisation methods," says Haasdijk. "But with other AI technologies, in particular data-driven technologies like machine learning, the relation between input and output is harder to explain. That can cause our imaginations to run wild."

But AI doesn't have to be as opaque as it may seem. The proverbial 'black box' of AI can be opened, or at least, it is possible to explain how AI models get to a decision. The point of transparent AI is that the outcome of an AI model can be properly explained and communicated, says Haasdijk. "Transparent AI is explainable AI. It allows humans to see whether the models have been thoroughly tested and make sense, and that they can understand why particular decisions are made."

Transparent AI isn't about publishing algorithms online, says Haasdijk – something that is currently being discussed as mandatory for algorithms deployed by public authorities in the Netherlands. "Obviously there are intellectual property issues; most companies like to keep the details of their algorithms confidential," says Haasdijk. "But more importantly, most people do not know how to make sense of AI models. Just publishing lines of code isn't very helpful, particularly if you do not have access to the data that is used – and publishing the data will often not be an option because of privacy regulations." According to Haasdijk, publishing AI algorithms will not, in most cases, bring a lot of transparency. "The point is that you have to be able to explain how a decision was made by an AI model."

Transparent AI enables humans to understand what is happening in AI models, emphasises Haasdijk. "AI is smart, but only in one way. When an AI model makes a mistake, you need

human judgment. We need humans to gauge the context in which an algorithm operates and understand the implications of the outcomes."

The level of transparency depends on the impact of the technology, adds Haasdijk. The more impact an advanced or AI-powered algorithm has, the more important it is that it is explainable and that all ethical considerations are in place. "An algorithm to send personalised commercial offerings doesn't need the same level of scrutiny as an algorithm to grant a credit or to recommend a medical treatment," says Haasdijk. "Naturally, all AI models should be developed with care, and organisations should think ahead of the possible ramifications. But AI models that make high-impact decisions can only be allowed with the highest standards of transparency and responsibility."

Detecting hidden bias

So how do you create transparent AI? First, says Haasdijk, there are technical steps. "The technical correctness of the model should be checked, all the appropriate tests should be carried out and the documentation should be done correctly," he explains. "The developer of the model has to be able to explain how they approached the problem, why a certain technology was used, and what data sets where used. Others have to be able to audit or replicate the process if needed."

The next thing to assess is whether the outcomes of the model are statistically sound. "You should check whether certain groups are under-represented in the outcomes and, if so, tweak the model to remedy that." This step can help to detect hidden biases in data, a well-known problem in the world of AI, says Haasdijk. "Suppose you use AI to screen job applicants for potential new managers in your company. If the model is fed data from previous managers who were mostly white males, the model will replicate that and might conclude that women or people of colour are not fit for management roles."

A challenge here is that most data sets have not been built specifically to train AI models. They have been collected for other purposes, which might result in skewed outcomes. "AI models are unable to detect bias in data sets," Haasdijk explains. "Only humans, who understand the context in which the data has been collected, can spot possible biases in the outcome of the model. Checking training data for possible bias therefore requires utmost care and continuous scrutiny."

Finally, AI models should be validated to enable organisations to understand what is happening in the model and to make the results explainable. Deloitte's GlassBox proposition offers a variety of tools – both open source and developed in-house – to help companies validate advanced or AI-powered algorithms. The toolkit allows organisations to look inside the 'black box' of AI, to expose possible bias in training data and to help explain the decision-making process of AI models to employees and to customers. You can read more about GlassBox in the article *Unboxing the Box with GlassBox: A Toolkit to Create Transparency in Artificial Intelligence*.

'Computer says no'

There are a couple of reasons to pursue transparent AI. An important one is that companies need to understand the technologies they use for decision making. As obvious as this sounds, it is not always a given. "The board room and higher management of a company are often not really aware what developers in the technical and the data analytics departments are working on," says Haasdijk. "They have an idea, but they don't know exactly. This causes risks for the company."

Paradoxically, as open source AI models are becoming more user-friendly, there might be more AI applications built by people who do not completely understand the technology. "There are open source models that are very easy to use. Someone might feed it with data, and get a result, without really understanding what is happening inside the model and comprehending the possibilities and limitations of the outcomes," says Haasdijk. "This can cause a lot of problems in the near future. A risk of AI becoming more user-friendly and more widely available is that someone in a company might be using AI irresponsibly, and not being aware of it – let alone their bosses knowing about it."

It can be hard for companies to keep track of all the AI models that are used within their organisation, says Haasdijk. "A bank recently made an inventory of all their models that use advanced or AI-powered algorithms, and found a staggering total of 20,000." Some of these algorithms, like capital regulation models, are under strict scrutiny from regulators. But in most cases, advanced or AI-powered algorithms are not subject to any kind of external and internal regulation – think of algorithms used in marketing, pricing, client acceptance, front office or automated reports. "Transparent AI can help companies to regain control over the variety of AI models that are deployed in their organisation," Haasdijk says.

Transparent AI can give organisations more insight in when and why AI algorithms make mistakes, and on how to improve their models accordingly. "AI models do make mistakes – in many instances they make fewer mistakes than humans, but still, you want to know when and why that happens," says Haasdijk. "Take the example of the self-driving car that hit a lady who was walking with her bike, because the algorithm misjudged the situation. It is essential that companies understand when and why mistakes like these happen, to avoid similar accidents in the future."

Finally, transparent AI can help organisations to explain individual decisions of their AI models to employees and customers. And that's not all that customers expect from the organisations; with the GDPR ruling that recently came into force, there is also regulatory pressure to give customers insight into how their data is being used. "Suppose a bank uses an AI model to assess whether a customer can or cannot get a loan," says Haasdijk. "If you deny a loan, the customer probably wants to know why that decision has been made, and what they need to change in order to get the loan. That means the bank must have a thorough understanding of how their AI model reaches a decision, and to be able to explain this in clear language. 'Computer says no' is not an acceptable answer."

Re-establishing trust

The potential ramifications of AI have become a concern on the level of politics. By now, more than 20 countries are working on a national AI strategy, and various national and regional AI-related policy measures have been announced. The European Commission installed a High-Level Expert Group on Artificial Intelligence to work on Ethical Guidelines for Trustworthy AI. On an international level, the United Nations has created an AI and Global Governance Platform to explore the global policy challenges raised by AI, as a part of the Secretary-General's Strategy on New Technologies.

