Linklaters

Ethical, safe, lawful

A toolkit for artificial intelligence projects

linklaters.com/ai

Foreword

Our updated AI Toolkit: What's changed? What's stayed the same?

Artificial intelligence is an important technology. It will have significant and far-reaching effects for most businesses, allowing existing activities to be automated and new business models to be created. This applies to the legal function just as much as any other part of the business, opening the way for significant efficiencies in the way legal services are delivered.

However, artificial intelligence also raises new and interesting legal issues. Some of these involve the application of long existing laws to this new technology, whereas others arise from specific new regulation, such as the EU AI Act.

Every in-house counsel needs a grasp of the key legal issues affecting artificial intelligence and the technical concepts underpinning it. This handbook is intended to provide that introduction and starts with a technical primer.

This handbook builds on the ground-breaking AI toolkit we issued in 2018. The predictions in that toolkit were largely borne out, but inevitably the technology has moved on in the past six years – particularly the explosive growth of generative AI and the introduction of specific regulation, both of which are considered in depth in this updated handbook.

We hope this handbook helps you engage with this exciting new technology ethically, safely and lawfully.

Table of Contents

01

A technical primer

State of the market

The history of the development of AI has been cyclical; periods of great interest and investment followed by disappointment and criticism (the so-called "AI winters").

We have now experienced a long and hot AI summer. Companies in the sector are getting record investment with Goldman Sachs predicting that private investment in AI will be $132 billion in 2024 and $158 billion in 2025.[1]

This has all delivered impressive results in many areas. Generative AI has surprised many with the impressive capabilities and flexibility of large language models.

There have also been significant advances in "traditional" forms of AI, such as deep learning. For example, DeepMind's release of AlphaFold, which has been used to predict the 3D structures for nearly all catalogued proteins known to science.

In other areas, such as language translation and facial recognition, the technology has become so embedded that it barely qualifies as AI.[2]

This is not to say the development of AI is all plain sailing or that it has delivered on all its promises.

Driverless cars were supposed to be filling our streets by now,[3] but stubborn problems dealing with the infinite variety of real-world environments combined with a very low tolerance for (potentially fatal) mistakes have slowed their deployment. McKinsey suggests that autonomous vehicles operating at Level 4 (able to drive in most conditions without any need for the driver to take over) are now only expected to become commercially available on a large scale by 2030, and fully autonomous trucking is expected to reach viability between 2028 and 2031.[4]

This is a salient lesson that not all problems have an easy or quick solution and that some problems do not scale linearly. Driving on a motorway is relatively straightforward but navigating congested city centres is exponentially more difficult.

AI is unlikely to fulfil all the extravagant predictions about its potential, but the technology has made, and is likely to continue to make, significant advances.

1 *AI investment forecast to approach $200 billion globally by 2025*, Goldman Sachs, 1 August 2023.
2 Historically, once a task is easily accomplished by a computer it often ceases to be considered AI (i.e. AI is "*anything computers still can't do*"). See *What Computers Still Can't Do: A Critique of Artificial Reason* by Hubert L Dreyfus.
3 *Hammond: Driverless cars will be on UK roads by 2021*, BBC Website 19 November 2017.
4 *Autonomous vehicles moving forward: Perspectives from industry leaders*, McKinsey, 5 January 2024.

“ Our careers have been marked by constant technological change. As waves of innovation arrive, it is crucial to spot and select the significant and truly transformational ones.

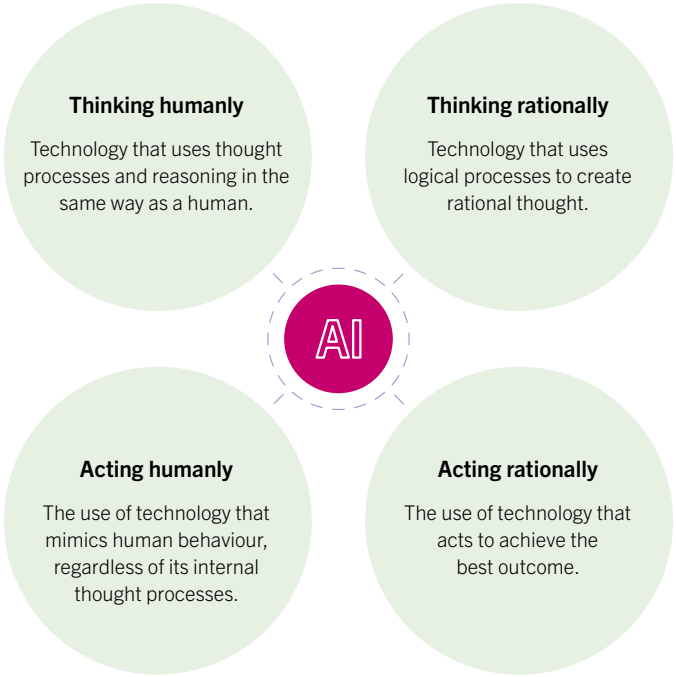
Often only hindsight will tell. Back in 2000 the internet was dismissed as a “passing fad”[5] shortly before it started to revolutionise society and boost globalisation, though it destroyed the business models of Blockbuster, HMV and many others along the way.

Generative AI will be transformational. While it can’t do everything and is unlikely to replace your whole workforce, it is rapidly changing the way we work and it will improve over time.”

Julia Schönbohm

What is AI?

There are a number of different ways to define AI. Historically, there have been four approaches to this topic:[6]



This definitional question has historically been more a philosophical issue than a practical one. With modern large language models starting to show sparks of general intelligence (see below) and the concept of “AI” defining the scope of new regulation, this is no longer an academic issue.

