KPMG

Trust in Artificial Intelligence

Transform your business with confidence

Contents

The era of AI is well and truly here – with huge implications for businesses across all sectors.

Many businesses are currently developing and operationalising Robotic Process Automation (RPA)[1] solutions and are beginning to experiment with true Artificial Intelligence (AI).[2] These are systems that can both interpret natural language and also learn to find the right answers without them having been programmed.

In their '*Hype Cycle for Emerging Technologies in 2017*' Gartner have identified that AI, as a transparently immersive experience and digital platform, is a trend that will enable businesses to survive and thrive in the digital economy over the next 5 to 10 years.[3]

This degree of innovation comes, however, with a heightened level of risk. Whilst traditional risk and control frameworks and IT process models can still help, we believe that there are new risks and different ways to control some of the existing risks.

Businesses urgently need to recognise this new risk profile and rethink their approach to the risks and controls relating to this technology in a structured way. They also need to ask 'what does it mean for my risk appetite?'[4]

This is essential for two main reasons:

1

The use of such advanced technologies will become material for many organisations, possibly sooner than anyone expects. When the time arrives it will not be possible to get the right controls in place overnight and have the capability to manage the risks effectively, or to provide assurance. Hence it is key for governance, risk and compliance practices and capabilities to develop alongside the evolution of the usage of such technologies.

2

AI will allow systems and businesses to become much more complex (to the point that it exceeds the capacity of the human mind to comprehend). The nature of this increased complexity is also selfperpetuating and although it might appear as simplification, it could well introduce 'technical debt'.[5]

Embedding controls in a system to mitigate technical debt after its implementation is typically far more costly than designing in the right controls at the start. Opportunities to build risk and control consideration by design will inevitably diminish over time and hence now is an optimal time to consider taking a positive and dynamic approach to building in control.

Entanglement:

"Machine learning systems mix signals together, entangling them and making isolation of improvements impossible". This is referred to as the CACE principle: Changing Anything Changes Everything.

Undeclared consumers:

"Without effective access controls, some of AI's consumers may be undeclared, silently using the output of a given AI instance or model as an input to another system. Undeclared consumers are expensive at best and dangerous at worst" as having them can impact relationships that are "unintended, poorly understood, and detrimental". Furthermore, "undeclared consumers may create hidden feedback loops".

Unstable data dependencies:

"Some input signals are unstable, meaning that they qualitatively or quantitatively change behaviour over time. This can happen implicitly, when the input signal comes from another machine learning model itself that updates over time". "It can also happen explicitly, when the engineering ownership of the input signal is separate from the engineering ownership of the model that consumes it". "This is dangerous because even "improvements" to input signals may have arbitrary detrimental effects in the consuming system".

Dealing with changes in the external world:

"One of the things that makes machine learning systems so fascinating is how they can interact directly with the external world. Experience has shown that the external world is rarely stable".

"We always overestimate the change that will occur in the next two years and underestimate the change that will occur in the next ten. Don't let yourself be lulled into inaction"[5]

– Bill Gates

Purpose of the paper

Although this paper looks at the subject through an Internal Audit lens, it is designed for anyone tasked with the safe delivery of AI.

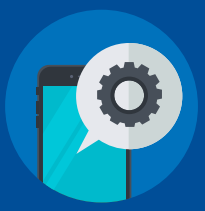
This includes:

Heads of Internal Audit and IT Internal Audit

Risk Managers

CIOs and their direct report

AI practitioners



Heads of Digital

Chief Information Security officers (CISOs)

How we produced this paper

Over the course of 2017, a number of professionals met in London across a series of workshops. The group was brought together by KPMG and consisted of IA professionals from five major UK corporate entities, across several industries, together with KPMG professionals drawn from several disciplines including IT Internal Audit, Technology Risk, Data Science/ Architecture, Data Analytics and Software Testing. Additional input was provided by a leading AI vendor and a renowned data science professor.

Based on our research, we concluded that whilst there were a number of instructive papers (for example *Hidden Technical Debt in Machine Learning Systems* by Sculley et al[5], and *A Model for Types and Levels of Human Interaction with Automation* by Parasuraman et al[6], there was no clear model or framework setting out the main risks and potential controls around effective use of AI.

We identified other publications that focused on the risks and controls of AI, but in our view they typically lacked the detail to allow them be used in a practical sense. Given the rapid and inevitable proliferation of such technologies, we decided to design a risk and control framework ourselves.

After considering AI-related activities in our own organisations, we created a high level risk list and more detailed set of risk statements – and defined outline controls pertinent to the risks. We then reviewed these for completeness and consistency and clustered these in a set of categories.

We went on to consider a number of widely used frameworks covering governance, standards and good practice, to identify a means of organising our thinking in a way that would be widely recognised and accepted. Amongst others, we considered COBIT, COSO, ISO27XXX, NIST, ITIL and TOGAF. We selected COBIT as it addresses both enterprise government and the governance of enterprise IT, which, arguably, is where AI is best located.

Terminology

As with all technology, particularly in the early stages of development, there is ambiguity in the language used. As a working group, we worked with Professor Mark Kennedy at Imperial College London to develop a catalogue (see Appendix) of AI and related terms, which provide context for the use of those in this paper.

Why does it matter?

Keeping expertise on board

AI systems will be conducting business processes – or elements of them. That means an organisation must be able to:

- Retain a way of managing without the AI system, in case it breaks down
- Re-perform or validate either the AI system or its components, in order to supervise and manage it. Thus demonstrating that it understands the outcomes the AI system produces, particularly if these are subject to regulatory scrutiny. Organisations can't blame an AI system for an error, or tell a regulator "it was the bot"![7]

The lack, or loss of, human involvement and expertise means that, in the worst case, no one will know how processes work and retention of expertise will become increasingly difficult.

