SLAUGHTER AND MAY / ASi DATA SCIENCE

Superhuman Resources

Responsible deployment of AI in business

Foreword

Since the turn of the century, there has been an exponential increase in the creation and management of data. Technologies that can sift, analyse and deploy this growing wealth of digital information are being rapidly adopted across government and business, as well as by consumers themselves. Artificial intelligence now promises to transform the application of these resources.

The effect will be highly disruptive to existing ways of doing things, but promises huge opportunities for consumers and for the entrepreneurs that realise their dreams. The argument that artificial intelligence will cause mass unemployment is as unpersuasive as the argument that threshing machines, machine tools, dishwashers or computers would cause mass unemployment. "The bogeyman of automation consumes worrying capacity that should be saved for real problems," said the economist Herbert Simon in the 1960s.

None the less, genuine concerns about the misuse of artificial intelligence, and the unintended consequences, need to be taken very seriously. I am therefore delighted to see this paper published because it is critical that the risks and liabilities associated with AI adoption are equally well recognised as businesses increasingly expose machine learning applications to their clients and customers, and the public at large.

Matt Ridley, *Times* columnist, author of *The Rational Optimist* and member of the House of Lords.

"businesses are
coming to realise
that there are
some unique risks"

Background

Great promise, but strings attached

Advanced artificial intelligence (AI) systems are already being used to enhance our lives and to transform the way businesses operate. Gains in both computational capacity, and our understanding of how to exploit that capacity, mean that a form of "general artificial intelligence" – a truly cognitive system – could be created within our lifetimes. This will be revolutionary in a way few can presently imagine.

While the holy grail of general AI is still just out of reach, right now businesses across a broad spectrum of industries are exploring the potential efficiency gains offered by AI systems when applied to specific tasks. Understanding AI's potential, and how to exploit it, is no longer the preserve of an elite cadre of data science academics and engineers; AI and machine learning tools can now readily be developed and deployed by public bodies and by private businesses and entrepreneurs. The use of AI systems is already widespread in areas such as transport, finance, defence, social security, education, law and order, public safety and healthcare.

Instances of businesses exploring and adopting AI systems are increasing exponentially. Whilst this comes with clear upsides, businesses are also coming to realise that there are some unique risks associated with these systems and technologies. We believe they are risks which to date have been underappreciated and in many cases unaddressed.

The costs of getting AI implementation wrong could be great – and this could include human, social and political costs as much as economic costs: organisations risk meaningful losses, fines and reputational damage if the use of AI results, for example, in unintended discrimination, misselling or breach of privacy.

It is thus increasingly critical for boards and leaders to consider carefully not only how adoption of technology such as AI can deliver efficiencies and cost savings, but also to consider carefully how the associated risks can be managed properly. In an AI context in particular, this is undoubtedly a challenge for managers who typically will not yet have the knowledge and tools available to evaluate the size and shape of the risks in an AI system, let alone to manage them effectively. The absence of best practices or industry standards also makes it hard to benchmark what the safe and responsible use of AI looks like in a business environment.

We are keen advocates of AI and we believe it has truly transformative potential in both the public and private sector. We would like to see organisations take a thoughtful and responsible approach to their implementation of AI systems. Unfortunately, while much airtime has been given to the potential benefits of AI technologies, there has yet to be significant attention devoted to the risks and potential vulnerabilities of AI.

This paper, published jointly by ASI Data Science and Slaughter and May, brings together technical and legal expertise to provide a unique analysis of this topic. We recognise that where businesses fail to lead the way in developing best practice, this may give rise to regulatory and governmental responses that cannot but function as blunt tools in the face of AI across such a variety of sectors. Therefore, to avoid this pitfall and help businesses retain their ability to exploit AI to the full extent, we suggest a suite of concrete, practical principles to assist businesses and other organisations as they responsibly explore, create and deploy advanced artificial intelligence systems.

Key terms

Artificial Intelligence (AI)

AI can be hard to define. Alan Turing described it as “the science of making computers do things that require intelligence when done by humans”. It is important to remember that, notwithstanding any mystery, artificial intelligence is (for now at least) a form of human-developed software running on human-designed hardware, executing a series of human-originated commands.

General v Narrow Artificial Intelligence

We subdivide the field of AI into two categories: **General Artificial Intelligence** and **Narrow Artificial Intelligence**. General Artificial Intelligence refers to AI that has such broad applicability that it could successfully perform any task or solve any problem requiring human intelligence. Narrow Artificial Intelligence refers to algorithms that are designed to solve one particular problem, for example chess-playing algorithms. The distinction between Narrow and General AI is a continuous spectrum rather than binary – some algorithms can be more general than others while still not being fully general (for example DeepMind’s DQN algorithm that played multiple different video games at super-human performance). No truly General Artificial Intelligence – or more accurately Artificial General Intelligence – has yet been created, and expert estimates for achieving that ultimate goal range from 10 to 100 years, although consensus seems to be shifting towards the closer side of such a window.

Machine Learning

Machine Learning is a subset of the wider field of AI. It is improvements in machine learning that have driven the remarkable progress in AI performance over the last 15 years. Machine learning refers to algorithms where the performance of a task improves with experience. Imagine a chess-playing algorithm; we can refer to the algorithm as a narrow AI if it is able to play chess effectively. If the performance of the AI improves as it plays more games, it is also a machine learning algorithm. Machine learning algorithms work by learning from data patterns, and are often contrasted with **Expert Systems**, which are computer programmes that simply follow rules explicitly pre-programmed by humans. For all their power, machine learning algorithms are not perfect. Because they excel at tasks that involve detection of subtle patterns or correlations in large datasets, they can sometimes be hard to understand, interpret and audit. This is particularly true of the more sophisticated algorithms, such as artificial neural networks, where decision-making can be so complex as to become essentially opaque.

Static v dynamic machine learning

A dynamic machine learning system is one that continues to learn and develop its model in real-time based on the data it is exposed to, whereas a static machine learning system is one that is trained with a dataset, but then operates statically in production so that it cannot continue to develop and refine its operation on the basis of new datasets. A static machine learning system may of course intermittently have its model updated with more or better training data.