A lot of these initiatives are addressing the public anxiety concerning the loss of control to an AI revolution. With every new incident of AI going awry – of which a couple examples were mentioned at the beginning of this article – the demand for transparent, and particularly for responsible, AI is rising. "Not only do people want to understand how AI-based decisions are made," says Van Duin. "They want to be reassured that AI is used to benefit mankind and is not causing harm."

So far, not one commonly accepted framework for how to assess AI has been set, notes Van Duin. "For privacy issues, organisations must adhere to GDPR regulation, the European legal framework for how to handle personal data. But there is nothing similar for ethics in AI, and so far, it looks like there will only be high-level guidelines that leave a lot of room for interpretation. Therefore, there is a lot of responsibility for companies, organisations and society to make sure we use AI ethically."

In order to set an industry standard for responsible AI & Robotics, the non-profit organisation Foundation for Responsible

"Technically, a lot is possible. Now companies have to decide how far they want to go"

Robotics (FRR) and Deloitte are developing the FRR Quality Mark for Robotics and AI. This is a recognisable quality mark for consumers that ensures robotics and AI are designed, developed and used responsibly, comparable to the Fairtrade quality mark for products that have been produced to Fairtrade Standards. The FRR Quality Mark for Robotics and AI aims to re-establish a sense of trust in robotics and AI technology among consumers. It will indicate whether a robotics or AI product adheres to the ethical standards that come with the quality mark. You can read more about the quality mark in the article *FRR Quality Mark for Robotics and AI: A Recognisable Quality Mark for Responsible Design, Development and Use of Robotics and Artificial Intelligence.*

Embedding ethics in the organisation

It's not just consumers who are demanding a clear vision on AI; employees are demanding it from their company too. This makes the stakeholder landscape around AI more complex. As previously mentioned, last year Google decided not to renew a contract with the Pentagon involving automatic recognition software for drones, after large-scale protests by employees. Over 4,000 employees, including some of the company's top AI researchers, signed a petition demanding "a clear policy stating that neither Google nor its contractors will ever build warfare technology". The protests resulted in Google publishing a value statement on AI, in which the company, among other things, pledged not to use its AI technology for weapons. A few months later, Google decided to drop out of bidding for a $10 billion Pentagon cloud-computing contract, citing that the contract would clash with the company values on AI.

"A company pulling out of a bid for a $10 billion-dollar contract for ethical reasons is a clear sign of the times", says Tjeerd Wassenaar, partner at Deloitte Risk Advisory with a focus on ethics and corporate values. With fast-developing technologies like AI, organisations are forced to think about the ethical consequences in a structural way, argues Wassenaar. "Technically, a lot is possible," he says. "Now companies have to decide how far they want to go. Currently, there is no clear policy on this in most companies. The case of Google employees protesting against their own company indicates that many people in the organisation do not feel that these boundaries have been set."

The demand for transparent and responsible AI is part of a broader debate about company ethics, says Wassenaar. "What are your core values, how do they relate to your technological and data capabilities, and what governance frameworks and processes do you have in place to keep up with them? Those are questions that companies currently must address. If they don't, they risk their reputation, legal issues and fines, and worst of all, the trust and loyalty of their customers and employees."

Deloitte's Digital Ethics proposition helps companies to treat digital ethics in a structural way. It helps companies to define their core

principles, to establish governance frameworks to put these into practice, and to set up benchmarks to monitor whether the principles have been effectively implemented. You can read more about Digital Ethics in the article *Digital Ethics: Structurally Embedding Ethics in your Organisation.*

"This is the moment to ensure that AI-powered algorithms are built the right way"

Realising the positive potential of AI

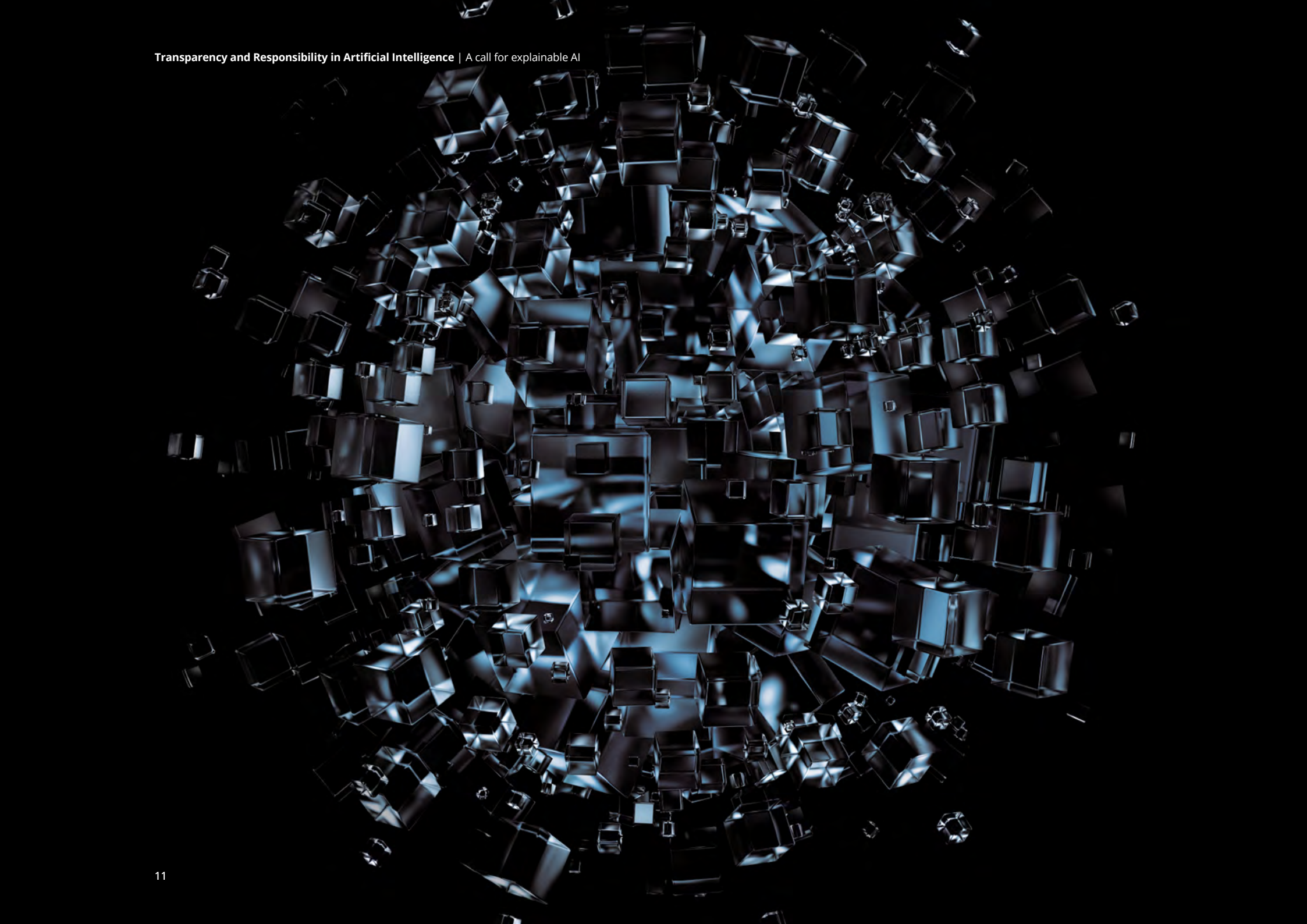
Transparency and responsibility in AI will help to ensure that advanced or AI-powered algorithms are thoroughly tested, explainable, and aligned with the core principles of the company. In short, it will help organisations to regain control over the AI models that are deployed. "We need to feel that we are in control," says Van Duin. "A lot of people feel the opposite: that we are losing control. By creating transparency and establishing clear guidelines and procedures for creating AI applications, we can make sure that advanced or AI-powered algorithms function as intended and that we can capture the value that it promises."