Organisations will also need to consider the risks associated with dependence on third parties. In other words, how much more difficult it will be to exit from a provider when it not only runs infrastructure or hosts applications, but hosts AI which is learning and changing over time, potentially in a 'black box'. And equally, "who owns the intellectual property when a third party AI system has learned from your data?" Organisations need to think in terms of a system owner, a data owner, and a learning owner.

In different cases, the extent to which the business needs to understand the machine's decision making is different, and the approach to risk management should be tailored accordingly.

Supervising the systems

Companies have been dealing with immense transaction volumes for many years. However, while in the non-

AI world, a system always does what it has been programmed to do (subject to appropriate change control), this is not necessarily the case with machine learning capability. If the system does something different, how would you know whether or not the processing or the outcomes are still right?

Also, a move from many 'programmed' systems to fewer AI-enabled automated processes will inevitably make control harder, or the impact of a control failure more widespread.

This boils down to one key question: "how do you achieve effective human supervision of AI?" Can control velocity – the speed with which a control or suite of controls must operate – keep pace with risk velocity, which is the speed with which the risk materialises?

Avoiding unintended consequences

This topic warrants plenty of space in its own right. However, for the purposes of this paper, here's one example: an AI system may access data not envisaged by the system designer and, as a result, learns (or infers) something that it is both invalid and beyond correction.

A well-known example of this is Microsoft's Tay chatbot: Tay was an AI bot originally released by Microsoft via Twitter in March 2016. Controversy ensued when the bot began to post inflammatory and offensive tweets through its Twitter account, 'taught' by other Twitter users, forcing Microsoft to shut down the service only 16 hours after its launch. The problems were apparently caused by trolls who 'attacked' the service as the bot made replies based on its interactions with people on Twitter[8].

Dealing with unknown unknowns

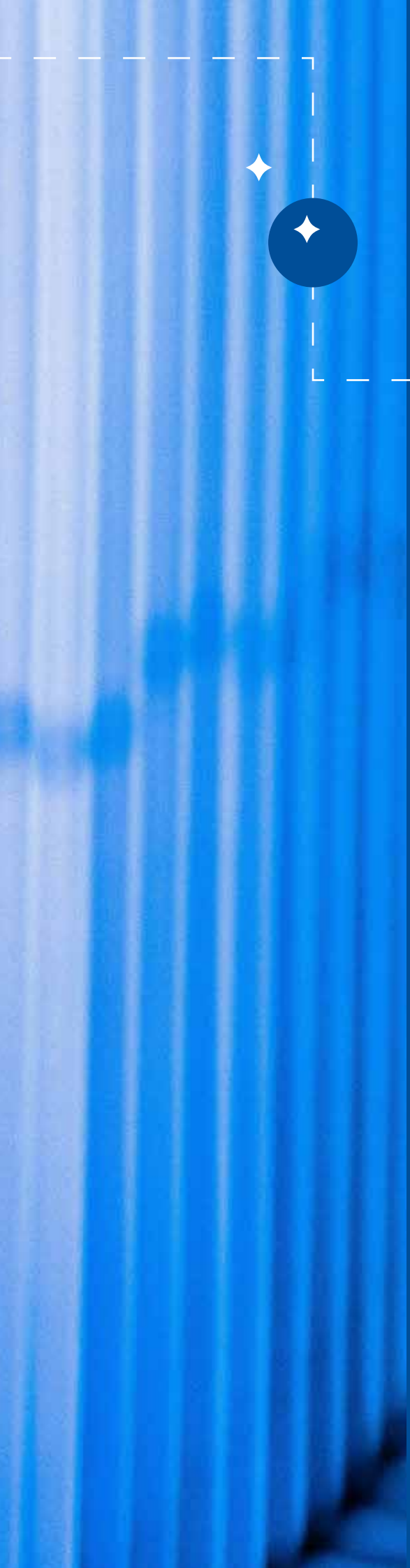
Just like people, a machine doesn't know what it doesn't know. Take a 'black swan' event – apparently unexpected and not predicted, yet which, once it's happened, has disproportionate impacts. Clearly such an event would mean the context in which a person or a machine makes decisions has changed.

A good example is the global financial crisis where systems continued to operate and make trades without amending behaviours until humans intervened. Again, where the humans do not have the capacity to intervene in time because of a lack of retained expertise, or because of lacking automated safety stops to prevent things evolving too fast for humans to cope, the outcome could be disastrous.