Supervised v unsupervised machine learning

Supervised machine learning involves training an algorithm with data consisting of an input and a corresponding output, where a human teacher has confirmed that the input corresponds to the particular output. For example, an algorithm might be trained to recognise pictures of cats and dogs by being provided a series of pictures of cats and being told these are cats and then a series of pictures of dogs and being told these are dogs. By contrast, unsupervised learning involves an algorithm attempting to discern a model based on input only training data. For example, an algorithm could be devised so that if it were provided with a series of pictures of cats and dogs, it would seek to discern similarities and differences so as to be able to group separately all of the cats and all of the dogs.

Development of machine learning

From rapid to exponential

Before the 1980s, machine learning was considered to be at the fringes of the field of artificial intelligence. It was even debated whether learning would be a necessary feature of artificial general intelligence at all. Only as computational power grew, and more sophisticated algorithms were developed, did it become evident that learning systems could solve impactful problems.

Arthur Samuel's creation in 1959 of an algorithm that learned to play checkers is often identified as the first in the lineage of modern machine learning, although the roots of the subject can be traced back beyond this, through Alan Turing to the fathers of statistics and probability. The term "machine learning" was coined by Arthur Samuel in his 1959 paper titled *Some Studies in Machine Learning Using the Game of Checkers* and by 1962, Samuel's learning algorithm was able to beat the Connecticut State checkers champion, the fourth best player in the US.

Another pioneer of machine learning was Frank Rosenblatt, who invented the 'perceptron' in 1957. The perceptron was the first design of what we now call an 'artificial neuron', the individual entity that when connected together with others forms an artificial neural network. The work led to the creation of a machine for the US navy that could learn to spot patterns through experience. Rosenblatt's work fell out of fashion when it was demonstrated that there were simple problems for which it could never learn to distinguish the correct answer. It was later found that by tweaking the perceptron design, creating networks and using different training algorithms (essentially, developing modern artificial neural networks) these problems could be solved. Breakthroughs were also made in the 1960s around the architectures – the wiring diagrams – of neural networks which reduced the time required for them to train.

There were then other powerful classes of algorithm that were developed in the late 20th Century, including clustering algorithms, decision trees and random forests, and support vector machines, offering a broader variety of tools for data scientists. Often, algorithm choice involves balancing different demands, for instance, balancing the need for predictive accuracy with a desire for transparency, i.e. should the answer be 'black box' (not easily interpretable by a human) or 'white box' (human interpretable). Other trade-offs involve reducing training time (and computational expense) at the expense of predictive accuracy.

The recent excitement around deep learning – the name for a particular set of architectures of artificial neural networks – started around 2005, when the work of people like Hinton, Bengio and LeCun started to show human level performance in the task of image recognition, something that computers had struggled with historically. This was followed by surprising performance leaps in a variety of areas including tasks as diverse as speech recognition and playing computer games. Generally, the feeling is that deep learning will, at some point, allow computers to surpass humans in most perceptual tasks, and allow the automation of those elements of any job where such skills are required.

Why now for AI?

The recent explosion of machine learning technology is really a product of two things: tremendous increases in computational power and enormous volumes of accumulated data. The cost of performance at this level has also dropped dramatically.

We expect the ability of computers to continue to grow. Fundamentally, the human brain is a computer on a biological substrate, and so ultimately we should not expect there to be any tasks performed by humans that remain outside of the capability of computers. This looming breakthrough may be closer than intuition suggests. The World Economic Forum reported in its Global Risks Report for 2017 that global investment in AI start-ups has risen astronomically from USD282 million in 2011 to just short of USD2.4 billion in 2015. Figures published recently by Bank of America Merrill Lynch suggest that the global market for AI-based systems will reach a value of USD153 billion by 2020; more money will be invested into AI research in the next decade than has been invested in the entire history of the field to this point.

Global investment in AI start-ups

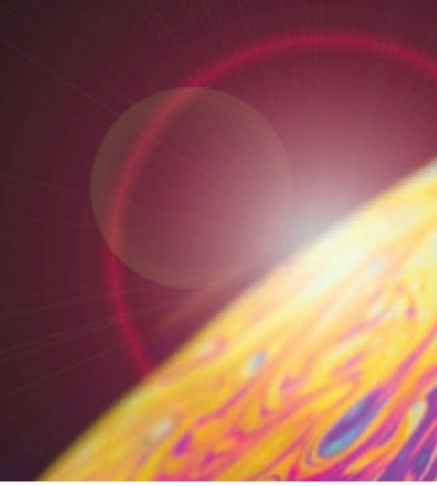
$153bn	**2020** ▲
$ 2.4bn	**2015** ▲
$ 0.3bn	**2011** ▲

Source: WEF / BAML

If we could plot a graph to show the rate at which AI technology has developed over its relatively modern existence as a field of scientific endeavour, we would see a hyperbolic curve; it seems highly likely that we are now entering the near vertical phase of that curve. Google and Nvidia, for instance, have both recently announced special purpose processors for AI that are capable of processing at tremendous speeds, massively faster than anything seen to date.

Transformative potential

Much to play for...



The technological revolution has seen organisations automate repetitive, high volume, sometimes complex but typically rule-based (“if X then Y”) processes, and has delivered incredible increases in efficiency and productivity. The immense impact that technology has had, and continues to have, in the world has transformed the human experience. Cast in this light, AI – a system that can simulate human cognitive processes – has the potential to generate efficiency advances at a multiple rate of anything we have experienced to date.

There are some more evident aspects of AI that indicate its enormous potential:

1. **Ability to synthesise large volumes of complex data quickly**
   With the exponential growth in the volume of data available to organisations it is becoming increasingly difficult to use conventional means of analysis to fully exploit the value in that data. Machine learning systems can synthesise tremendous amounts of data to develop complex models that are able to realise practical value at far greater speeds than any human, or indeed team of humans.

   This has relevance in at least two situations. First, where existing decision-making processes depend on the knowledge and experience of skilled professionals who draw on banks of knowledge accumulated over many years; and second, where decision-making can only take place (or may become more accurate or reliable) after analysis of a significant volume of data points. AI has the potential to substitute for the first and accelerate the second.

   It can take many years to achieve mastery in specialised professional fields of knowledge – actuarial science, medicine, law. Yet machine processing of complex non-linear relationships in data can now be achieved with levels of efficiency and reliability that mean AI can begin to make insurance underwriting decisions, discover new medicines and carry out due diligence on complex businesses in fractions of the traditional timescales for high-skilled professionals.