Despite the *Frankenstein*- or *Terminator*-inspired narratives you can read in media articles, most advanced or AI-powered algorithms that are being developed right now are relatively innocent and not about high-impact decisions, says Van Duin. "But we should start to think transparency and responsibility in AI right now." Self-driving cars are the perfect example, he says: they will not be widely available any time soon, but companies are currently working on this technology. And the self-driving car will, by definition, have life and death situations programmed. "This is the moment to ensure that these algorithms are built the right way."

With all the talk about the potential risks of using AI, one might almost forget about the fact that AI offers genuine business opportunities. Early adopters in business have shown amazing results in cost reduction, better service, better quality or even completely new business models. Deloitte's AI-Driven Business Models proposition helps companies to assess how they can leverage AI in their company right now and what capabilities they need to build. You can read more about AI Driven Business Models: a strategic approach to capture the full potential of AI.

In the long run, transparency and responsibility in AI will not only help to avoid disasters, but will also help to realise the positive potential of AI in making major contributions to the good life, says Van Duin. He sums up: "There is a huge potential in healthcare. AI can support medical diagnoses, it could support treatment and save lives. It could reduce our energy consumption, by optimising processes. It could reduce the need for us to own a car, which will have a positive environmental impact and will save us money. It could reduce the number of traffic accidents. The access to information will get better, it will help us to work more effectively. That could lead to us having more spare time – maybe at some point in time, a 30- or 20-hour work week will become the norm rather than 40 hours. In short, the quality of life could increase massively."

"The impact of AI can be enormous", says Van Duin. But ultimately, he says, AI will need the trust of the general public to have the most impact. "If all the considerations around transparency and responsibility are in place, AI can make the world a better place."



Unboxing the Box with GlassBox

A Toolkit to Create Transparency in Artificial Intelligence

Artificial Intelligence (AI) models can become so complex that we no longer understand the output. This undermines the trust of companies and customers. Therefore, Deloitte has developed GlassBox: a toolkit that looks inside the proverbial 'black box' of AI-powered algorithms.

Artificial Intelligence models, in particular data-driven models like machine learning, can become highly complex. These algorithms are typically presented as a 'black box': you feed them with data and there is an outcome, but what happens in the meantime is hard to explain.

This lack of understanding of AI technology causes large risks for companies, says Roald Waaijer, director Risk Advisory at Deloitte. "AI-powered algorithms are increasingly used for decisions that affect our daily lives. Therefore, if an algorithm runs awry, the consequences can be disastrous. For a company it can cause serious reputational damage and lead to fines of tens of millions of euros." Worst of all, he adds, it may hurt customers, for instance by unintentionally treating them unfairly if there are biases in the algorithm or training data. "This may lead to a serious breach of trust, which can take years to rebuild."

To help companies look inside the proverbial black box of AI, Deloitte has developed GlassBox. This technical toolkit is designed to validate AI models and to expose possible bias and unfairness – in short, to check whether AI-powered algorithms are doing what they are supposed to do. "It's just like bringing your car to the garage," explains Waaijer. "You occasionally need to look under the bonnet to see whether everything is working properly. That is what GlassBox does: we look under the bonnet of an algorithm to check the AI engine."

Moreover, Deloitte offers tools to help explain the decision-making process of AI models to employees and customers, for instance by visualising how an AI-powered algorithm came to a decision. "With the GDPR regulation that recently came into force, consumers have the right to receive a meaningful explanation of how their data was used to get to a decision," says Waaijer. "'Computer says no' is not a sufficient answer. You have to explain the decision and give insight into what happens inside the black box."

"If an algorithm runs awry, the consequences can be disastrous"

"Computer says no is not a sufficient answer"

Inspection methods

There is a wide variety of AI models, like neural networks, discriminant-based methods, tree-based methods and others. The GlassBox toolkit has different inspection methods for each of them (see illustration). Some tools have been developed in-house; others, like ELI5 or LIME, are open source. Bojidar Ignatov, junior manager Financial Risk Management at Deloitte with a focus on advanced modelling, has been involved with assembling the GlassBox toolkit. "There are various ways to open the black box and get an idea of how these algorithms operate," he says.

Take for instance random forest models. With this AI technology, you randomly generate a lot of trees – a forest. All the trees have a different combination of variables that interact, and the algorithm tries to find the tree that is most representative for the data. The GlassBox toolkit offers various ways to validate random forest models: for example, it is possible to reconstruct how features are selected, to take a deep dive on different features or to understand the interaction of the features.