Validating the outcomes

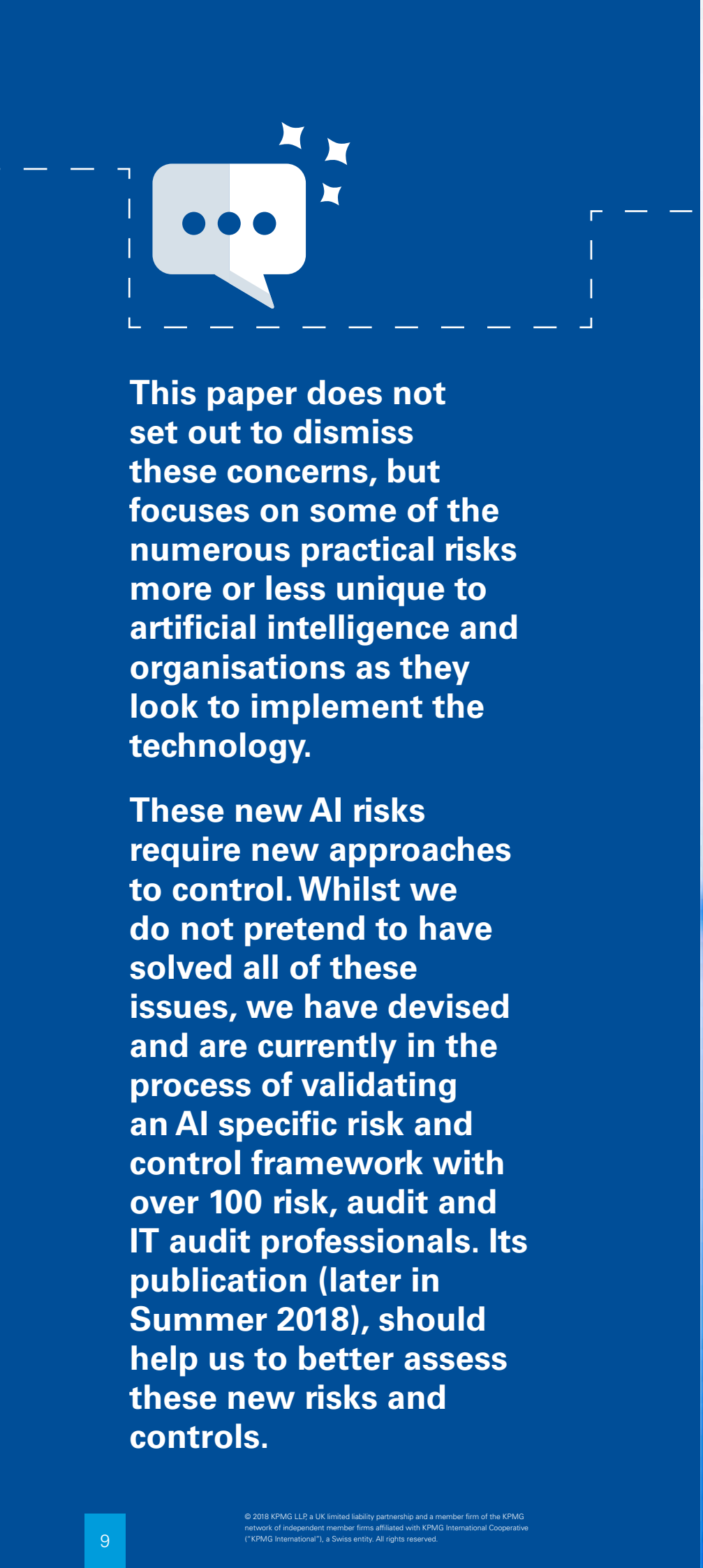
One of the major challenges for the audit process will be validation of outcomes or decisions made by an AI system. It's similar to the challenge faced by management of how to demonstrate to others that the AI's outcomes are correct and appropriate.

We've already considered the impact of a loss of expertise. Unchecked this could reduce the ability of organisations – and their auditors and other assurance providers-to validate the outputs from AI systems.



Expectations are high for the potential value of these new technologies – but, equally, some people also express scepticism and even fear about our ability as humans to manage the risks and assert control over the technology in the long term.

Witness, for example, the issues outlined by Elon Musk[9], and by Nick Bostrom in his book *Superintelligence*[10], raising the prospect that if AI develops into something that surpasses human brains, that may present significant threats to humanity. Many other commentators, such as the late Stephen Hawkins[11], have highlighted the social impact of AI, which could mean the loss of millions of jobs and incomes.



This paper does not set out to dismiss these concerns, but focuses on some of the numerous practical risks more or less unique to artificial intelligence and organisations as they look to implement the technology.

These new AI risks require new approaches to control. Whilst we do not pretend to have solved all of these issues, we have devised and are currently in the process of validating an AI specific risk and control framework with over 100 risk, audit and IT audit professionals. Its publication (later in Summer 2018), should help us to better assess these new risks and controls.

It could also become increasingly difficult to independently verify outputs or decisions made by an AI system if supporting data and/or calculation methods used at the time of the decision are not captured: for example, because capturing such 'evidence' continuously would require vast data storage.

Mastering the change process

Controls over system change have been a staple of IT control environments for decades. Without effective controls, the likelihood of erroneous code, system outages and successful cyber-attacks are high – as numerous examples demonstrate.

That said, many controls around how entities change their technology systems over time will prove obsolete for managing changes as a result of continuous self-learning capabilities of AI. That would be the case, for example, where changes are being made by the AI, e.g. to the weightings within the model that determine the answer that the model produces, without seeking or obtaining approval for making such changes.

Whilst we can envisage alternatives – such as AI suspending 'code release' until testing and approval has taken place – it may not be possible to prove that change has not taken place without approval, and it may be virtually impossible to conduct parallel running in an AI system. And how would you implement the traditional segregation of duties (SoD) between the development system and the live systems if the AI solution can deliver it all with no human intervention? Or would you no longer need SoD for non-human processing?

Building in your values

Many organisations have a set of values or an ethical code which guides their behaviour, creating a sense of shared identity, and defines what it stands for and how it operates. Even if this is not codified, every entity has a culture.

For instance, some businesses choose not to buy from certain vendors while others do, or e.g. some lenders provide mortgages to people that others don't. That is not down to having different quantities of data or smarter systems, but because one organisation has a different ethical or strategic view of what is right. Embedding such aspects within an AI solution in a sustained manner could prove challenging, especially because values are dynamic.

Dealing with the data

There are a number of thorny challenges around data. For example:

How can the organisation avoid their AI system taking on board the wrong learning, and how can it 'unlearn' data? If for example, a certain legal position was previously accepted as appropriate but no longer is, as the result of a specific case, you must find a way of excluding that data point from the AI's 'memory'. To some extent one can design solutions with such flexibility in mind, but that would only apply to aspects that people at that time consider to be a variable. E.g. to unlearn a bias that you may only realise years after the solution has been introduced is likely to prove to be a challenge.

How can the organisation be satisfied that the data feeding an AI system's learning and decision-making is complete and of a sufficiently high quality? There can be a measure of confidence and rigour around the selection and quality of the data used to train an AI system. Yet it is far more challenging to ensure that data used to learn over time – which may not be curated or controlled by the organisation – is also of an appropriate standard. The 'entanglement' topic mentioned earlier could also add infinite levels of complexity, after all, there could be a long chain of AI systems that feed in to one another where it becomes increasingly difficult to assess the quality of the ultimate input data.