2. **Adaptability and scalability**
   Machine learning systems often learn to perform well at their task by discovering some truly important but obscured features of a large data set. A striking recent example is Google Translate’s use of a single neural network for translating between languages. The system was apparently able to perform well even when translating between two entirely unfamiliar languages (languages the system had never seen translations between), suggesting that it had learned not only the languages involved but had also devised itself a process for learning the language translation process.

   The ability to solve a new problem by learning from a related task is indicative of how adaptive machine learning systems can become, and is a significant step on the path towards achieving forms of artificial general intelligence. From seemingly basic inputs and processes AI has the potential to generate hugely scalable solutions.

   Perhaps the greatest promise of machine learning so far comes not from the specific tasks that AI systems can now perform but from the promise of those systems developing new skills and applications from the knowledge and processes they have already learned. Such advances have already begun: DeepMind has been able to repurpose a system originally designed to optimise performance when playing strategy games so as to optimise the cooling of Google’s massive data centres – an adaption that will generate

meaningful financial savings. In another example of adaption, an AI circuit was shown to have learned to keep time by converting itself into a radio that picked up the recurring radio frequency of a nearby PC.

3. **Autonomy**
The sophistication of tasks undertaken without human oversight has seen a step-change with the advent of machine learning systems. Previous "expert systems" relied on logic that performed poorly in novel situations for which the system had not been prepared, and other autonomous systems depended on simple sensors and constrained environments.

But because modern AI is capable of learning and accumulating knowledge without human supervision, it promises a more reliable future autonomy. Autonomous cars are one of the more conspicuous deployments of AI at the moment, and significantly more complex than non-manned trains with which we have lived for some years; yet have proven to be safer so far (admittedly based on relatively limited exposure to the real world) than human drivers.

4. **Creativity**
There is some ineffable aspect of creativity that we struggle to ascribe to machines. Yet when confronted with the art, writing and designs that machines have produced it is becoming evident that they are capable of novel and sometimes impressive creations. For example, there are hundreds of new applications stemming from computer vision, some of which are capable of producing machine art: original creative images in an environment free from human supervision.

Similar creativity as applied to language is now giving machines the ability to write both descriptive and creative prose: news outlets Yahoo and Reuters have already published machine-written articles. When combined with systems that can understand and interrogate speech, so-called chatbot systems are able to sustain a level of conversation sufficient to provide viable call-centre services. Another example is JukeDeck, a London-based AI group using AI to produce music tailored to the listener or use – in essence a purely artistic creative application of AI.

Machine learning systems are especially effective at learning how to maximise performance through creativity. This feature of AI has already been exploited in design processes to reimagine and optimise, for example, the construction of light-weight components for bicycle frames or the structure of an automotive engine block.

5. **Consistency and reliability**
In addition, machine learning creates opportunities to achieve greater consistency and reliability in processes that at present depend heavily on human judgment. Reliable, time-pressured decision-making systems could have a material beneficial impact in a diversity of sectors: in transport, accidents could be avoided with more rational split second avoidance decisions; in medicine, A&E patients could be more rapidly diagnosed on the basis of a rapid synthesis of data points; and in disaster situations, more rapid and thought-through safety recommendations could be made in the immediate aftermath of natural disasters or terrorist attacks.

These opportunities come from the ability of machine learning systems to cut through complexity while exploiting the speed and reliability of machine software relative to the performance of humans in pressurised situations. Outside of high pressure environments, one can equally see potential for material gains, for example for investment firms making discretionary financial recommendations, where inconsistencies in human decision-making (whatever the cause – fatigue, 'document blindness' or straightforward negligence) can give rise to costly outcomes: penalties and remediation.

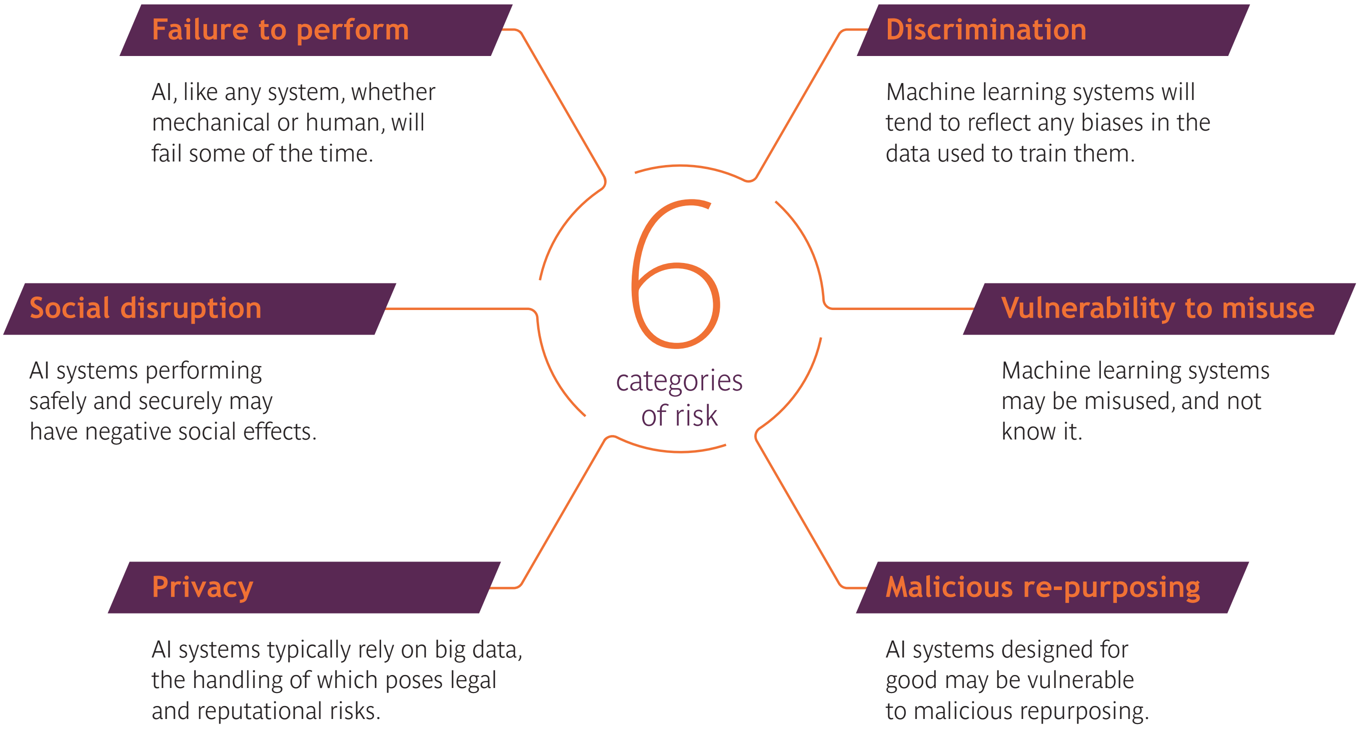
Deployment risks

…still plenty to lose

As the potential benefits multiply, so too do the risks. There are certainly risks associated with reliance on machine learning systems, as well as some acute moral and ethical considerations.