One example of a feature deep dive of random forest models is to replicate a local optimum of the model used for a single decision of the model, and to estimate the global optimum. "By comparing these, you can figure out what features have a strong impact on a decision, and which are the key features overall," explains Ignatov.

In order to get an understanding of the interaction of features in random forest models, Deloitte developed the 'Interaction Matrix'. The Interaction Matrix is a tool that shows which features are relatively often placed together within the trees. "These interactions can be visualised in a heat map, so that it is easy to see which combination of factors often contribute to the outcome," says Ignatov. "The warmer the heat map, the more often two factors are connected." In the end, a human expert that understands the context of the algorithm can judge whether these features are indeed important for the decision, or if the model needs tweaking.

Model Family	Examples of inspection methods
	Neural Networks • Weight visualization in neural network • Garson's method on relative node importance • Hyper-parameters assumptions validation
	Tree-based Methods • Variable importance plot (mean decrease accuracy) • Variable interaction matrix (bases on proximity of variables in the trees) • Tree interpreter (show individual prediction determination); LIME
	Discriminant-based Methods • Analyze SVM weights to show feature importance • Sensitivity analysis of a specific variable ceteris paribus
	Instance-based Methods • PCA to identify the most important risk drivers • Sensitivity analysis of a specific variable ceteris paribus • Individual prediction explanation: investigate the neighboring data points
	Generative Methods • Validate assumptions implied by the Generative PDF, do they hold and are they realistic • Validate class imbalance remediation method

Overarching Dimensions

Data Quality

AI algorithms are highly dependent on the model inputs and data provided.

Examples of focus areas in scope of data validation:

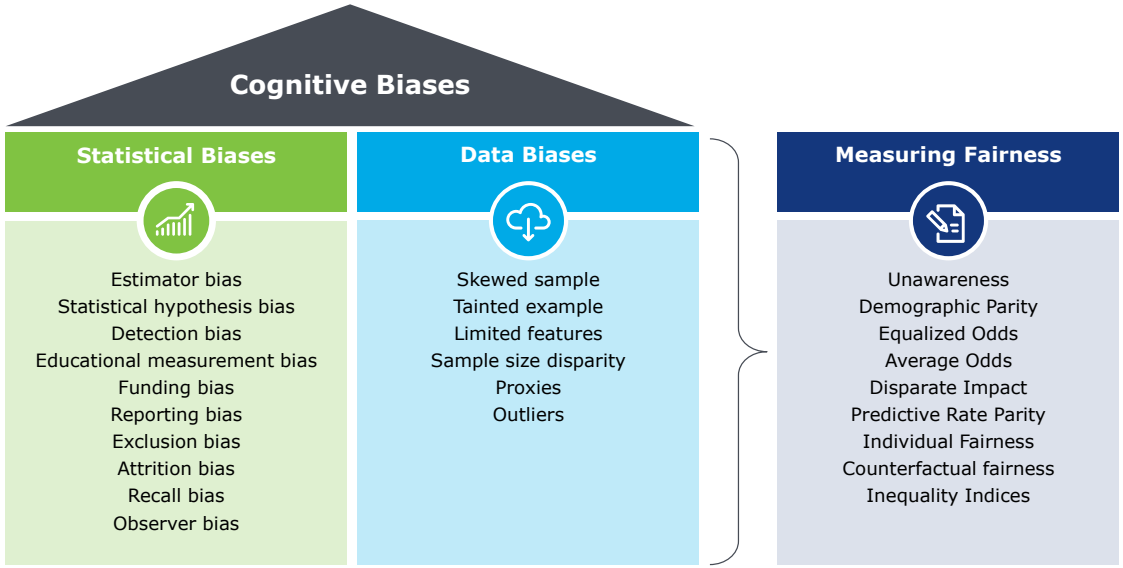
- Scaling of input data
- Data transformations
- Data enrichments and enhancements
- Manual and implicit binning
- Sample balancing
- Creeping (fresh) data biases

Avoiding bias in data

An important dimension in validating AI models is checking the data quality. An AI model can work perfectly fine, but if the input data is flawed or biased, this will affect the outcome. Even if you don't use race or gender explicitly in your model, it doesn't mean these factors do not play a role in the output, explains Waaijer. "AI models are famous for proxying features via multiple factors, like postal code, height, etcetera. When the data input is biased, the AI model will find a way to replicate the bias in the outcomes, even if the bias isn't explicitly included in the variables of the model."

The GlassBox toolkit has different tools to expose bias in data sets. A part of the data validation is to check whether all the required steps have been made before data goes into the AI model, such as scaling the input data, data transformations, data enrichments and enhancements, manual and implicit binning or sample balancing. Next, bias is checked against a predetermined set of biases that are possibly present in the data set or for creeping bias in fresh data. It all comes down to testing different scenarios, says Waaijer. "For instance, if you want to test whether a data set has a gender bias, you can test how the distribution of men and women plays out with a certain combination of parameters. If the outcome is very different than expected, something is wrong."

"One difficulty is that, in most cases, it is only possible to detect bias when a model is in use", adds Waaijer. "You only really see it happening in practice. For instance, research has shown that a combination of seemingly innocent factors like height and postal code can disadvantage people of a certain background. This is not something you would expect, you can only discover this when the model is in use." This means that ethical considerations