Getting to grips with the challenge

So far, we have focused on the audit *of* AI, as that is the primary focus of this paper. However, it is important to acknowledge that an additional set of challenges exist in relation to auditing *with* AI.

There is a number of challenges for IA regarding audit with AI. Few IA functions have adopted technology to deliver audits; whether in data analytics, scripts or RPA tools, the IA profession in general has consistently been a late adopter. IA therefore faces the risk of a growing expectation gap between the business use of technology and IA's ability to keep in step with these developments. The pace of change, particularly in AI, will not slow down and it will be increasingly difficult for IA to catch-up, in skills, tools or methodology. If not addressed soon, IA functions may become less relevant.

In addition, automation may replace the bulk of compliance-focused work before too long, which might put pressure on the traditional 'three lines of defence' model.

In order to deal with such new developments, the first challenge is that the IA function may well be unaware of the AI activity currently under way within their organisation. Whilst most large businesses are rolling out RPA tools and many are already experimenting with AI already, IA must get involved and engage with the issue.

A second challenge is that, in some ways, AI is not a well-defined new major risk. It has a lot in common with people, in the sense that we can't always see what's going on in their 'thought processes and previous performance may not be a good indicator of future outcomes. In the same way as with people, where AI systems 'learn' from experience and amend their algorithms, flaws in the training data may result in poor or inconsistent performance outcomes. One look at today's IT environments shows that there are plenty of legacy systems already in use where we don't necessarily know when or why they go wrong – but they still do.

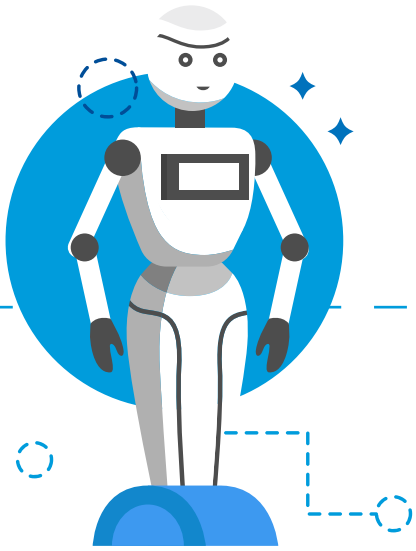
Another characteristic of AI that is not new is its all-pervasive nature. Just as IT is everywhere in our organisations, the same will be true of AI. And just as the risks relating to IT (such as badly managed change, unauthorised access, cyber-attack or poor business continuity planning) have the potential to affect most if not all of the entity, this will also apply to AI. However, where AI replaces humans, for reasons outlined above, a single mistake may have a much more profound impact, and potentially at such speed that significant damage might have occurred within an enterprise before the mistake is detected, stopped and corrected.

In a poll of more than 120 internal auditors in November 2016 and 2017, nearly half of delegates said that AI is already being used, to at least a limited degree, by their organisation, but:

> 80% were not confident about the governance in place around it.

> 45% of delegates in Nov 2016 planned to perform an audit on their AI solutions by 2018

> 70% admitted that they weren't clear what their audit approach would be.

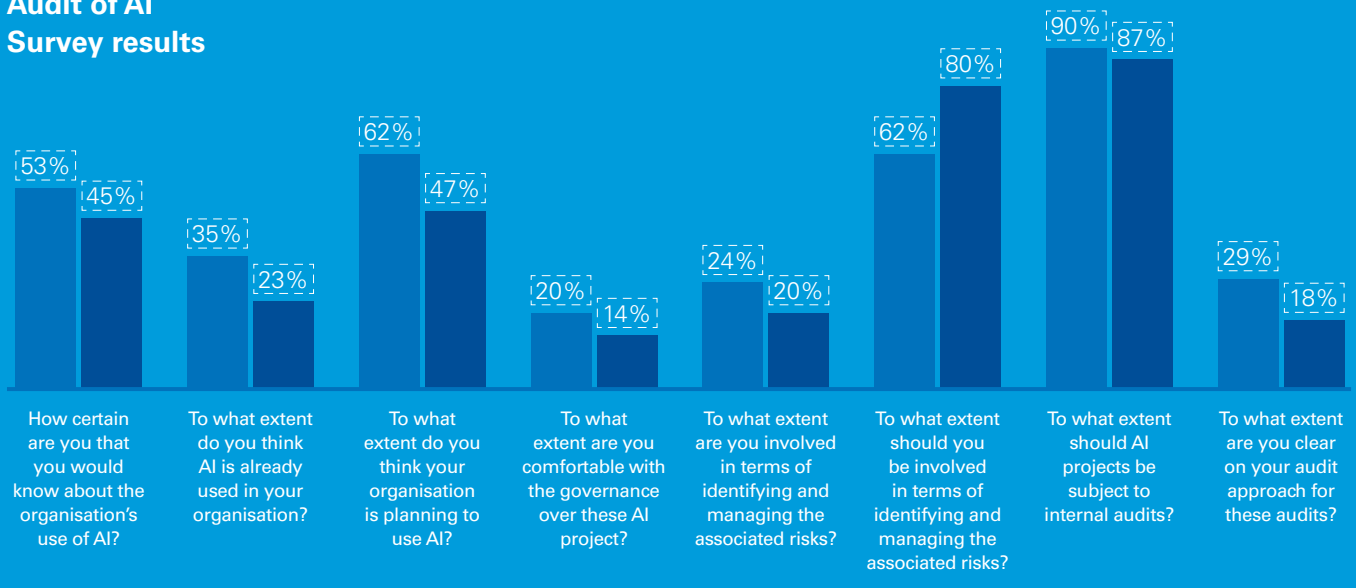


Throughout the history of technology in business, a common thread has been technology outpacing the risk and control requirements of the business. Many organisations have suffered financial and reputational damage through being unaware of, or responding late to, the risks associated with emerging technologies. We have seen this in the proliferation of ERP systems where security and control features often had to be retrofitted – say, to better manage segregation of duties and access control. We also saw this more recently with effective cyber security controls or, for example, the introduction of cloud and the subsequent design and implementation of 'service auditor reports' to provide assurance over outsourced control environments.