Commentary on, and indeed investment in, AI and machine learning seems so far to have focussed on the potential upsides. This is understandable, but misses a critical point: AI can only ever be useful if it can be deployed responsibly and safely. Safe means either no harm or an acceptable risk of harm: civil engineers do not design a bridge and then later contemplate what safety features should be implemented; safe deployment is critical to the design process. The same reasoning applies to AI.

To manage risk, you first need to understand it. We have identified 6 categories of risk that are particularly acute for, and should be top of mind for, the responsible design and deployment of AI systems:



6 categories of risk

Failure to perform

AI, like any system, whether mechanical or human, will fail some of the time.

Discrimination

Machine learning systems will tend to reflect any biases in the data used to train them.

Social disruption

AI systems performing safely and securely may have negative social effects.

Vulnerability to misuse

Machine learning systems may be misused, and not know it.

Privacy

AI systems typically rely on big data, the handling of which poses legal and reputational risks.

Malicious re-purposing

AI systems designed for good may be vulnerable to malicious repurposing.

1

Failure to perform

AI, like any system, whether mechanical or human, will fail some of the time. There can be any number of cases of failure, ranging from a typo in source code to a fundamental flaw in the overall design of a system. The following are three common causes of failure for AI systems:

A. Bad design

AI systems often pose novel design challenges, even for experienced scientists and expert software engineers. Machine learning systems are fed examples and autonomously learn a model that generates outputs from input data. This makes it much harder to predict how the system will behave in practice than for other software where the model is prepared in advance by a human. Traditional software is deterministic, whereas machine learning systems are probabilistic and can learn surprising strategies to perform well at a task.

For example, an AI call-centre service could be trained to communicate with customers and be given the goal of minimising the number of complaint calls. Expected to learn the most effective strategies for dealing with customer problems, the system will in fact learn any strategy that more effectively reduces calls, such as keeping the line active to prevent in-coming calls, or learning that communicating rudely reduces repeat callers.

Whereas in traditional software some common sense can be explicitly encoded and its limitations made clear, the flexibility of AI to learn sophisticated behaviour can lead us to forget that it need not learn common sense solutions to the problems it is tasked with solving.

These problems are compounded when a machine learning system operates dynamically, continually updating its model in real-time as it processes data, or when the model is difficult to interpret, as with deep neural networks. While any system can fail because it is poorly designed, the fact that it is particularly difficult to predict how an AI system will perform in practice means it can be hard to spot that the system has been badly designed until it fails.

B. Bad data

AI systems are only as good as the data they are designed to handle. One of the great benefits of machine learning algorithms is that they can be trained to handle a wide variety of inputs by using large training data sets. Problems arise where the training data set contains not enough data, not the right data, not enough real-world data, incorrect data or data that is skewed/biased in a way that does not reflect the environment in which the system will be operating. How accurately a machine learning system will respond to a given input is directly related to how similar that input is to what it has seen before.

A system trained with a limited data set will usually encounter more inputs that are unfamiliar and will therefore perform poorly compared to a system trained with a broader and more diverse data set. For example, take a photo app that has been trained on a large variety of animal faces, but a limited number of human faces not representative of the diversity of photos it would see in practice. When deployed to automatically label images, this could lead to the embarrassing and potentially insulting mislabelling of people as animals.

C. Bad application

With ordinary software it is usually quite easy to know what the system can and cannot do. It is relatively simple to look at the algorithm and determine that the system will handle these inputs, but not those inputs. With machine learning systems, whether a system will handle certain inputs is a product of both the algorithm and its training data. Therefore, the reasoning process or model employed by an AI system is often much more opaque than that of conventional software. This also means that AI systems can operate in a way that is much less predictable than other types of software. This lack of transparency and of predictability does not necessarily mean the system is less reliable or less accurate. Indeed, many innovative applications of AI technology The capabilities and limitations of conventional software are typically quite evident. With machine learning systems, on the other hand, whether a system will handle certain inputs appropriately is a product of both the algorithm and its training data. Thus the model employed by an AI system is often much more opaque, and its operations less predictable, than conventional software. This lack of transparency and predictability does not necessarily mean the system is less reliable or less accurate. Indeed, many innovative applications of AI technology in medical diagnosis and autonomous piloting of vehicles are proving to be significantly more reliable than most humans performing the same tasks. It does, however, mean the task of determining where and when it is appropriate to deploy a particular AI solution becomes more of a challenge.

To decide to deploy any system you first need to understand both the environment in which the system is being deployed and the level of risk you are comfortable with in that environment. Second, you need to know what the system does and how well it does it. This ultimately is for you to decide: yes, the solution will work in my environment and yes, I am comfortable with the risks.

It is the second question – what the system does and how well it does it – that is often trickier to answer for AI systems than for conventional software. The relative lack of transparency and of predictability makes assessing both the types of possible failure and the risk of that failure occurring, more difficult. For this reason, where risk tolerance is low, extensive testing is typically necessary. In medical diagnosis, for example, extensive testing is required and a threshold for confidence from the machine has to be set.

It is also important to test for unexpected scenarios. For example, if a system is diagnosing using medical scans, it needs to respond appropriately, perhaps with an alert, if shown the wrong type of scan on which it has not been trained to make decisions. Better systems should be capable of recognizing 'unexpected' inputs.

Discrimination

Closely related to the problem of biased training data, a particular concern in the context of AI systems can be discrimination. As noted above, machine learning systems will tend to reflect any biases in the data used to train them. This can mean subtle discriminatory biases are not always immediately apparent.