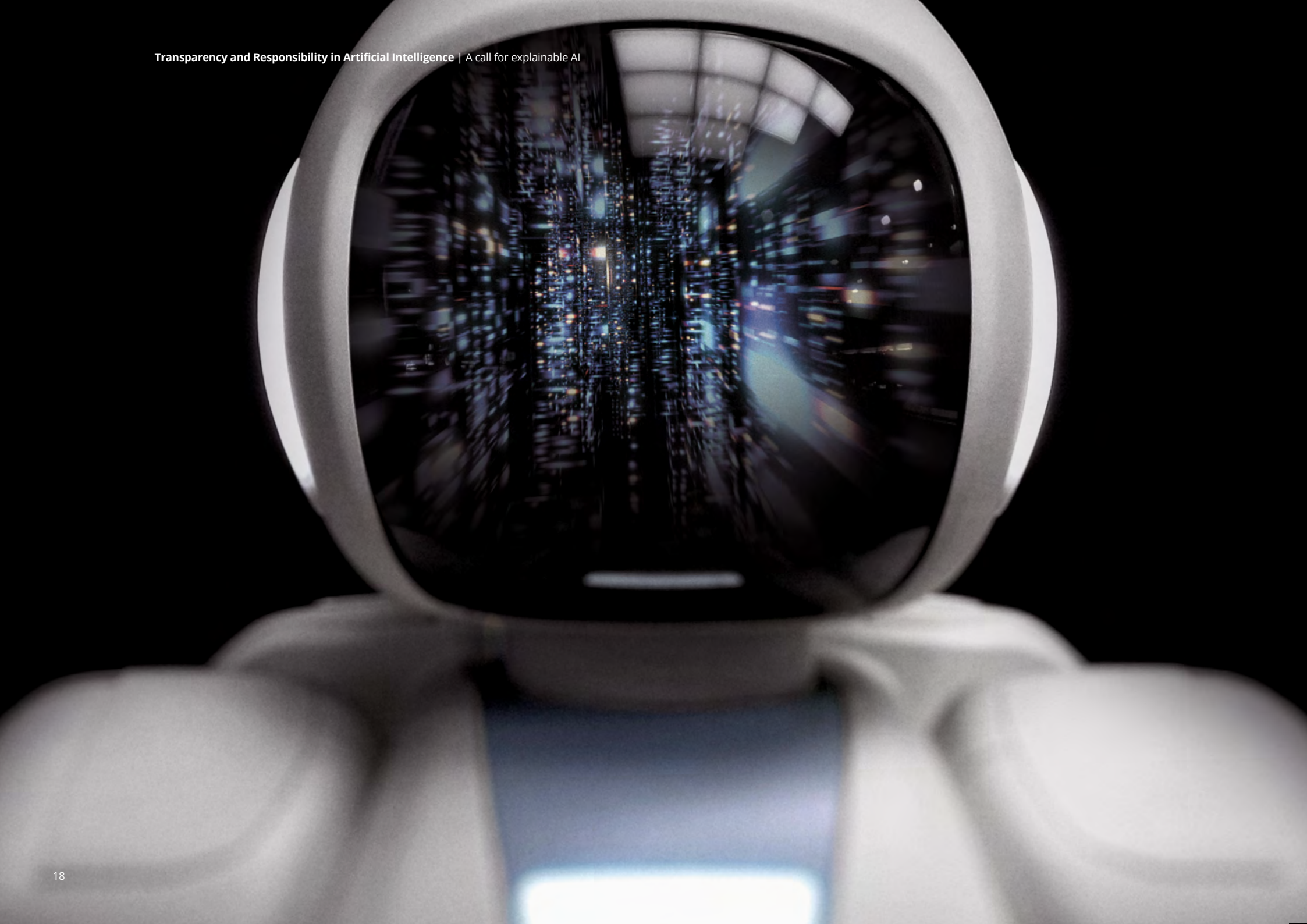
in AI models should be in place not just at the beginning, but continuously, says Waaijer. "You need a monitoring cycle in which you continuously, or at least periodically, monitor the results."

Human judgement

The GlassBox toolkit offers various tools to gain insight into advanced or AI-powered algorithms. However, human judgment is needed to make sense of the results. "Only humans can gauge the context in which an algorithm operates and can understand the ramifications of an outcome," says Waaijer. "Unboxing the algorithm is step one; step two is to be sure that the algorithm operates in line with the values of the company. For that, you need human expertise."

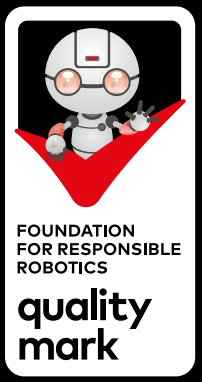
The GlassBox toolkit enables organisations to take control over the various AI models they have in use. It allows them to ensure the outcomes of AI-powered algorithms are explainable and make sense. "Opening the black box of AI will become a business priority," says Waaijer. "If you can make sure you use AI-powered algorithms correctly and responsibly, not only will you avoid risks, but you will also be able to realise the full potential with AI."

"Opening the black box of AI will become a business priority"



FRR Quality Mark for Robotics and AI

A Recognisable Quality Mark for Responsible Design, Development and Use of Robotics and Artificial Intelligence



The Foundation for Responsible Robotics and Deloitte are developing the FRR Quality Mark for Robotics and AI to ensure that robots and AI-powered products are developed and used in a responsible manner.

Robots and AI-powered products are increasingly used for tasks that can have a high impact on our daily lives. AI-powered products are, for instance, changing the way hiring procedures, credit ratings, or fraud detection are being carried out. Robots are being used to support and replace humans in dull, dirty or dangerous tasks: drones for search and rescue, warehouses powered by robots, or surgical robots assisting surgeons.

While robots and AI-powered products may be working behind the scenes today, we can expect them to become a more pervasive part of our lives in the near future. But for most people it is hard, if not impossible, to understand how these products operate. How can you, as a consumer, make sure that robots and AI-powered products are being made in a responsible manner and are doing what they are supposed to do?

To address these questions, the Foundation for Responsible Robotics (FRR) and Deloitte are developing the FRR Quality Mark for Robotics and AI: a quality mark that ensures that robots and AI-powered products are made in a responsible way, with attention to human rights and values. The aim is to create a recognisable quality mark for consumers, comparable with the Fairtrade quality mark for products that have been produced to Fairtrade Standards. "As a consumer, you should be able to rely on a standard," says Marc Verdonk, partner Emerging Technology at Deloitte. "You don't have to understand all the details of the technology, you just have to know that someone with the right expertise using the right framework can ensure it is being made and used in a responsible way."

Developing a framework

FRR is a non-profit organisation with a charity status that aims to create a future of robotics and AI that are responsibly designed, developed and used. The board of the FRR consists of experts in ethics, robotics and AI, under the leadership of Aimee van Wynsberghe, assistant professor in Ethics of Technology at the Delft University of Technology. "Our aim is to create a culture of responsible development of robotics and AI, to promote public well-being for us and for the coming generations," says Van Wynsberghe. Deloitte supports FRR through the Deloitte Impact Foundation and is contributing its AI and auditing expertise to FRR to help create the FRR Quality Mark for Robotics and AI.