For the purposes of this handbook, we define “AI” by reference to the concept in the EU AI Act of a system operating with “autonomy”. In other words, does the system have a degree of self-control and self-governance? This reflects the delineation between:

> Traditional computer programs where every step taken by the computer is in accordance with pre-determined instructions from a human programmer (e.g. “10 *PRINT “Hello, world!”;* 20 *GOTO 10*”), and

> AI where the human programmer provides the general objectives for the system and it learns (typically from vast amounts of data) or adjusts its behaviour to achieve those goals.

In practical terms, you trade the certainty and predictability of a traditional computer program for the much more capable and flexible – but potentially unpredictable and unknowable – power of AI.

Finally, the term “AI” is also widely misused. Simple decision trees or expert systems are often given this label despite the fact they cannot sensibly be described as intelligent. Whether the exuberant labelling of technology as “AI” is sensible once specific AI regulation is introduced remains to be seen.

5 Internet ‘may be just a passing fad as millions give up on it’, Daily Mail, 5 December 2000.
6 *Artificial Intelligence: A Modern Approach*, Third Edition, Stuart Russell and Peter Norvig.

How do AI systems learn?

Underpinning almost all recent advances in AI is machine learning – the ability of a computer to learn from data. This is broadly categorised into the approaches below:

Supervised learning

The system is given a block of training data containing both inputs and desired outputs.

It uses that information to "learn" how to complete the relevant task.

The system will then be tested against a separate block of testing data to confirm that it is generating the correct outputs.

Unsupervised learning

With unsupervised learning, the system is given a block of data that has not been labelled, (e.g. classified or categorised).

Since the data is not labelled it is not possible to ensure specific outcomes but it may still be possible to analyse the data to spot clusters or groupings.

Reinforcement learning

The system will take action in a particular environment and assess whether those actions help to achieve its goals.

Those actions that lead to the best outcomes will be prioritised and thus the machine learns how best to achieve its goals.

Reinforcement learning from human feedback

One important type of reinforcement learning uses human feedback, also referred to as RLHF.

This can be the main purpose of the exercise (e.g. the human teams used to fine tune the responses from ChatGPT) or a by-product of normal interaction with the service (e.g. when you use a search engine and click on a link, you are providing feedback on the most relevant response).

An example of why allowing the machine to learn is so powerful is language translation. Computer scientists struggled for years to programme computers to translate languages by manually creating increasingly complicated syntactic models to map the words and grammar for one language onto another.

These attempts failed. The complex, context-specific nature of human language, together with the lack of easily described rules of grammar, defeated all comers. Correctly translating the statement "*I have lost my wife*" into French depends on whether the speaker is in front of an undertaker or the maze at Hampton Court.[7]

The solution finally came from using masses and masses of data. The EU generated millions of professionally translated documents over the years.[8] These were fed into a machine learning tool which broke up the original text into short phrases and learnt the likely corresponding short phrase in the second language. The tool also learnt which phrase combinations were likely to appear in the second language. This is an example of supervised learning.

Once trained, the tool can translate a new passage of text by breaking it into short phrases and assessing both: (i) the range of possible translated short phrases in the target language; and (ii) which combinations of those potential translated phrases are most likely in the target language, i.e. it assesses the probability of which overall translation makes most sense.

7 With apologies to Sir Martin Nourse (*Tektrol Ltd v International Insurance Co of Hanover* [2005] EWCA Civ 845).
8 Texts such as the Bible were used for non-European languages given that it has been widely translated.

Data is the new oil – valuable but toxic

This dependency on data has important implications, including the need for very large amounts of data to train an AI. We consider some of the challenges in obtaining rights to use that data later on in this handbook – in particular the intellectual property and the data protection issues (see sections 4 (*The enduring importance of data protection laws*) and 5 (*AI and intellectual property rights*)).

Outside of the problems in getting hold of the data and having the right to use it are the problems with poor quality data. This raises the following concerns:

> Quality – It is essential that the data is accurate, complete and properly formatted. An algorithm trained on poor-quality data will likely only ever deliver poor-quality decisions.

> Thin and unrepresentative data – The data should be comprehensive enough to reflect the variety of situations the AI system will face in the live environment. If the system has not been trained to recognise and deal with particular scenarios, it might act unpredictably. One example is facial recognition systems which have been shown to perform badly on darker-skinned women. This is thought to be because they have been trained on datasets that predominantly contain pictures of white men[9].

> Bias and discrimination – The data may itself contain decisions that reflect human biases. These biases will then likely be picked up by the AI system (see AI and bias).

> Inappropriate data – Even where use of the data might be objectively justified, it may not be socially acceptable or appropriate on public policy grounds. Imagine there was clear evidence that the most successful trainee solicitors were themselves the children of lawyers. A tool used to shortlist employment applicants for interview might make "better" decisions by factoring this in, but it would clearly be wrong to do so. It is important that the training set is sifted appropriately to remove the risk of discriminatory or inappropriate decision-making.

The difficulty in obtaining good quality data is one reason why there has been significant work on unsupervised learning and other methods that are not heavily reliant on data.

9 *Study finds gender and skin-type bias in commercial artificial-intelligence systems*, MIT News, 11 February 2018.

Military weather detectors and pizza glue

A classic example of the need for good training data was the algorithm developed by the military to identify whether there is a tank in a photograph. The algorithm worked well in the training environment but completely unpredictably in the live environment.