Take the topical example of a machine learning system assigning recidivism risk scores to an individual (Durham constabulary have announced a trial of such an AI system). A system like this may take into consideration the number of previous interactions the relevant individual has had with police. This may appear to be a sensible consideration, but could, for instance, systematically bias the system to identify those who have lived in areas with high police presence. If these individuals then face further extrinsic circumstances which make reoffending more likely, a cycle could arise wherein the system never learns to correct this bias. Such unintended and unexpected scenarios risk entrenching forms of bias and discrimination.

Importantly, it is not always possible to avoid problems of discrimination by simply excluding sensitive data from entering a machine learning system. Often one or more other pieces of data can either reveal or correlate to these sensitive details. Details about a person having attended a particular school, for example, may indirectly disclose the person's gender if that school is gender-selective. In some places, a person's address may make it highly probably that the person is of a particular race or ethnic background or socio-economic status. Details about a person attending a particular event or a person's membership of a club or group may expose any number of characteristics about that person. The more data points about individuals that are fed into a system, the more likely it is that some of them will serve proxies for sensitive personal details.

3

“Discrimination, even if inadvertent, can cause substantial reputational damage to an organisation”

Consider the example of an AI system deployed to determine car insurance premium based on a wealth of digital data concerning the policy holder. Even if characteristics such as that individual’s gender, age and race are excluded from the input data, any number of data points about the individual may still correlate with gender, age or race in a way that a human cannot foresee. The AI system may therefore produce results which illegally discriminate against individuals on the grounds of gender, age or race, notwithstanding well-intended attempts to avoid this outcome.

Discrimination, even if inadvertent, can cause substantial reputational damage to an organisation. Recent press has featured a number of stories about AI systems that have been seen to be “racist”, “sexist” or otherwise discriminatory. Moreover, discrimination based on certain characteristics including gender, age, disability, pregnancy, race/ethnicity, religion or sexual orientation can be unlawful in many contexts.

Vulnerability to misuse

Malicious use is a noteworthy vulnerability of AI systems because, specifically in the case of dynamic machine learning systems, inputs fed into the system through ordinary use have the ability to change the system’s model. Malicious data feeds can modify an AI system’s model so that the system produces ’bad’ results. This type of misuse of an AI system can impair the functionality of the system’s model and can, in a similar way to issues of discrimination, be reputationally damaging for an organisation. A recent, high-profile example was the Microsoft Twitter bot, Tay. Tay employed a machine learning algorithm to learn from tweets it received on Twitter to inform tweets that it would post. Not long after Tay was made open to the public, it began posting racist and misogynistic tweets, mimicking tweets that had been maliciously sent to it by certain Twitter users.

A potentially more significant risk is conceivable in connection with face recognition systems being used in vital processes such as border control. Very small changes to what such a system sees can completely change how the recognition system labels it. With some knowledge of the system, deliberate changes can make an image indistinguishable to the human eye but totally mislead image classifiers.

It may also be possible that the repeated input of data into an AI system could function to discover the underlying model that the system us using. In many cases, a significant amount of the commercial value in the AI system will be in the model. This category of risk could give rise, then, to an unexpected window on the valuable inner workings of an organisation.

4

Malicious re-purposing

As highlighted earlier, one of the great features of some AI systems is their adaptability. Adaptability in terms of an AI system being able to adjust its reasoning and responses based on exposure to new data, but also the relative ease with which generic AI algorithms can be re-purposed and applied in different contexts. While this paper focusses primarily on systems designed to achieve benign or beneficial purposes, it is worth bearing in mind the risk that a system could be repurposed in the wrong hands from a benign to a malignant use. For example, recently a facial recognition system used to identify biomarkers of disease was controversially repurposed for identifying biomarkers of criminality. One can equally see that systems for analysing financial market activity might be capable of repurposing to disrupt those same financial markets.

> "a system could be repurposed in the wrong hands from a benign to a malignant use"

5

Privacy

Wherever data is being processed, compliance with privacy/data protection laws should be front of mind. Often the larger the data set, the more likely it is that the data set will contain personal data (information or an opinion about an identifiable individual). Many applications of AI involve working with supersize data sets. As we explored earlier, AI systems are particularly useful in big data analytics, due to their ability to synthesise and extract value from enormous amounts of complex data. We also discussed the benefits of using large training data sets to improve the quality of a machine learning system's model and therefore the accuracy of its results. The likelihood that personal data is being processed also depends on the nature of the data. For example, a machine learning system designed to achieve accurate facial recognition would usually require training with many photographs of people's faces, which are personal data. Conversely, an AI system used to detect patterns in securities trading, may not involve any use of personal data.

We have identified three aspects to the privacy risk that are particularly salient when working with AI systems that process personal data. The first two are common to most forms of big data analytics.

First, proper consent must be obtained from each individual (or there must be some other basis under applicable privacy/data protection law) for all processing that takes place. Auditing consent can be challenging where the data set is vast, particularly where if the data set has been compiled from multiple sources. The EU General Data Protection Regulation coming into effect in early 2018 will make this even more critical, as it: imposes obligations directly on data processors, includes more onerous requirements around obtaining consent and significantly increases the penalties for non-compliance (up to the greater of €20 million or 4% of global annual turnover). However, one of the interesting challenges of AI that the Information Commissioner in the UK has already noted in its guidance is that AI may well undermine the traditional binary, yes/no approach to consent: “This is seen as incompatible with big data analytics due to its experimental nature and its propensity to find new uses for data, and also because it may not fit contexts where data is observed rather than directly provided by data subjects. However, there are new approaches to consent that go beyond the simple binary model. It may be possible to have a process of graduated consent, in which people can give consent or not to different uses of their data throughout their relationship with a service provider, rather than having a simple binary choice at the start.”[1]

Secondly, for voluminous data sets, it can be trickier to ensure compliance with data protection requirements around storing personal data securely, keeping personal data up to date, permitting data subjects to access their personal data, complying with requests from data subjects for their personal data to be deleted and actively deleting personal data once it is no longer required for the purpose for which it was collected.

Thirdly, and more peculiar to machine learning, is the risk that some AI systems can be used in a way that permits personal data to be reverse-engineered from the system’s model. In some applications, personal data may be observable by analysing the outputs generated from certain inputs. Some systems are even capable of being operated in reverse, so a user could give the system something that would ordinarily be one of its outputs and ask it to generate an input that would produce that output.