There is a permanent 'arms race' between those charged with providing protection and the hacking community focused on finding gaps in our 'armour'.

Recognising that this will be no different in the case of AI systems, we have developed a risk and control framework to guide entities, as they start to implement increasingly advanced AI systems, to proactively consider the risks and ask some important questions. While some of the risks and controls defined in this framework are new, many others build on existing risks and controls.

Nov 2017 vs Nov 2016
Audit of AI
Survey results



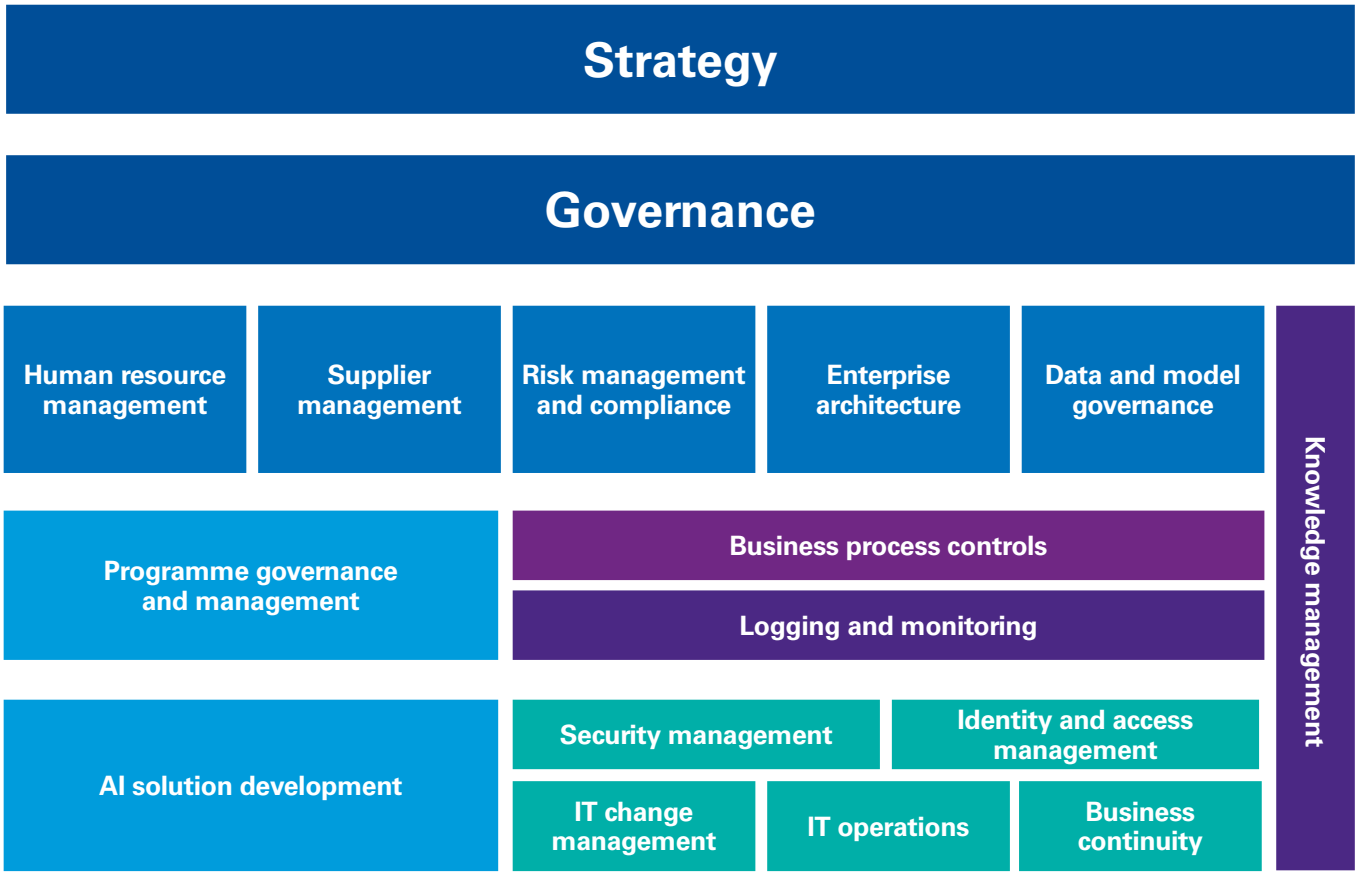
KPMG Audit of AI poll

At KPMG's IT Internal Audit conferences in Nov 2016 and 2017, we polled over 120 internal auditors on their awareness of and readiness for Artificial Intelligence. We asked several survey questions to assess IA's involvement with managing risks around their organisation's AI solutions. Around a third of the delegates said that AI is already being used, to at least a limited degree, by their organisation, albeit that only half were certain that they would know. However, over 80% were not confident about the governance in place around it. And whilst around a half were planning an audit of AI in 2017 or 2018, well over three quarters were not clear on how to approach this.

Risk and Control framework

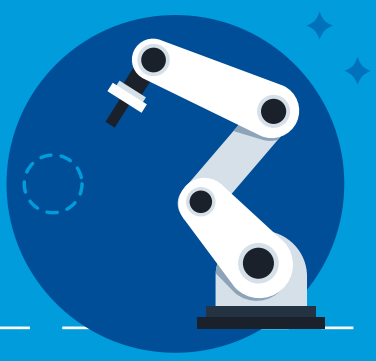
The risk and control framework we have developed is designed to help those tasked with the safe delivery of AI. An excerpt is set out in detail in the appendix. We have developed a risk and control framework specific to AI as a guide for IA professionals to use when confronted with the increasing use of AI in organisations and across different levels of maturity. However , the guide might also be helpful for AI practitioners.