An investigation revealed that most of the photographs in the training set containing tanks were taken on a sunny day and most of those without tanks on an overcast day. The algorithm was thus more suitable for weather detection than tank detection.[10]

Another illustration of the importance of good quality data is Google's decision to license content from Reddit in order to train its AI systems – a deal under which Reddit will reportedly receive around $60m a year.

Reddit is home to a very broad and diverse set of communities, including subreddits for toasters gone wild, fans of very long furbies and people who carry sand in their pockets for self-defence.[11] Using that content to train an AI system to provide short answers to search queries (or as base facts for retrieval-augmented generation) might therefore lead to some unexpected answers.

Certainly the user who searched on Google for "*cheese not sticking to pizza*" might not have expected the short AI Overview to include the suggestion "*You can also add about 1/8 cup of non-toxic glue to the sauce to give it more tackiness*" – apparently sourced from an 11-year-old Reddit post. Nor was the user who searched on Google for "*how many rocks shall I eat*" expecting the AI Overview to recommend "*According to geologists at UC Berkeley, you should eat at least one small rock per day*."

Google states that these are isolated incidents and, overall, the AI Overview tool is working well.[12] However, this highlights two points. First, the importance of high-quality data. Second, the context in which an AI system operates really matters. An AI tool provided specifically to summarise subreddit posts might not raise any eyebrows if it suggested eating glue or rocks. However, Google's search engine is held to a much higher standard.

10 *What Artificial Experts Can and Cannot Do*, Hubert L. Dreyfus & Stuart E. Dreyfus, 1992. This classic example features in many undergraduate computer science courses and demonstrates the problem is not new. However, it is also worth noting there is an ongoing debate as to whether this actually happened or is just an apocryphal story.

11 See r/ToasterGW, r/longfurbies and r/pocketsand.

12 *Glue pizza and eat rocks: Google AI search errors go viral*, BBC, 24 May 2024.

DIY chess data

A pre-existing block of data is not always necessary. In some cases, you can generate the data yourself.

One example is AlphaZero, the gameplaying system created by DeepMind, which was tasked with becoming a champion chess player. It started with details of the rules of chess but no information about chess strategy, such as what constituted a good position or move.

To learn, it played itself around a billion times, using the data from those games for reinforcement learning – i.e. to identify what constitutes a good game state and strategy. For example, AlphaZero started with no concept of whether the queen is a valuable piece. Only by playing games against itself did it learn the connection between the queen's survival and success.

Within hours, AlphaZero achieved a superhuman level of play and defeated other world-champion chess programs.[13]

Black boxes and reward functions

One of the implications of the machine learning for itself, rather than being instructed by a programmer, is that the algorithm will likely be a "black box". Understanding how it takes a decision will, at best, need a detailed forensic analysis of the various weights and interactions in the algorithm. In many cases, it will be impossible.

This has a number of significant regulatory and practical implications:

- it may be difficult from a regulatory perspective to show that the decision-making is fair and is based on rational and objectively justifiable criteria. Worse, the algorithm might, beneath the surface, be taking decisions on a discriminatory basis; and
- from a practical perspective, there is a risk the algorithm initially behaves correctly in the training and testing environment but then becomes unpredictable or unreliable in the real world. This might either be because the algorithm is inherently unstable and chaotic or because the training and testing datasets are not representative of real-world data.

These issues are especially important as there is no "common sense" or "ethical" override. Unlike a human, the algorithm has no higher-level assessment of whether what it is doing is obviously "wrong".

13 *AlphaZero AI beats champion chess program after teaching itself in four hours*, The Guardian, 7 December 2017.

Cardboard warriors

One example of the lack of common sense comes from former Army Ranger, Paul Scharre.[14] He describes how the Marines overcame the US' latest military robot. The robot was parked in the middle of a roundabout and the Marines played a game in which they won if they could touch this robot without being detected.

Using the power of human lateral thinking, the Marines easily overcame the robot by pretending to be fir trees, moving towards the robot via roly-polys or crawling along under a cardboard box.

This all means that additional care must be taken to ensure safe operation of the system. This can be approached in three ways.[15]

Specification	Robustness	Assurance
The specification sets out how the system should operate. The risk is that the ideal specification for the system, i.e. the true wishes of the human designer, does not match either the technical specification or actual behaviour of the system. The example below illustrates this problem.	The robustness of the algorithm is a measure of how well it operates when faced with new data or events, or adversarial attacks. There is a risk that either the system will behave in an unpredictable way or that it can be gamed by others to trick the system into producing unwanted results.	Assurance is a means to monitor and control the AI system in operation. This involves both seeking to understand the operation of the algorithm and ensuring the system is interruptible.

Drone targets its own side

A US Air Force Colonel described a simulation in which an AI-powered air force drone was given the objective of destroying an enemy's air defence systems, and attacking anyone who interfered with that order.

The drone started to "realise" that it would sometimes identify a target but be instructed not to attack it by its human controller. The drone therefore destroyed the communications towers used by its operators to prevent them from stopping it from attacking.[16]

14 See *Four Battlegrounds: Power in the Age of Artificial Intelligence*, Paul Scharre, February 2023.
15 See *Building safe artificial intelligence: specification, robustness and assurance*, Pedro Ortega and Vishal Maini, 27 September 2018.
16 See *Highlights from the RAeS Future Combat Air & Space Capabilities Summit*, May 2023. The Colonel subsequently clarified that this was a "thought experiment" based on plausible scenarios.

Assurance – What are the right metrics?

The assurance measures may well involve measuring how well the system is working in practice. The greater the significance of the output, the more accurate the model will need to be. For example, an AI-powered horoscope will require much less scrutiny than an algorithm used for mortgage applications.