6

Social Disruption

AI systems performing safely and securely may still have dramatic effects. Much has already been written on the scope for AI and automation to displace the human workforce and this form of efficiency-driven disruption is already well-recognised in certain sectors. Examples include AI models which can underwrite life insurance applications, provide robo-advice or conduct due diligence on countless commercial and legal contracts. This can manifest itself as a risk for a business in two particular ways. Firstly, businesses which are set to replace manual processes with artificially intelligent machine processes will of course need to consider the impact on workforce morale and on labour relations; automation is of course not always bad news for employees in view of its ability to liberate humans from processes. Secondly, at a more strategic level, businesses will need to consider how they begin to reallocate and re-focus their resources as their operational processes change.

Alternatively, from a more radical perspective, we may need a complete fiscal rethink if ideas such as a robot tax gain more momentum in the near future. Figures as varied and influential as Bill Gates, Benoit Hamon (the socialist candidate in the French presidential election) and Elon Musk have argued that the development of a technology like AI, if it is to replace human workforces, necessitates a paradigm shift in our economic attitude.

> “AI may necessitate a paradigm shift in our economic attitude.”

[1] ICO (2017) Big data, artificial intelligence, machine learning and data protection

Regulation as society's approach to managing risk

A response but not a solution

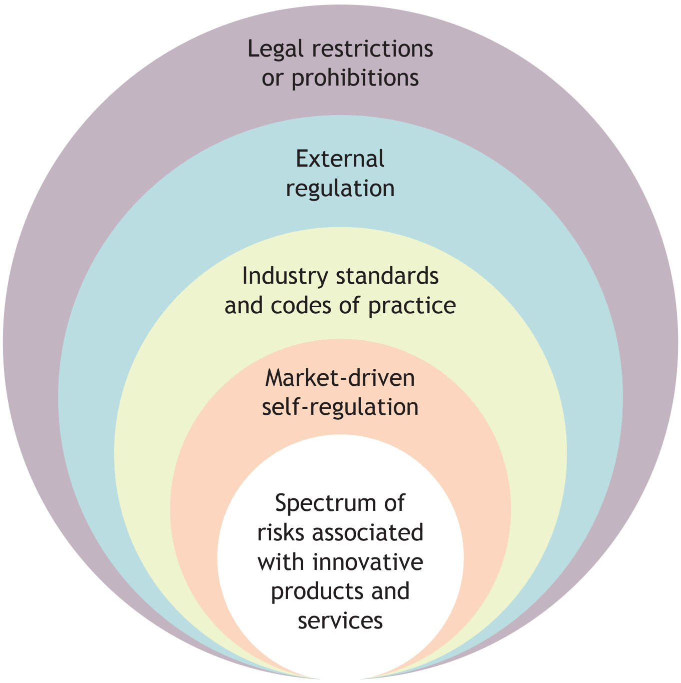
New technologies commonly raise questions of regulation. The ethics and laws surrounding new fields of innovation have often developed alongside the potential of the new field itself – biotech and genomics is a good example where ethical concerns have needed to take precedence over technical progress. Artificial Intelligence is no different in raising questions about governance but, because it relates to thought and the agency of human judgement, it presents a unique challenge.

There is one aspect of machine learning that stands out as a distinct challenge for policy makers and corporate governance. Statistical models that accurately predict an outcome or classify an object have traditionally been transparent in their reasoning. We have been able to see their workings and interrogate their internal logic. For some AI algorithms, this is simply not possible[2]. An algorithm might be able to prevent more car crashes than any human driver, but it won't be able to explain why a crash happened. It might be able to diagnose cancer more reliably than any human doctor, but with humans neither understanding why it is more reliable, nor why it might sometimes make a mistake.

Automated processes can still carry 'an aura of objectivity and infallibility'[3]. The norms and assumptions relating to their operation over the last few decades have generally been valid when dealing with static algorithms or rules because we are used to computers following instructions quickly and perfectly. When things have gone wrong, it is usually because the rules have not been set right: it has been people at fault, not the machine.

States and societies are already equipped to navigate human error and typically have a range of escalating options for responding to the types of risks which could be associated with new products or services that are not responsibly deployed.

At one end of the spectrum, where a new technology or innovation is used by businesses in a relatively benign area, **market-driven self-regulation** may be sufficient to manage any associated risks – if a product simply does not work, consumers are likely to regulate with their feet – by using a different provider of the same product, or by not buying into the product at all. For example, a dating app based upon a faulty algorithm is not likely to survive for long if users are paired with unsuitable matches. Alternatively, some algorithms simply do not have the capacity to be seriously harmful. For example, Dijkstra's algorithm to route traffic across the internet, or the MP3 sound compression algorithm, or Auto Tune to make off-key singing more bearable.


Legal restrictions or prohibitions
External regulation
Industry standards and codes of practice
Market-driven self-regulation
Spectrum of risks associated with innovative products and services

[2] Dave Weinberger (2017) Alien Knowledge

[3] Ian Bogost (2015) The Cathedral of Computation

Where there is an increased level of risk, commercial actors may self-regulate, for example by agreeing **industry standards and codes of practice**. This may include having certain entry requirements to membership of an industry body, which would then allow members to use a "kitemark" on their products. Alternatively, there could be a kitemark applied on a product-by-product basis following assessment against relevant industry body standards, following trends in respect of other safety-oriented products such as crash helmets, windows and smoke alarms.

In some AI fields, it is implausible that purely internal regulation will be sufficient to provide a consistent and acceptable level of public control for learning algorithms: where human life and liberty is involved, in medical, transportation, or military applications for example. Therefore, at the next stage along the spectrum, governments may empower independent **regulatory bodies**, which are external to a market, to influence and oversee market conduct. The role of regulators here is typically to protect consumers as end-users within a given market, maintain stability and integrity in the market and promote healthy competition between the relevant market actors. The presence of an external regulator often engenders a greater level of trust and confidence in the relevant market and its products; this can be as simple as an easily accessible ombudsman and a binding industry code (for example, regulatory bodies in the legal or advertising spaces).