Although the overall scope for the FRR Quality Mark for Robotics and AI is still being discussed, an important premise of it is to ensure that human rights and societal values are dealt with throughout the entire chain of creating the product. All products using robotics or AI will be able to apply for the quality mark. "That can be a wide variety of products," explains Verdonk, "like algorithms, robots, smart devices and connected toys."

During the assessment, the FRR Quality Mark for Robotics and AI will take a close look at the product, such as the controls related to actuators, communication interface, sensors, data storage and firmware. Moreover, it will take a close look at the AI engine: How are the algorithms trained and tested? Can the algorithms be changed without authorisation? The quality mark will also look at the policies and procedures used by the company that creates the product. "In all these aspects, the principles of creating something in an ethical manner have to be safeguarded," says Verdonk.

Ensuring security and privacy

One aspect the FRR Quality Mark for Robotics and AI will take into account is security. "If you want to promote the responsible use of robotics AI, security is very important," says Verdonk. "With large robots or drones, security should always be a priority. Or think of a doll that uses AI to interact with children. If the doll can be hacked and used for something completely different, the consequences can be disastrous."

Another aspect is privacy. "Take for example delivery robots," says Verdonk. "They need cameras to see the road. In theory, these cameras are able to film people who are passing by. But you can also consciously design the robots in a way that the camera is unable to film anyone above their knees, like the self-driving delivery robot of Starship Technologies. Respecting people's privacy through this design choice is an ethical choice and these are the kinds of choices we want to encourage." The FRR Quality Mark for Robotics and AI will therefore not only look at if the product is working properly and if it has been thoroughly tested, but also at the considerations that have been made throughout the design process.

An important aspect of the responsible use of AI is avoiding bias, says Verdonk. There have been many instances in which algorithms unintentionally replicated bias. A recent example is an AI-powered recruiting tool that showed bias against women. "The algorithm was actually properly trained, but it was fed with data of historical profiles of candidates that had been hired," explains Verdonk. "And since the company had hired more men than women in the past, the system taught itself that male candidates were better. The output of AI-powered algorithms is determined by the quality of the input data as well as by what the algorithm is doing. Therefore, the approach to both will be considered in the quality mark."

"The FRR Quality Mark for Robotics and AI sets a standard that companies can adhere to"

Steering industry in the right direction

For consumers, the FRR Quality Mark for Robotics and AI can help them trust a technology that can be hard to understand. It can also help them to make a well-considered decision about what kinds robotic or AI-powered products they want to buy or use. For companies, the FRR Quality Mark for Robotics and AI can help to develop robots and AI in a transparent and responsible way. "The FRR Quality Mark for Robotics and AI sets a standard that companies can adhere to," Verdonk explains. "Companies want to show their customers that they actually do care about ethics. An independent quality mark can help to demonstrate that they take ethical considerations seriously."

Verdonk is convinced that the FRR Quality Mark for Robotics and AI will help to lift the industry standards for creating robots and AI-powered products, and that it will steer the industry in the right direction. "I do think that robotics and AI will be very impactful technologies," he says. "But in order to make that a positive impact, we need thoroughly tested products that have been developed and are being used in a responsible manner, as well as the trust of the general public. If we can create a trusted standard that ensures the responsible use of robotics and AI, we can create a future in which these technologies can have a tremendous positive impact."

Digital Ethics

Structurally Embedding Ethics in your Organisation

The use of advanced and Artificial Intelligence-powered algorithms can lead to serious ethical breaches. Deloitte's Digital Ethics proposition helps organisations to develop a clear vision of their business principles, and to create a governance framework to ensure that their use of technology is aligned with their values.

A few years ago, a teenage girl in Minnesota received a booklet from a US department store with coupons for baby gear. Her father was furious: Why was this company sending these coupons to his daughter? A couple of days later, the girl admitted to her father that she was, in fact, pregnant. The department store had developed an algorithm that was able to assess the likelihood that a customer was pregnant based on her shopping behaviour – in this case, the algorithm happened to be spot on.

The story sparked great controversy. The department store hadn't breached any of the US privacy laws of the time, but for the general public, it had crossed an ethical boundary. Living in a world in which companies know more about our children than we do ourselves made people feel uneasy. Since then, the number of incidents in which the use of data and advanced or Artificial Intelligence (AI)-powered algorithms caused controversy has exploded. Every week there is something in the news: hacks, leaks, data breaches, algorithmic bias and privacy infringements. By now, large technology companies are even formally warning investors of ethical uncertainties around AI.

"Companies are actively developing and exploiting their technological capabilities, often without considering the ramifications," says Tjeerd Wassenaar, partner at Deloitte Risk Advisory with a focus on ethics and corporate values. "This can lead to serious reputational damage, legal issues and fines, and worst of all, the loss of trust and loyalty of customers. Trust is a company's most valuable asset: if you betray your customers' trust for short-term profit, you will lose in the long run."

Safeguarding ethics in the use of advanced or AI-powered algorithms has become one of the most prominent questions of this era. Deloitte's Digital Ethics proposition addresses this question. It provides a framework to help organisations develop guiding principles for their use of technology, to create a governance structure to embed these principles in their organisation, and finally, to monitor progress and see whether these principles have been effectively implemented.

Leadership needs to step up

When companies run into ethical breaches, it is usually not because employees deliberately want to hurt the trust of customers. It's simply not clear to them what the values of the company are and how they relate to their work, says Wassenaar. "Take the example of the department store: the data analytics department and the marketing team were probably just trying to do their jobs as well as possible. Nobody had instructed them to think about ethical aspects of their campaign, such as: How do our customers feel if we target them with advertisements for baby products if they didn't inform the company explicitly about their pregnancy? Should we have an age barrier with these promotions?"

Marketing and sales departments are usually instructed to reach their targets, which might incentivise them to utilise their company's data capabilities to the max, says Wassenaar. People working on technology are often driven by technological innovation and want to push the boundaries of what is possible. Wassenaar believes that this is where the leadership of a company needs to step up. "You need to have a clear vision of your values, communicate them deeply within your company and make sure your use of technology is aligned with these values."

What makes this issue more complex is that legislation is typically slow to catch up, certainly with rapidly developing technology like data analytics and AI. "It is not simply a compliance issue that you can leave to your legal department," Wassenaar says. "It is an ethics issue: What are your business principles? What are the boundaries you set for yourself?"