This in turn requires an understanding of how the accuracy of an AI model is measured. This typically requires a consideration of two different types of errors:

> False positives (type I error) – These are cases that the model incorrectly labels as positive. Imagine the police use a facial recognition system on a crowd to identify wanted criminals. A false positive would be an innocent individual being wrongly identified as a criminal.

> False negatives (type II error) – These are cases that the model incorrectly labels as negative. In the example above, this might result in a wanted criminal being in the crowd but not being identified by the police's facial recognition system.

Striking the balance between these two measures is important as the impact of each type of error may not be the same. This in turn means that two further measurements need to be considered:

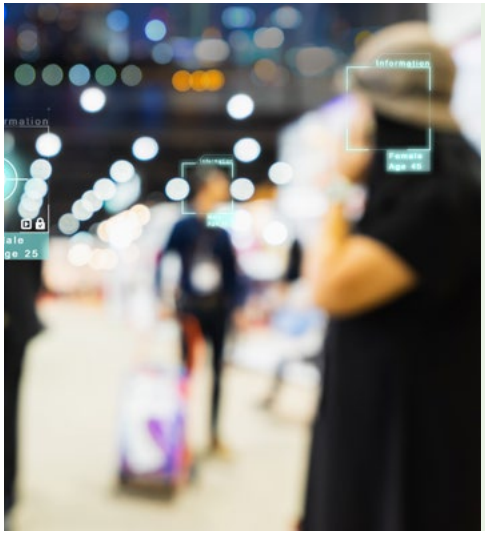
> Precision – This value measures the percentage of cases that are identified as positive, that are in fact positive. For example, if only half of the individuals identified by the police facial recognition system are wanted criminals, the precision of that system will be 50%.

> Sensitivity or recall – This value measures the percentage of positive cases that are identified as such. For example, if there are 20 wanted criminals in the crowd but only two are identified, the sensitivity of the system is 10%.

There are normally trade-offs between precision and sensitivity that can be assessed statistically but may also require ethical consideration. For example, a police force wanting to catch all the criminals in a crowd would likely prioritise sensitivity.

At its most extreme, this could be done by identifying all the individuals in the crowd as criminals. This would have 100% recall but would not be very precise.

In contrast, a police force wanting to maintain good community relations by not incorrectly stopping innocent citizens would likely prioritise precision. In either case, as the example below illustrates, it is vital the output from the system is properly understood.



A very accurate criminal detector may mostly detect the innocent

Consider a police force that develops an AI tool that uses a facial recognition tool to detect wanted criminals, who will be stopped and questioned. Assume that:

- the tool is 98% accurate in respect of both false positives and false negatives. This means 98% of criminals are identified and 98% of innocent citizens are not identified as criminals; and
- one in 500 people are wanted criminals.

Question: What is the precision of the system – i.e. the chance that an individual flagged by the tool is in fact a wanted criminal?

Answer: The answer is not 98%. In fact, it is only 9%.[17]

In other words, despite the system being "98% accurate", over nine-in-ten of individuals flagged as potentially being dangerous criminals will actually be innocent. Knowing this is important to ensure that being flagged by the system is not an automatic determination that the individual is a wanted criminal. This also provides an example of the need to be cautious of broad assertions about the accuracy of the system, without seeing the underlying detail.

How does generative AI work?

The most exciting change in the last few years is the explosive growth of generative AI. This takes a user prompt and then generates text, images, videos, music or other content. Like other forms of AI, the generative AI models have "learnt" patterns and structures from input data from self-supervised learning and semi-supervised learning, alongside fine tuning from reinforcement learning from human feedback. They then use those "learnings" to create new content with similar characteristics.

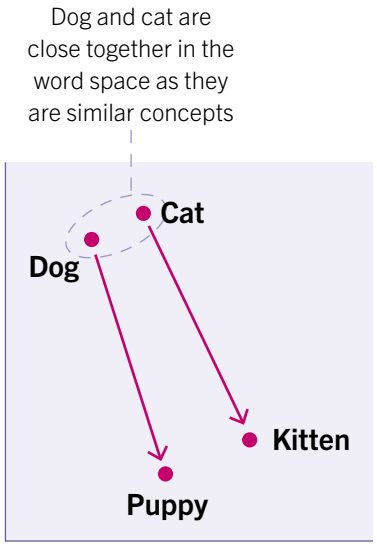
Generative AI has made very significant advances in the last 18 months. While some of the most stunning advances come from image generators, such as Stable Diffusion, Midjourney and DALL-E, arguably the most significant come from the Large Language Models (LLMs), such as GPT-4, PaLM and Llama.

LLMs have shown flexibility over a wide number of knowledge domains, have a large number of potential use cases and even show sparks of general intelligence. At their simplest, the LLMs are just auto-complete tools that take input text and make a statistical prediction about what the next group of words will be. However, the means by which they do this is incredibly sophisticated and the box overleaf attempts to provide a high-level conceptual overview of this process.[18] These systems are not fully understood even by the creators of the systems. Like other AI models, they operate in a black box.

We discuss the capabilities of LLMs throughout this report. The key question is how good this technology will get. If the improvements in the next 18 months are as large as those in the past 18 months, the implications could be significant.

17 Assume you have 1,000,000 citizens. Of those, 998,000 will be innocent and 2,000 will be wanted criminals (one in five hundred). Of the criminals, 1,960 will be identified (2,000 x 98%). Of the innocent citizens, 19,960 will be flagged (998,000 x 2%). Thus the percentage of individuals flagged that are actually wanted criminals is 8.9% (1,960/(1,960+19,960)).