At the top end of the risk-management spectrum, governments can apply **legal restrictions or prohibitions**. At the most draconian end would be a total ban on certain types of activity where a very high risk of harm exists, backed up by sanctions. Before an outright prohibition, however, may come laws and regulations that limit or control particular activities and establish frameworks within which close monitoring and supervision takes place. A good example of this would be the aviation industry, where the barriers for entry are high and regulatory requirements stringent, with in-depth investigations into any failure – this is not surprising, given that any small failure could affect very many lives.

AI is at its heart simply a tool which can be used in existing public and private sector functions – the level of risk associated with any particular AI system should be a function of the level of harm that would be caused by its failure, and so the responses of legislators and policy-makers to use-cases for AI systems will inevitably fall across a spectrum, depending on the context of the uses.

Legislation and regulation tend to be blunter tools than industry-driven regulation and codes of practice, and will typically have a more restricting effect on innovation, creativity and productivity. It is certainly not the case that AI is free from legal and regulatory oversight today and so businesses seeking to deploy AI systems already must have regard to the impact of those existing legal boundaries.

The key point we seek to make in this paper, however, is that businesses across all sectors have an opportunity to shape and influence the future development of legal and regulatory frameworks as they begin to adapt to address machine-based processes, products and services.

By being forward-thinking and, perhaps above all else, **responsible in the design and deployment of AI** businesses, sectors and whole industries can mitigate the restricting effect that external regulation of AI may have in the longer term. Getting out ahead of the policy-makers by designing AI systems that can take full account of the risks to which they may expose individuals, communities or society, and as far as possible mitigating those risks must be the best means to avoid the need for legislators and policy-makers to feel they must take restrictive or prohibitive action.

With that in mind, in the final section of this paper we have identified three key themes for businesses to focus on when contemplating the design and deployment of AI systems, and a series of practical reflective questions that are intended to enable those businesses to identify and minimise the risks of harm associate with AI system deployment and thereby maximise the ability to exploit the gains that AI promises to deliver.

Key principles for responsible deployment of AI

Good Governance

The ability to understand and challenge decisions and processes within your organisation is crucial to discharging your duties of good governance. As with other aspects of technology in business, such as cyber risk, the assessment and responsible deployment of AI requires both good governance and good technical execution.

1. **Has the business accepted AI as part of its risk register?**
   AI should be a core component of an organisation's risk register. Assuming your model could fail, it is vital to assess the likelihood and the impact of such a failure. What could a failure look like? In the event of a failure, what contingency plans exist if the operation of the algorithm needs to be suspended? Have these been tested with a fire-drill in a live environment?

2. **Does the business have real-time monitoring and alerting for security and performance?**
   Firms commonly require IT functions to monitor the state of the network, but it is currently less common to monitor the health of the algorithms, and the underlying data, that are driving the decision making across that network. This is a capability that would provide firms with significantly greater ability to respond quickly to problems if they arise.

3. **Is the causal process auditable?**
   The lack of transparency often associated with AI systems is not insurmountable – human beings are often opaque as well. If the full context of the algorithm can be logged, its decisions could even be more reproducible than a human example. Do you have in place the strategy and capability to reconstruct decisions by these algorithms? Your approach must be sufficient in the context of your particular market environment.

4. **Is there an accountability framework for the performance and security of the algorithm?**
   Every significant algorithm the business relies upon should have an internal owner, even where the algorithm is supplied by a third party as a 'black box'. A central register or record of such algorithms and their owners can be a significant asset to a business trying to navigate a complex internal data landscape. In addition to a named owner, ideally the record should contain basic information relating to the purpose and structure of the algorithm, and the services that are dependent on it. There should be a clear reporting line from these key individuals to senior executives, with careful avoidance of closed feedback loops to help mitigate any lack of clarity around responsibility for algorithms.

Quality Assurance

AI systems now work sufficiently well to deliver real business value, but they are not yet a mature technology well understood throughout all levels of business. The value of AI will accrue to businesses who develop confidence in their technology and succeed in deploying AI in practice. To be assured of performance from a system that is complex and often difficult to interpret will require a broad picture of what's going into the system and how it has performed under a range of scenarios.

5. **Are the algorithmic predictions sufficiently accurate in practice?**
   The performance of your algorithm is a critical and often dynamic metric that can change over time as the model learns from new data. The levels of risk, and type of errors, that are acceptable will depend on the context and what currently counts as "good enough". The best metric to assess the performance of the model will vary depending on the application (e.g. classifying events into groups, detecting anomalies, predicting demand). Do you understand your risk tolerance in the deployment environment? Is your organisation clear on the benchmark level of error for your particular application that would constitute unacceptable accuracy and whether the model operates above or below that level, or is there a need to look to external professional support?

6. **Has the algorithm been fed a healthy data diet?**
   Garbage in, garbage out always applies to algorithms. How robust is the model to junk, partial, or poisoned data? Data is often overlooked for potential biases, especially if they are the result of pre-existing biases within human systems. A series of straightforward checks would help satisfy this concern: was the training sample of sufficient scale, and checked for bias? Has it been tested against adversarial data? Is data used in developing the model equivalent to the data it will see once deployed? Are statistics on the data being monitored, as well as the overall model performance?

7. **Is the algorithm's objective well-specified and robust to attack or distortion?**
   It is important that the objective the AI system tries to achieve is well specified to avoid accidents, unintended misuse or malicious repurposing. Have you been sufficiently specific in defining your objective, to capture what you want the system to do? Is there some undesirable behaviour that could accidentally fulfil the objective? Have you gone beyond best-case scenarios in testing the algorithm?

Legal

In light of the variety and complexity of AI and its deployment across businesses, there may be a limit to the extent to which existing regulatory systems are able adequately to supervise and control the risks involved. Responses to this new technology may be unexpected and tricky to navigate. In many verticals, from financial services to civil aviation, medicine to consumer credit, sectoral regulatory regimes can already present complexities that will only further complicate the legal challenges for businesses looking to deploy AI systems. Similarly, the application of AI tools in cross-cutting areas, such as pricing, may well raise issues under more generally applicable regulatory regimes such as competition law. Responsible deployment of AI requires careful and specialist legal engagement. Under this heading, we have identified three legal focal points that will be critical to the responsible deployment of most AI systems.