"Large technology companies are even formally warning investors of ethical uncertainties around AI"

Digital Ethics: a structural approach

Deloitte's Digital Ethics proposition helps organisations deal with ethical considerations in their use of technology in a structural way. "We take into account eight factors that are relevant for Digital Ethics: accountability, transparency, privacy, inclusiveness, bias awareness, informed consent, proportionality, and individual data control," Wassenaar explains. "We offer a list of possible steps organisations can take to get a grip on these complex issues."

First, Deloitte assesses a company's current situation with a 'Digital Ethics Quick Scan'. Based on this, an improvement plan is made to address immediate areas of attention. Deloitte can help to define the core principles of an organisation through vision and strategy sessions with senior leadership and by interviewing stakeholders inside and outside the company. A Digital Ethics Survey can explore how employees feel about these topics.

The next step is to establish governance frameworks to put these values into practice. This can involve launching awareness campaigns and training for the workforce. Other innovative options include e-learning, gamification or an ethics escape room in which people need to solve an ethical dilemma within a time limit. "It is important that a company's core principles do not end up somewhere in a brochure that no one will read," says Wassenaar, "but that they are truly understood and lived up to throughout the entire organisation – from the board room to the interns."

Deloitte can help to draft new policies and procedures, like an appeal procedure for customers who object to an AI-based decision. It is also possible to create new functions, such as an ethics department. "More and more companies have an ethics department that helps to develop training programmes, and that can investigate when there are reports of things going wrong," Wassenaar says. A crucial factor here is that this department should report directly to someone on the board, he adds. "Empowering the ethics function sends a strong message to the organisation that ethics matter, and it ensures that the ethics department can have an impact."

Finally, Deloitte can help to set up benchmarks to see whether the principles have been effectively implemented. "It makes sense to return after a couple of months or years to see whether the principles have been applied throughout the organisation, or whether there are some areas that are still pretty rogue," Wassenaar says. Moreover, since technology is developing rapidly, organisations might be confronted with new possibilities and ethical dilemmas that require new thinking about principles and related policies and procedures. "Embedding Digital Ethics in your organisation requires continuous attention."

Competitive advantage

Exploring potential ethical implications when deploying advanced technologies has several business advantages, says Wassenaar. First of all, it will help organisations avoid legal issues. Secondly, it is more efficient to think about ethics early on: "If you start thinking about digital ethics when you have a finished product, you might have to redevelop the entire product," says Wassenaar. "Thinking about these issues at an early stage helps you to get to the market faster and in line with your business principles."

Third, Digital Ethics will help build to create a strong reputation. "If customers do not feel comfortable using a service, if they feel it's too aggressive or insensitive, they seek alternatives," Wassenaar says. "On the other hand, having the customer's trust and loyalty ensures a long-term business profitability." The same goes for employees: a clear vision and strong reputation will attract the right workforce, which is an important business advantage.

Most importantly, embedding digital ethics will help to make a positive impact on society. Advanced technologies can be very impactful – both in a positive and a negative way, Wassenaar explains. "If you make sure you think of all ethical considerations in time and have all policies and procedures and a governance framework in place to live up to them, not only will you gain a competitive advantage and make profits, but you'll also create a positive impact on the world."

AI Driven Business Models

A strategic approach to capture the full potential of AI

Many companies have started to explore the potential of AI for their future business. Yet only a few have actually managed to capture the value in real-life business use cases. Where to start and how to configure for success?

In the last few years we have seen enormous and rapid developments in the world of Artificial Intelligence (AI), and it is hard to overestimate its potential. Big tech giants like Alphabet, AWS and are developing science fiction like applications in their labs; algorithms that teach themselves winning game strategies, can recognize human emotions or mimic a human conversation.
At the same time, some multinationals are implementing impressive new applications that help automate and robotize back-office processes, that can scan millions of knowledge heavy documents to find particular insights or automated intelligent chatbots in customer services processes. But besides the good news, we also see many AI initiatives getting stuck in proof of concepts and pilots, making a lot of companies struggle with the question where to start and how to scale.
"A lack of a sound vision and right prioritization causes a lot of AI projects to stall", says Naser Bakhshi, senior manager Artificial Intelligence at Deloitte. "Many initiatives start with a technology that sounds cool, without thinking how it really can make an impact on the organisational goals. They should start with forming a vision on AI that is aligned with the company's strategy, rather than just letting the one that shouts the loudest experiment freely."

AI Value Assessment

Deloitte has developed a proven methodology to facilitate the discussion on 'where to play' and 'how to win' with AI. "By taking a company's strategy and long term vision as a starting point, and look at the goals and aspirations and where AI can actually make an impact", explains Bakhshi. "One of our clients is acting in a highly competitive market, where margins are under pressure. They should focus on automating their processes, leveraging AI and Robotic Process Automation. Whereas another client is a highly specialized firm, depending on high-end expert knowledge. They benefit most from an AI-powered expert system that can process many unstructured documents."

The AI Value Assessment (AIVA) is an assessment, designed to assess the strategic themes, the existing processes in an organisation and unique data sources that can deliver tangible value by applying Cognitive and AI technologies. The AIVA follows a structured three step approach to test if generated ideas are desirable, feasible and viable for execution. Bakhshi: "It boils down to three simple questions: 'do we want this?', 'can we build it?' and 'does it make sense?' At the end of the AIVA you will have a thorough understanding of the exact cases that could benefit from AI and the related value."

"Getting the right inspiration is an important prerequisite in these discussions", says Stefan van Duin, Partner at Deloitte. "In so called 'art-of-the-possible' sessions we bring examples and ideas from all over the world. These examples may come from the same industry as our client's, but often the best ideas are coming from completely different industries." These inspirational sessions are helping to make the potential of AI more tangible. "Often AI stays very abstract," says Van Duin. "People may have expectations that the current technology just can't deliver, or they think too much in small incremental steps."

"It boils down to three simple questions: 'Do we want this?', 'Can we build it?' and 'Does it make sense?'"