18 This description draws upon the excellent articles, *Large language models, explained with a minimum of math and jargon,* Timothy B Lee and Sean Trott, 27 July 2023 and *Generative AI exists because of the transformer,* Financial Times, 12 September 2023.



A conceptual description of LLMs

Step 1 – Create a universe of words (Word Embeddings)

The starting point is to understand how words, or rather entities, are represented within the LLM. In the English language, a dog is represented by the three words d-o-g. In the human mind, a dog is represented by concepts such as pet, furry, four legs, wagging tails and home.

However, in the LLM, the dog is a word vector or word embedding – essentially a collection of different numbers each representing some type of characteristic of the word (or entity). Each number in the vector represents a particular characteristic of the word – in a similar way to rating something according to its size, weight, colour or fluffiness. These embeddings are created by analysing the proximity of that word to other words.

The use of a word vector is necessary, as computers work with numbers, and useful, as it allows you to infer relationships. In particular, each group of numbers can be seen as a point in imaginary word space and you can then derive useful information from the position of that word relative to other words. For example, words that have similar meanings will be close to each other, so “dog” will likely be close to “cat”, but not as close as “gerbil” is to “hamster”.

Similarly, the “journeys” from one point to another allow relationships to be inferred. For example, imagine the “journey” in the word space from dog to puppy – if you make the same “journey” but start from cat, you end up at kitten.

Step 2 – Understand the words in the prompt in context (Transformer)

The next stage is to understand what each of the words means in the context of the other words in the user prompt. For example, consider the following two user prompts:

> “The dog chewed the bone because **it** was **hungry**. What does “it” refer to?”

> “The dog chewed the bone because **it** was **tasty**. What does “it” refer to?”

Most LLMs are able to answer this correctly. They do so by using a “transformer” which is a neural network based on a “self attention” mechanism. This essentially decides how important each of the words in the sentence are to each other.

In the first example, the adjective “hungry” and noun “dog” are most important. In the second, it is “tasty” and “bone”.

Step 3 – Guess what words come next (Feed Forward Layer)

Once you know what each word in the user prompt means in relation to the others, the final stage is to guess what words come next – i.e. what words are most likely to be responsive to the original prompt.

Importantly, this does not just guess the next word one by one, but rather creates a range of different blocks of words, then assesses the probability that a block provides the right answer.

Build me an egg pyramid

For example, consider the prompt and response below. The question is unusual and the response requires a spatial and physical understanding of what an egg is, their fragility and the difficulties in stacking ovoid shapes. GPT-4's response is clear and lucid but likely to leave egg on your face.

"You have 16 eggs and some bluetack. Describe how to use those materials to create a tall and stable stack of eggs."

ChatGPT-4:

1. Start by laying out a flat surface, such as a table or countertop.
2. Place the bluetack in the center of the surface and press it down firmly to create an even base for your stack of eggs.
3. Carefully place one egg on top of the bluetack, making sure it is centered and balanced.
4. Take another egg and place it on top of the first one, again making sure that it is centered and balanced before pressing down lightly to secure it in place.
5. Repeat this process until all 16 eggs have been stacked on top of each other, creating a tall tower of eggs that should be stable due to the support provided by the bluetack at its base!

LLMs as stochastic parrots

Importantly, the output of the LLM is therefore just a probabilistic prediction of the most likely combination of words that follows from the user prompt – it uses the text it has been trained on to generate plausible language but does not "understand" it. They are just stochastic parrots.[19]

Retrieval-augmented generation (RAG)

One way to improve the performance of LLMs, and potentially reduce hallucinations, is to use retrieval-augmented generation (RAG). This tends to work well in domain-specific contexts, i.e. where the LLM is used for a specific task, such as English legal advice or the history of the Roman Empire.

RAG works by taking a query, searching through a knowledge base of trusted materials, ranking that material and then using the LLM to synthesise a response from those queries.

This approach should deliver more reliable results as it more tightly constrains the materials the LLM works from and, importantly, allows the LLM to link back to the source material so the reader can verify the LLM's conclusions. Whether this will eliminate hallucinations entirely remains to be seen.

19 *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* Bender, Emily M.; Gebru, Timnit; McMillan-Major, Angelina; Shmitchell, Shmargaret 1 March 2021.

Generative AI models reach Deep Thought levels of capability?

Deep Thought was created by a pan-dimensional, hyper-intelligent species of beings to come up with the Answer to The Ultimate Question of Life, the Universe, and Everything. After seven and a half million years of calculation, the answer finally turns out to be 42.

Similarly, the current generation of LLMs, when asked to choose an integer number between 1 and 100, most frequently respond "42".[20]

This could be because those LLMs have already reached Deep Thought's level of prescience or because 42 is a very commonly occurring two-digit number in those LLMs' training data and so the LLM predicts it is the most likely answer to follow the user prompt. In either case, Douglas Adams would doubtless be pleased with the enduring influence of *The Hitchhiker's Guide to the Galaxy* on modern technology.