8. **Is the algorithm's use of personal data compliant with privacy and data processing legislation?**
All learning algorithms depend on data, and typically substantial quantities of data, and often this data includes personal data relating to individuals. There are well established laws governing the use of personal data, but regulation in this area is generally becoming stricter and the penalties for breach are becoming more severe (the EU General Data Protection Regulation is already a new benchmark in this field). Whenever you are looking to deploy an AI system, it is essential to undertake a proper data protection assessment to ensure that any personal data processed in the system is dealt with lawfully. In particular, there must be a lawful basis for all processing of personal data, and the operation of the AI system may necessitate changes to that basis over time. Your organisation's data protection officer should therefore be heavily involved in the deployment and specialist legal advice may be required to provide assurance that the business can remain compliant with its legal obligations.

9. **Does the algorithm produce discriminatory outputs?**
All responsible businesses will avoid intentional bias. As we noted in the previous section of this paper, a particular concern with AI systems is that they can develop unintentional discriminatory biases quite easily. While you may reduce the risk of discriminatory biases by carefully selecting the data used to train a particular algorithm, it is often difficult, if not impossible, to identify all the data that may cause a discriminatory bias in a system. This is a legal point that requires technical execution, lawyers who understand AI and the potential for discriminatory effects will need to work closely with data scientists who can offer effective solutions. To protect against the problems of discriminatory biases in your use of AI, you should:

    (a) fully understand the potential legal as well as reputational risks associated with discrimination in the area where your AI system is being deployed;

    (b) seek validation that your algorithm has been subjected to comprehensive quality-testing, either by an appropriately independent internal team or if necessary by an experienced external resource, to ensure that it does not engage in illegal or otherwise damaging discrimination; and

    (c) if an undesirable discriminatory bias is detected, seek help to correct for this; this may entail feeding the AI system tailored input data to negate the discriminatory correlation or alternatively building into the system an active correction filter for the discriminatory bias. Such fixes will become increasingly straightforward to achieve in data science terms, but the fix may only be possible if the if the issue is addressed up front.

10. **Have you considered what is a fair allocation of risk with your counterparties?**
For now one of the more knotty problems to solve in the field of AI is the allocation of liability. It is increasingly common for AI systems to be deployed to undertake tasks that would formerly have been performed by a person. Whereas previously a person could be held responsible for any failures in the performance of those tasks, it may not always be clear who is responsible when an AI system fails in the performance of its tasks. It may be difficult to discern whether the problem originated with the design of the algorithm, the coding of the software, some other element of the coding of the surrounding software, the initial training data, or how the algorithm has been deployed. The critical questions from a governance and risk management perspective, therefore, are: who is responsible for failure of the AI system to perform? In what circumstances will they be responsible? And what are they responsible for? As a business seeking to deploy an AI solution, the answers to these questions should direct key aspects of contracts with investors, developers, suppliers, implementation partners, external consultants and customers. It is essential that the legal and commercial teams involved in preparing and negotiating these contracts understand the unusual complexities around liability and risk for AI systems.

The reflective questions which appear in this section are intended to be used as a practical touchstone for General Counsel, Chief Data Officers and other senior executives as part of a process of assurance for AI design and deployment. We have grouped these questions into three broad categories, albeit there will inevitably be overlap between them.

"Artificial Intelligence could be the most transformative technology of the 21st Century"

Conclusion

Handle with care

We believe that Artificial Intelligence could be the most transformative technology of the 21st Century. These practical opportunities are not the stuff of the future – they are here today. AI is already driving down the cost of doing business, improving the quality of decision making, and increasing the personalisation and responsiveness of services. This is no longer just an opportunity for the R&D team or the Innovation lab, but for mainstream business leaders in every industry and sector.

However, while learning algorithms have enormous positive potential, they also carry significant legal, security, and performance risks that, if not managed well, could jeopardise reputations, or worse. Just as AI may transform the success of businesses that use it well, it may also be at the root of future corporate failures and social harms. If knowledge is power, then AI has the potential to give superpowers to our human resources; and those superhuman resources will have to be used with care.

Technology issues have historically been the preserve of the back office, not the board. In the last decade, with the rise in significance of technology and of related cyber risks, the seniority of technology leaders has increased, with Chief Digital Officers and Chief Data Officers commonly now being appointed to the main board.

But AI isn't just another technology. Because of its ability to alter and in some cases replace human processes, the way we have traditionally supervised and assured the judgement of our human resources will have to be reinterpreted if it is to be applied effectively to machines.

The scalability of digital services means that a single algorithm could soon (or may already) affect the lives of millions of customers, suppliers and counterparties, and be responsible for decisions worth billions. Just as with cyber security, the reputation of entire organisations may hang in the balance. AI holds enormous promise, but it requires responsible deployment. If business does not take on that responsibility, then other bodies may feel obliged to step in.

In this report, we have set out a series of practical questions for boards to use to assure themselves that they are deploying AI effectively and responsibly. Some of these questions may require specialist legal or technological knowledge and support to answer, but we hope that our report has at least clarified the right questions to ask.

And if our report has provoked further thoughts or questions for you, we would be delighted to hear from you.

The authors

Slaughter and May

Slaughter and May is a leading, full service, international law firm headquartered in London. Our fintech team, led by Ben Kingsley and Rob Sumroy, supports clients across the full spectrum, ranging from established international financial institutions and global technology, media and telecoms providers, to investors and high growth start-ups with the potential to become market disrupters and leaders. We advise on the legal implications of developments in the fields of technology, intellectual property, data use, and financial regulation, but our interest spans any and all legal implications for innovation and growth in the fintech sector.

Ben Kingsley
ben.kingsley@slaughterandmay.com

Rob Sumroy
rob.sumroy@slaughterandmay.com

Ian Ranson
ian.ranson@slaughterandmay.com

Matthew Harman
matthew.harman@slaughterandmay.com

Harry Vanner
harry.vanner@slaughterandmay.com

The authors

ASI Data Science

ASI Data Science is a high-growth British technology firm working to build the capability of other organisations to use Artificial Intelligence effectively. We coach and support firms of all sizes and sectors to become AI ready by providing first class data science software, training, project and advisory services.

Richard Sargeant
rcs@asidatascience.com

Marc Warner
marc@asidatascience.com

Nick Robinson
nick@asidatascience.com

This material is for general information only and is not intended to provide legal advice.
For further information, please speak to your usual Slaughter and May contact.

June 2017