Moonshot

The longlist of ideas needs to be evaluated against the potential impact. "When looking at an innovative idea for the application of AI, the big question is how it will move the needle," explains Van Duin. "We don't always have to look for a positive business case in terms of profitability. If a company has high ambitions in zero-footprint operations, an AI solution that helps reducing carbon emission may have a very high impact." Bakhshi adds: "Small and incremental improvements may be relatively easy to accomplish with AI, but we like to look at big impact: what is the 'moonshot' idea for your company?"

Moonshot ideas are transformational and radically change the way of working. They could potentially imply new business models or service offerings that could change the competitive landscape. An example is the US based insurance company Lemonade, who based their whole client interaction and processes on digital channels and AI, completely changing the way customers interact with their insurer.

AI initiatives should not be perceived in isolation but must connect to the long term strategy of the company and be part of digital transformation. As an example, think about "customer centricity" as strategic theme. Having a fact-based sense of your customers needs and how to serve them in a differentiated and optimal way, is very important but can be challenging. How could AI contribute to become a more customer centric company and improve service towards customers? This is a question that triggers thinking along the strategic priorities instead of taking technology as starting point. The answers, often use cases, will be then attached and valued against the strategic direction of the organisation.

Leadership enablement

"Another important dimension to capture the full value of AI is around leadership", says Jorg Schalekamp, Lead Partner for the Analytics practice in EMEA. At executive level there is not always a good understanding of what AI can deliver and how to scale and embed it in the organisation. "It is essential that the executive leaders understand the technology to the extent that they are comfortable taking (investment) decisions for it to implement. Helping leadership teams to understand the fundamentals of AI so they can have the right conversations with their teams is an important task that should be initiated from the very beginning of any AI initiative. It also takes an entrepreneurial mindset to invest in AI and drive it past the Proof of Concept phase into real adoption. Giving employees the room and support to work on such innovative projects, is key. They should feel supported and rewarded, also in the case of AI failure, but only if failure comes fast", explains Schalekamp.

Getting things done

Defining a vision and finding the high-impact ideas is one thing, but getting these ideas actually implemented proves to be highly challenging. Bakhshi: "Of course you need the technical capabilities to develop your idea into a workable solution; AI specialists, data specialists, engineers, designers et cetera. But it doesn't stop there. People may need to learn a new way of working. There may be new operating models or process redesign involved."

Many organizations make the mistake of keeping the AI innovations locked in a lab. But a successful approach requires a roadmap that links the strategy to people, processes, data and technology. Van Duin: "In our experience, to really make an impactful change, you need to be prepared to take radical steps. We supported one of our clients in a global training program, in which all management teams of all departments -we are talking hundreds of people- were given hands-on training in learning to work with these new technologies."

"A successful approach requires a roadmap that links the strategy to people, processes, data and technology"

Bakhshi adds: "We have found that the formula for success is to form a multidisciplinary project team with experts from Deloitte and the client to embed knowledge in the organization. AI driven transformation will impact many dimensions of the organisation. Therefore experts from different disciplines are required to ensure that a scalable, safe and valuable solution is implemented.

"Beside the technical profiles such as AI engineers and data scientists, you will need to think about factors such as GDPR, where privacy & risk expertise are required. Also change management specialists are needed to make the transition of workforce smooth and adoption of AI achievable. Moreover, strategy consultants are part of the team to assess value of the use-cases, define a roadmap for ROI and structure the new business models. Last but not least, legal professionals are essential to manage Intellectual Property and contracts with (external) data and technology vendors et cetera to prevent possible issues in legal space.

"To navigate through this complexity it is required to setup a so called Cognitive Control Tower, which ensures an aligned approach and methodology, cross leverages best practices and learnings, and brings in subject matter expertise as mentioned above were needed together."

Agility as a starting point

Deloitte has embedded a lot of elements of the agile methodology in their AI approach. This ensures for value based, relevant prioritization of tangible deliverables. Bakhshi: "Each phase ends with a clear cut-off point and a go/no-go decision for the next phase, allowing for the client to assess if they want to continue development. This way you can not only scale fast, but also fail fast if the idea turns out to be not as successful as originally thought. This is part of the job when working on extreme innovations."

If necessary, Deloitte helps in getting started. Bakhshi explains: "Deloitte's 'Asset Light approach' allows for the client to utilize Deloitte's assets as long as they need to postpone making large investments in tooling, hardware or people, until they are certain about the solution."

The really nice thing about Deloitte's Asset Light approach is that it enables companies to minimize the risk of losing money on long term commitments (e.g. licenses, hardware). Moreover, It also enables your company to explore different AI platforms and technologies before making a final decision for a preferred technology which can be a (private) cloud, on-prem or an hybrid model. Deloitte has partnerships with big tech firms (e.g. Amazon, Google, IBM, etc.) but also works closely with various niche players in the field of AI that develop many state-of-the-art AI solutions not necessarily available in the big-tech platforms.

No two organizations are the same

"We strongly believe in a tailored approach to specific situations," says Van Duin. "We have learned that the best approach is dependent on the ambition level of leadership, the current capabilities, the change readiness and the applicable rules and regulations. Therefore we always assess the situation and adapt our approach. After all, no two organizations are the same."

"We strongly believe in a tailored approach"

Contact

Netherlands contacts

Partner Tax & Legal – Frank Nan
Partner Consulting – Patrick Schunck
Partner Financial Advisory – Oscar Sneijders
Partner Risk Advisory – Marc Verdonk
Partner Audit & Assurance – Theo Jongeneel
Partner GSC – Liesbeth Mol
Partner Innovation – Richard Roovers

Contact Details

If you would like to learn more about Deloitte Innovation and our activities, please contact us via email NLInnovation@deloitte.nl

Colofon

Text: Eva de Valk by De Graaf & De Valk and Deloitte
Proofreading: Julia Gorodecky by Handblend
Design: Visual Heroes
Commissioned by: Deloitte Netherlands

Luxembourg contacts

Patrick Laurent
Partner - Technology & Innovation Leader
+352 451 454 170
palaurent@deloitte.lu

Jean-Pierre Maissin
Partner - EMEA FSI Analytics Leader
+352 451 452 834
jpmaissin@deloitte.lu

Nicolas Griedlich
Director - Technology & Enterprise Application
+352 451 454 052
ngriedlich@deloitte.lu

Nadia Andersen
Manager - Innovation
+352 451 453 817
naandersen@deloitte.lu

Deloitte.