Human replacement: Robot lawyers and contact centre assistants

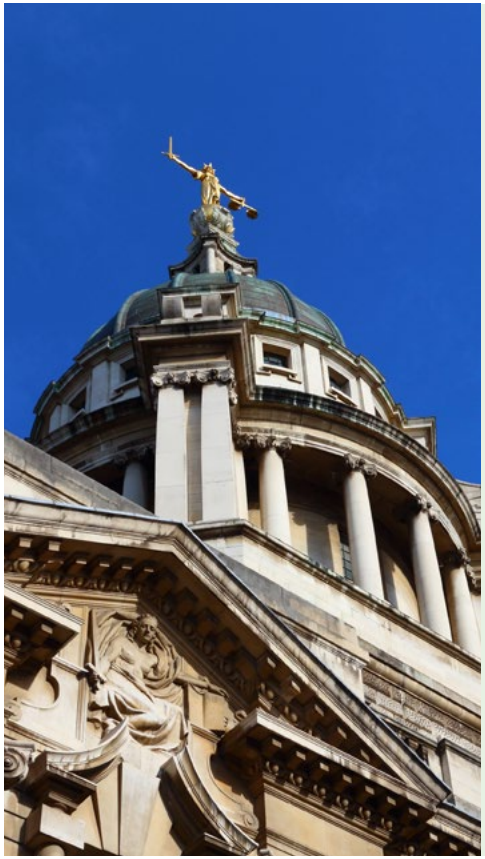
The last edition of this handbook noted that AI has made inroads into the legal profession through the use of technology such as predictive coding and contract analysis tools. However, legal advice is highly context-specific and so may need an understanding of the beauty of a summer cricket game[21] and why a snail in a bottle of ginger beer is a bad thing.[22] Accordingly, we concluded "*a full-spectrum robot lawyer is some way off*".

Six years on and AI is playing a growing role in the legal profession with LLMs commonly used for tasks such as summarisation, translation and ideation. LLMs can also provide superficially convincing answers to legal questions, particularly where that involves summarising a relatively well-known area of law. However, they struggle when asked to apply a fact pattern or to interpret a clause (see *The LinksAI Benchmark*).

20 *42: GPT's answer to Life, the Universe, and Everything*, Juan Ignacio Specht, 6 October 2023.

21 *Miller v Jackson* [1977] QB 966: "*In summertime village cricket is the delight of everyone. Nearly every village has its own cricket field where the young men play and the old men watch. In the village of Lintz in County Durham they have their own ground, where they have played these last 70 years. They tend it well.*"

22 *Donoghue v Stevenson* [1932] A.C. 562: "*For a manufacturer of aerated water to store his empty bottles in a place where snails can get access to them, and to fill his bottles without taking any adequate precautions by inspection or otherwise to ensure that they contain no deleterious foreign matter, may reasonably be characterised as carelessness without applying too exacting a standard*".



The LinksAI Benchmark

"*Can a pile of bricks be protected by copyright?*" "*What is the difference between reasonable endeavours and best endeavours?*"

These are just two of the questions you might ask your lawyer and expect an accurate answer in return. But what happens if you turn to AI for legal advice? We carried out a comprehensive benchmarking of the GPT 2, 3, 4 and Bard LLMs' ability to provide English law legal advice. The benchmark contains 50 questions covering 10 different areas of legal expertise under English law (such as tax, employment and privacy). Those questions were created by, and the answers marked by, a team of expert lawyers at Linklaters, who measured them against the response expected from a mid-level associate.

The report[23] revealed those systems frequently provide English law analysis that is incorrect and cite laws and cases that do not exist.

Since we published that report, we have benchmarked a number of other systems. Some of those new systems perform at a significantly higher level. One system we tested appeared capable of – for example – preparing a sensible first draft of a summary of the law in a particular area, albeit the summary still needs careful review to verify it is accurate and correct, and that it does not contain fictitious citations. However, even that system struggled with questions involving critical reasoning or clause analysis.

Other studies confirm these conclusions, such as the more recent Allens AI Australian law benchmark[24] which tested GPT-4, Perplexity, Llama 2, Claude 2 and Gemini-1. Even the best-performing LLMs were not consistently reliable when asked to answer legal questions and the citations showed "infection" from larger jurisdictions when citing laws. For example, one answer cited *Navitaire Inc v Jetstar Airways Pty Ltd* [2022] FCAFC 84 (which is fictional and likely to be an attempt to Australianise the English law case *Navitaire Inc v EasyJet Airline Co*. [2004] EWHC 1725).

Our view is still that LLMs should not be used for legal advice without expert human supervision. There are real risks to using them if you don't already know the answer.

23 https://www.linklaters.com/en/insights/blogs/digilinks/2023/october/uk---the-linksai-english-law-benchmark

24 https://www.allens.com.au/insights-news/insights/2024/05/testing-ai-competence-in-answering-australian-legal-questions/. The position in the US is considered in *AI on Trial: Legal Models Hallucinate in 1 out of 6 (or More) Benchmarking Queries*, Stanford University, 23 May 2024.

If full spectrum lawyers still appear to be someway off, other roles are at greater risk. The OECD estimates that, when considering all automation technologies including AI, 27% of jobs are in occupations at high risk of replacement.[25] These impacts are unlikely to be evenly distributed. Low-wage clerical tasks are most at risk of replacement (such as call centre staff), whereas highly skilled manual jobs are likely to be least affected (such as nurses or vets).[26]

It is also likely that in many cases the optimum solution is not fully human or fully robotic and instead is part human and part machine - also known as a "centaur". The shortcomings of AI can be overcome by augmenting human intelligence, and the human can be supercharged by the power of the AI. In other words, the higher-level understanding of a human can be partnered with the speed and bandwidth of a computer.

Organisations will need to consider the wider impacts of this displacement, alongside the need to ensure their workforce has the right skill set to adapt to a changing environment. This might involve reskilling existing employees or hiring employees with different skill sets in the future.