We have categorised risks into seventeen areas:



AI Risk & Controls framework – 17 categories for managing risks and controls for AI solutions. Note: the framework will continue to be refined over time, so the number of risks and controls listed are likely to change of time as well.

Category	Number of risks defined in the framework	Number of controls defined in the framework
01 – Strategy	1	2
02 – Governance	7	8
03 – Human resource management	4	5
04 – Supplier management	4	4
05 – Risk management and compliance	2	6
06 – Enterprise architecture	6	7
07 – Data and model governance	7	10
08 – Programme governance and management	3	4
09 – AI development	10	16
10 – Business process controls	TBD for each solution	TBD for each solution
11 – Logging and monitoring	2	3
12 – Security management	5	5
13 – Identity and access management	7	13
14 – IT change management	4	5
15 – IT operations	7	7
16 – Business continuity	6	8
17 – Knowledge management	3	3
Total	**78**	**106**



Model extracts

Strategy (1) and Governance (2)

Key areas for focus are strategic alignment and enterprise-wide governance considerations; and corporate values. Some of the key issues here include:

Business objectives – As with the introduction of all new categories of systems, it is critical that the investment and outcome is aligned to the objectives of the business and its values. What makes this so important with AI is that, for example, processes relying on individual AI systems need to comply with all relevant firm policies, procedures and external regulations, where the system may 'learn' from data outside of the entity's direct control and behave in a way that may not be compliant with such policies and regulations.

Policies and standards – Impacted processes need to continue to comply, and be seen to continue to comply, with relevant regulations, policies and procedures, irrespective of whether a particular function is performed by a human or an AI system. Also existing policies should be reviewed and where necessary amended for AI considerations. Segregation of duties requirements need to consider the AI instance, the AI developer and the AI accountable owner and custodian.

Ownership – The AI human owner needs to be clearly identified. As the accountable party, the owner needs to have appropriate monitoring and supervision controls in place to prevent or override AI decisions.

Corporate values, culture and ethics – Organisations need to implement controls to align the AI's decisions to the firm's cultural and ethical values. The firm is accountable for a bad or incorrect decision, whether that decision is made by a human employee or an AI system. The firm's corporate values need to be incorporated during the design stage of the AI solution, and the AI-owner needs to supervise that the AI behaves in line with the corporate values. Where corporate values are principles-based, rather than rules-based, clearly allowable or non-allowable activities should be defined and reflected in the AI design and operation.

The AI logic, and the data used by the AI system will need to be periodically reviewed to confirm that AI decisions continue to adhere to the firm's corporate values.

Human resource management (3)

The main topics in this area are knowing what skills are required, across business and IT skills, either to build, run or maintain the AI solution, or to use its outputs. Key points for consideration:

Capacity of the right skills – Processes are in place to recruit, develop and retain human resources required for an effective AI enabled environment. This includes both AI-skilled people, e.g. to develop AI solutions, and non-AI-skilled resources, e.g. to retain business knowledge.

Skills retention – Without consideration of strategies to retain key technology skills, in particular coding and mathematics, there is a risk that entities are unable to design or review algorithms. In addition, without appropriate resources able to oversee and manage AI processes, there is a risk that AI systems may not operate effectively, may cease to be culturally aligned to the organisation or make inappropriate judgments. This echoes considerations in the strategy of enterprise above.

Data and model governance (7)

The main topics in this area are hypothesis management, data and commercial bias, completeness and accuracy, audit trail and logic validation. Key points for consideration:

Hypothesis Management – A set of measurements should be in place to identify the hypothesis dependencies and ensure that the risks of starting and delivering the project are managed.

Data and Commercial Bias – Controls exist to ensure that the status quo is not changed by bias introduced from historical data, which would be incorrect to extrapolate, and to reflect sensitivities when dealing with different cultural groups.

Completeness and Accuracy – Completeness, accuracy and source of the data sets used to support the decision-making process can be demonstrated and proven whilst still adhering to privacy regulations.

Logging and monitoring (11)

The main topics in this area are monitoring that the solution does what it should do, and not what it should not do, as well as provide continuous insight in the performance against Key Performance Indicators (KPIs) and Key Risk Indicators (KRIs). Key points for consideration:

Audit Trail – Decisions and changes made by the system can be independently validated through quality checks based on log generated data demonstrating provenance of decisions and changes, and can be used to support legal requests in case of litigations.

Logic Validation – KPIs and KRIs are in place, and have been initially validated to record the usage of AI decisions (its performance) in critical process with application of confidence levels, thresholds or range to make decisions. These KPIs are monitored.

Security management (12)

The main topics in this area are securing AI platforms, its data (learning, input, and output data) and algorithms. Key points for consideration:

Protection – The introduction of any new technology opens up new threat vectors which hackers will look to exploit: where the workforce is predominantly an AI solution, this can make entities very vulnerable. The threats will increase beyond the direct attack, to using malicious data to corrupt the AI learning and subsequent outputs. Understanding the end to end controls to ensure security and resilience will be critical, with consideration made for standalone and separate AI solutions so that they do not get contaminated with bad data.

Security – Where there is shared infrastructure such as IaaS or PaaS hosted AI systems, there is a risk of data leakage which may result in behaviour change not in line with the expectations or culture/values of the entity.

Cross-contamination – This is an existing risk that the AI solution may become corrupted based on the data it ingests. This is made more unpredictable by the potential for unplanned changes to how the AI operates due to unsupervised learning.

IT operations (15)

The main areas for consideration are around AI inventory and IT Service Management, including AI-generated incidents. Key points for consideration:

AI inventory – Just as with managing traditional IT, having a complete and accurate inventory of all AI assets will be critical for managing its risks. Similar to what we have already seen with 'shadow IT' caused by easy access to cloud based technologies, this might be a more difficult topic than one initially expects.