EXAMPLE: A moment's violence

One example of the use of a centaur is subtitles for television programmes. People on television speak in a wide range of dialects, using idioms and non-lexical utterances ("*huh*", "*uh*", "*um*" and so on), and with varying degrees of clarity. This presents real challenges for any speech recognition technology. One way to deal with this is for a human to listen to the audio stream and "re-speak" the words. The speech recognition software can then recognise that clear and uniformly pronounced re-speaking much more easily than the original speaker.[27]

However, this solution is still not perfect. The BBC's subtitling of the Queen Mother's funeral informed viewers *"We'll now have a moment's violence for the Queen Mother"*.

25 *Employment Outlook 2023 - Artificial intelligence and jobs. An urgent need to act*, OECD (2023).
26 *The Potential Impact of Artificial Intelligence on UK, Employment and the Demand for Skills*, UK Government (BEIS), August 2021.
27 *The quality of live subtitling*, Ofcom, 17 May 2013.

“ The ultimate goal of most AI researchers is to create a machine that can match or surpass human intelligence across a wide range of tasks (so-called AGI).

Despite huge leaps forward in the capabilities of generative AI, we are not there yet. HAL 9000, Marvin and Ultron remain safely in the realms of science fiction.[30] However, when AGI does arrive, it may not conform to our anthropomorphic expectations and instead think and operate in a radically different way.”

Peter Church

Artificial general intelligence and the future

This chapter ends with the question of whether these systems are truly intelligent. Can an AI system match or surpass human capabilities across a wide range of cognitive tasks, so-called artificial general intelligence?

There is no doubt that the latest generation of LLMs is hugely impressive. A variety of studies[28] show they have remarkable capabilities across a variety of domains and tasks and are capable of solving novel and difficult tasks that span mathematics, coding, vision, medicine, psychology and more, without needing special prompting. Moreover, in many of these tasks, their performance is close to human-level performance.

Despite this, many consider the LLMs are not actually intelligent nor are they likely to achieve artificial general intelligence. They are just highly sophisticated text prediction models. They are not embodied so do not understand the physical world and cannot plan hierarchically or reason in any reasonable definition of the term.[29]

The future development of AI will likely come from different forms of technology and remains uncertain. However, given the huge investment into AI, it seems inevitable this technology will continue to improve and evolve.

Interestingly, AI models appear to show emergent behaviour that arises once the model achieves specific levels of complexity. In other words, the model’s power doesn’t increase smoothly as the size of the models increases, and instead, there are “breakthroughs” at which component parts of the system start to display self-organising and collective behaviours.

This leads to the disturbing conclusion that artificial general intelligence may eventually manifest itself simply as a result of the size and scale of future models - *de complexitate oritur vita*.

28 *Sparks of Artificial General Intelligence: Early experiments with GPT-4*, Sébastien Bubeck and others, March 2023.
29 *Meta AI chief says large language models will not reach human intelligence*, Financial Times, 22 May 2024.
30 See *2001: A Space Odyssey, The Hitchhiker’s Guide to the Galaxy and Avengers: Age of Ultron* respectively.

02

Overview of AI regulation

Global regulatory landscape

Global investment in AI is rapidly ramping up, transforming every business sector with AI's versatile capabilities. Global governments and regulators face the challenge of whether to rein in AI to manage the risks arising from its unpredictability, on one hand, or to take a balanced and utilitarian regulatory approach, on the other.

Given this juxtaposition, the global regulatory landscape in respect of AI is far from uniform. A patchwork of global rules roughly falls into three varying approaches to regulating AI: (1) comprehensive, risk-based regulations; (2) a voluntary, light-touch approach; and (3) developing incremental legislation.

See our colour-coded map overleaf for an at-a-glance overview of these varying approaches.

Difference in regulatory strategies

- Prescriptive risk-based regulation – Some jurisdictions put forward a comprehensive risk-based approach to regulating AI, including, for example, both the EU and mainland China. The EU AI Act boasts an extensive regulatory and liability regime which is poised to have a global impact in the same way as the EU GDPR. Meanwhile, China has implemented its own stringent requirements for generative AI, algorithmic recommendation management and deepfake programs, and is preparing a state-wide AI law. This state-led approach in China emphasises national security.
- Light-touch and voluntary approach – On the other end of the spectrum, we see regulators adopting a more flexible approach with non-binding guidance and having no immediate plans for AI-specific legislation, potentially waiting to see how efficient other jurisdictions are in enforcing their AI regulations. Markets like Singapore, the UK and Hong Kong SAR have adopted this light-touch strategy to promote safe, responsible AI use and drive innovation through softer voluntary guidelines.
- Incremental legislation – Although businesses tend to appreciate the benefits of a light-touch voluntary approach, there is an underlying move towards stricter regulatory oversight. For example, the United States was historically a follower in data privacy regulation, but now, spurred on by the Biden Executive Order in 2023 on safe, secure and trustworthy AI, it is notably aiming to lead global discussions on AI risk management and compliance. Additionally, individual states in the United States are incrementally rolling out new laws, such as facial recognition in California and Washington, AI hiring restrictions in New York and comprehensive state AI legislation in Colorado.