Incident and Problem Management – Typical ITIL processes – for example, incident detection or problem management – will likely be more complex and harder to manage. Processes may not be designed to identify AI-generated incidents such as misalignment with culture or minor errors in processing; by extension, problems may not be identified and managed because of a lack of identification or logging of incidents. This may result in outages or non-availability of systems or data, or information security breaches. For example, detecting that AI has turned off encryption in transit to fix a separate processing problem may result in identifying a problem that requires a different solution and/or result in amending the AI solution to reinstate encryption in transit.

Business continuity (16)

The main areas for consideration are around Resilience, Rollback, Business continuity planning and testing, and the ability to retrain the solution when required. Key points for consideration:

AI as a 'black box' – If the logic within the AI solution is not fully understood, it could impact the ability to recover services when issues occur, impacting business operations and resulting in financial loss or reputational damage.

Resilience – As the uptake of AI accelerates, entities will become increasingly dependent on AI and with that the need for AI solutions to be available 24/7. That is no different for organisations today; however, when systems are unavailable, often manual workarounds are available. If the workforce is completely automated - if resilience is not fully baked in and understood - then when there is an issue with the AI solution, you may no longer have a company at all, especially if it is managing critical end to end processes. As the AI uptake increases, so too will the resilience and availability requirements, which may increase costs and complexity.

Rollback – With continuously learning AI solutions, there will be times when organisations will want and or need to rollback self-made changes that have been made by the AI solution; for example when outcomes have become invalid or are outside of our risk appetite. It will therefore require detailed logging and reporting to be in place, supported by rollback functionality, to be able to reverse out changes.

Knowledge management (17)

The main areas for consideration are around cost, skills, security, AI as a 'black box' and impact on IT services.

Key points for consideration:

IT knowledge management and documentation – Due to an expected reduced number of human resources involved in a process, it becomes increasingly important to have complete and accurate documentation, including documentation of decisions and changes made by the AI itself.

Knowledge handover to 'sustain' – Specific individuals should be assigned to fulfil the 'BAU' / Sustain related roles. These individuals should have the right skills, have been handed over relevant knowledge from the Development team, and have been trained to fulfil their responsibilities.

Next steps

This framework represents an early attempt to provide a holistic approach to managing the risks around the use of AI, providing guidance to the audit community, amongst others. However, we need to continue to develop and mature our thinking.

We therefore invite fellow IA professionals or AI practitioners with an interest in this area to contact respectively Andrew Shefford (andrew.shefford@kpmg.co.uk) or Paul Holland (p.holland@kpmg.co.uk) for further information on how to contribute and participate in this project.

References

1 Robotic Process Automation (RPA) does not mean human-like physical machines, but automated tools to enable business automation; see Appendix for our proposed definitions

2 Artificial Intelligence (AI) refers to systems that demonstrate certain specific capabilities, in particular around learning capabilities and ability to handle natural language in a similar way to a human; see Appendix for our proposed definitions

3 https://www.gartner.com/newsroom/id/3784363

4 https://home.kpmg.com/uk/en/home/insights/2017/01/internal-audit-staying-in-step-with-new-technologies.html

5 https://pdfs.semanticscholar.org/1eb1/31a34fbb508a9dd8b646950c65901d6f1a5b.pdf

6 https://www.ida.liu.se/~729A71/Literature/Automation/Parasuraman,%20Sheridan,%20Wickens_2000.pdf

7 https://www.cio.com/article/3162239/leadership-management/banking-on-bots-the-move-towards-digital-labor-infinancial-services.html

8 https://en.wikipedia.org/wiki/Tay_(bot)

9 https://www.theguardian.com/technology/2017/sep/04/elon-musk-ai-third-world-war-vladimir-putin

10 https://en.wikipedia.org/wiki/Superintelligence:_Paths,_Dangers,_Strategies

11 http://www.bbc.co.uk/news/technology-30290540

Appendix: terminology

Artificial Intelligence, or AI

Artificial intelligence is the practice of employing advanced analytical techniques and algorithms to train computers how to use data from a wide variety of sources and formats to accelerate, automate, and augment decisions that drive growth and profitability[1].

Writ broadly, AI refers to the goal of machines that think like people and can therefore also work like people, or better, at many tasks. Inherently interdisciplinary in nature, AI has become a thriving field in data science that integrates ideas from philosophy, mathematics, statistics, linguistics, psychology and neuroscience. 'AI' is a concept mostly, often used as an umbrella term for advances in specific technologies and approaches such as Machine Learning and Natural Language Processing.

Strong/Wide and Weak/Narrow AI

Developing machines that truly think like people has proven much harder than AI's leading pioneers thought, so people now distinguish between strong and weak AI (also referred to as wide/ narrow AI) as the full breadth and flexibility of human-like reasoning versus systems that excel at narrowly-defined tasks that are hard to do even for human experts. As examples, weak AI systems excel at games like chess or Go, solving particular problems like identifying fraud detection, or recognizing particular features of images.

Augmented Intelligence

All AI augments human work and intelligence, so 'augmented intelligence' is mostly a redundant term, though sometimes used to emphasise the importance of the 'human in the loop'.

Expert Systems

Starting in the 1980s, computer scientists and software engineers built the first wave of commercially successful applications of weak AI were built by capturing the knowledge of experts as the rules and relations they use to make inferences and decisions in the course of complex tasks. Generally speaking, these systems featured inference engines that could reason logically using the rules and relations gathered by experts to emulate their work. Although some such systems excelled at tasks previously doable only by humans, the tasks were narrowly scoped and highly specialised, and such task-specific investments made it harder to get significant return on investment. Examples of expert systems included tools for medical diagnosis, loan approval, insurance underwriting, and the like. Expert systems are weak AI.

Machine Learning

As expert systems became practical, ambitious computer scientists began to develop a wide range of machine learning algorithms and conceptualise and build brain-inspired models of computation called neural networks. Unlike expert systems involved, machine learning solutions are trained, not programmed, e.g. to make decisions by recognising patterns.

Classification and Supervised learning

Most machine learning applications learn to classify inputs (stimuli or cases) as fitting or not fitting a pattern by exposure to training data sets in which each observation is pre-classified by humans as fitting the pattern or not. Patterns to be learned can be complex or simple. A memorable example of classification is the system that learns to tell the difference between Chihuahuas and blueberry muffins that look comically similar. Although that's an easy task for people, it's very hard to write a conventional computer program to do it by capturing the rules we use, as you would do in an expert system, to get the classification right. Classification is the term data scientists use to talk about learning to recognise when things fit a pattern, or not, and this kind of learning is called 'supervised' because the system is assisted, or supervised, through human pre-classifications. In the case of something like loan approvals, a training data set would include all the details of loan applications along with underwriter's decisions about whether the applicant's credit risk warranted approval and maybe also information about whether the loan performed or defaulted.

Classification and Statistics

Since classification is just a yes or no answer to whether a case fits a pattern or not, it is functionally equivalent to logistic regressions that predict predefined outcomes, with confidence levels, things like loan approval or denial, or whether a transaction is of no, low, medium, or high risk. To get accurate classifications with statistical methods like logistic regression, however, your data set needs to include the variables that matter to making the classification, or prediction. With logistic regression, the results will tell you how accurately the regression could match

[1] https://info.kpmg.us/artificial-intelligence.html

Data Reduction and Unsupervised Learning

Moving beyond supervised learning for classification tasks, unsupervised learning is the term for techniques that reduce complex datasets to potentially important patterns we may not yet know about or have names for. In the abstract, unsupervised learning involves recognising non-randomness in what might otherwise look like noise, and it is good for identifying patterns associated with things like risks or opportunities that have not yet been widely understood. Within this broad class of techniques, computer scientists have developed methods for anomaly and community detection to detect potentially important outliers or clusters. For the most part, the results of these methods are only useful once interpreted and explained by humans who know the domain the data are taken from. In business, community detection is useful for identifying, for example, customers with similarities that warrant recognising them as a distinct new customer segment. Anomaly detection is useful for detecting cases that fall outside norms, as would be the case, for example, with fraudulent transactions or cyber-attacks.

Data Reduction and Statistics

Data reduction is a little like what the brain does when you blur your eyes to ignore some details and bring others into focus, it is also functionally equivalent to statistical methods for grouping variables that move together, or co-vary, across cases. These methods are useful for reducing a larger set of variables to a smaller set of factors that capture truly different dimensions of what's going on in the data. Within data reduction, factor and cluster analysis are major families of methods, and there are many algorithms for specialised factor and cluster analysis. In data science, statistics, and machine learning, experts are comfortable using many models, and the best results come from knowing which specific method works best for the data set and problem you have. For complex problems, researchers experiment with different methods to learn which ones work best, and breakthroughs often come from using this learning to devise new algorithms for problem or class of problems that existing methods do not address well.

Data Engineering

One of the main challenges of machine learning is assembling the data sets for supervised and unsupervised learning. To get good results for tasks like classification or anomaly detection, one needs data sets that include all the variables that matter. Just as it is with humans performing these tasks, creative curation and analysis of inputs, or variables, is a major factor in outperforming the competition. In business tasks, this increasingly means going beyond traditional structured data to include unstructured data as well.

NLP (Natural Language Processing) and NLG (Natural Language Generation)

Increasingly, the unstructured text of media coverage and social media posts are proving valuable additions to data sets for supervised and unsupervised machine learning. NLP / NLG is the subfield of computer science in which researchers and engineers focus on developing tools for using natural language (text or speech) both as a data source for tasks and as a user interface to or from computers doing various kinds of work. Thus, such techniques provide data for tasks and a means of communication about tasks. Historically, computer science mostly focused less on natural languages and more on formal languages like logic and the various languages for writing computer software. Formal languages are good for precise descriptions of data and procedures, and they are easier for computers to take in because they generally follow very strict rules for syntax and semantics. Whereas formal languages require people to read and write code in languages that come and go, the rise of natural language techniques offers a path to interacting with computers that makes them more accessible to people, and vice versa.

Robotic Process Automation

RPA is the application of technology that allows employees in a company to configure computer software or a 'digital robot' to capture and interpret existing applications for processing a transaction, manipulating data, triggering responses and communicating with other digital systems. So this is about software (as opposed to the physical machines the word robot usually calls to mind) and administrative or service industry tasks as opposed to manufacturing work. Typically, RPA is used to refer to more advanced technologies that can be used to automate (parts of) processes where that was previously not possible, while still being coded/ configured and do not include 'intelligence' such as machine learning.

Cognitive

Started as an IBM buzzword but the terms Cognitive automation and Cognitive computing have received wider usage. These terms appear less defined than e.g. machine learning or natural language processing and are typically used to group technologies at a less holistic level as the term AI does. Example definitions:

- Cognitive automation leverages different algorithms and technology approaches such as natural language processing, text analytics and data mining, semantic technology and machine learning [2]
- Cognitive computing makes a new class of problems computable. It addresses complex situations that are characterized by ambiguity and uncertainty; in other words it handles human kinds of problems. The cognitive computing system offers a synthesis not just of information sources but of influences, contexts, and insights. To do this, systems often need to weigh conflicting evidence and suggest an answer that is "best" rather than "right"[3]

With thanks to the many KPMG contributors and to Rafael Bambino, Fayyaz Cheema, Mark Kennedy, Thomas Nowacki, Paul Thomas and others for their contributions to this report.

[2] http://www.expertsystem.com/what-is-cognitive-automation/

[3] https://cognitivecomputingconsortium.com/definition-of-cognitive-computing/

[4] https://irpaai.com/what-is-robotic-process-automation/

Contacts

Andrew Shefford
Global Head of IT Internal Audit
KPMG in the UK

T: +44 (0)20 7694 5507
E: andrew.shefford@kpmg.co.uk

Create | CRT096604