Increasing international cooperation

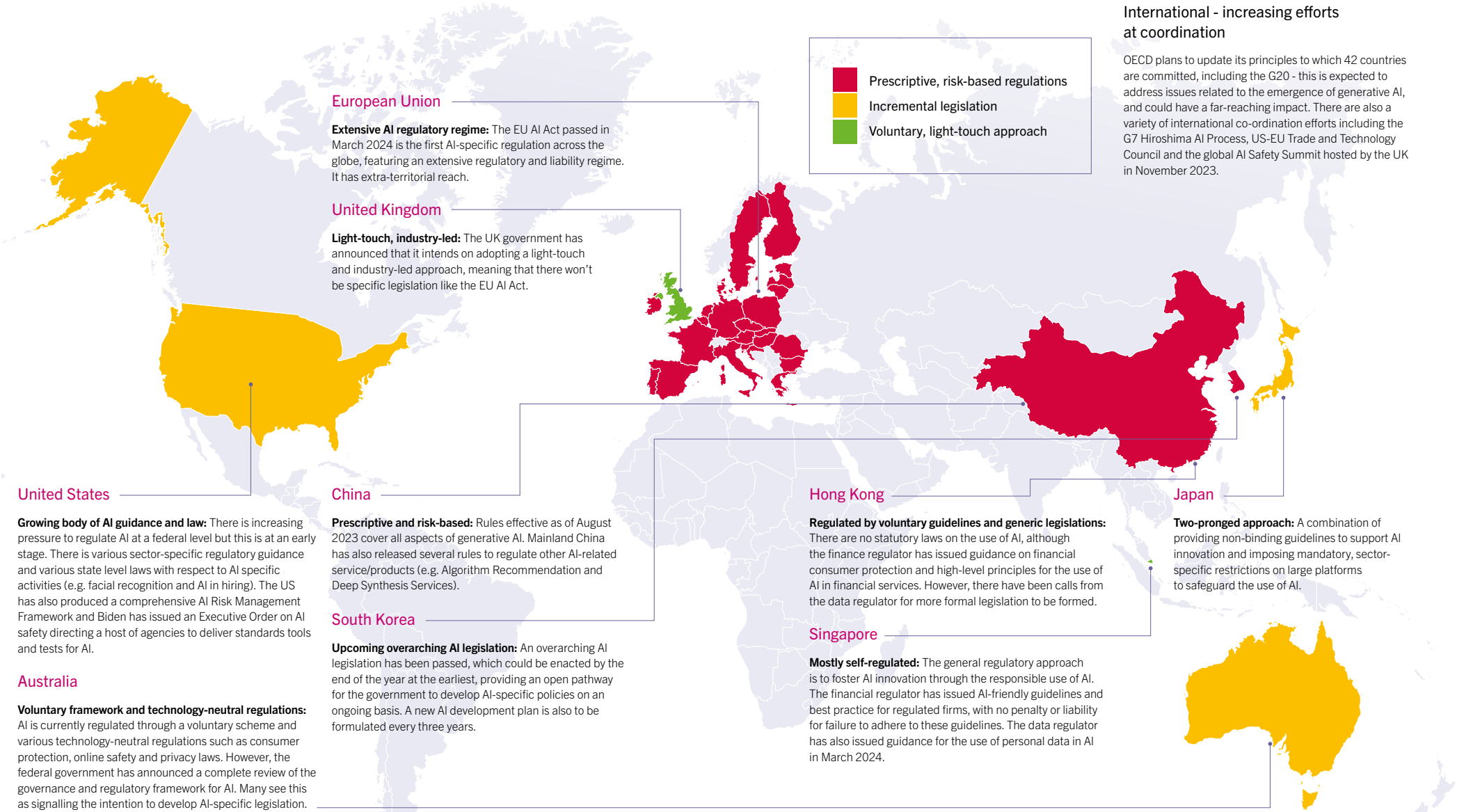
Despite the divergent and far from uniform global regulatory landscape, we are also witnessing converging regulatory themes that focus on transparency, safety, explainability, internal governance and risk management. With this, global governments and regulators are keenly embarking on international cooperation to chart consistent and uniform AI rules. This will be crucial for multinational businesses looking for global interoperability.

In May 2024, the Organisation for Economic Co-operation and Development (OECD) updated its AI Principles, to which 42 member countries are committed, to guide AI actors in developing trustworthy AI. Other international cooperation efforts include the G7 Hiroshima AI Process, US-EU Trade and Technology Council and the global AI Safety Summit hosted by the UK in November 2023.
It remains to be seen if an international consensus can be reached.

AI rules - Everything, everywhere, all at once

International - increasing efforts at coordination

OECD plans to update its principles to which 42 countries are committed, including the G20 - this is expected to address issues related to the emergence of generative AI, and could have a far-reaching impact. There are also a variety of international co-ordination efforts including the G7 Hiroshima AI Process, US-EU Trade and Technology Council and the global AI Safety Summit hosted by the UK in November 2023.



European Union

Extensive AI regulatory regime: The EU AI Act passed in March 2024 is the first AI-specific regulation across the globe, featuring an extensive regulatory and liability regime. It has extra-territorial reach.

United Kingdom

Light-touch, industry-led: The UK government has announced that it intends on adopting a light-touch and industry-led approach, meaning that there won't be specific legislation like the EU AI Act.

United States

Growing body of AI guidance and law: There is increasing pressure to regulate AI at a federal level but this is at an early stage. There is various sector-specific regulatory guidance and various state level laws with respect to AI specific activities (e.g. facial recognition and AI in hiring). The US has also produced a comprehensive AI Risk Management Framework and Biden has issued an Executive Order on AI safety directing a host of agencies to deliver standards tools and tests for AI.

Australia

Voluntary framework and technology-neutral regulations: AI is currently regulated through a voluntary scheme and various technology-neutral regulations such as consumer protection, online safety and privacy laws. However, the federal government has announced a complete review of the governance and regulatory framework for AI. Many see this as signalling the intention to develop AI-specific legislation.

China

Prescriptive and risk-based: Rules effective as of August 2023 cover all aspects of generative AI. Mainland China has also released several rules to regulate other AI-related service/products (e.g. Algorithm Recommendation and Deep Synthesis Services).

South Korea

Upcoming overarching AI legislation: An overarching AI legislation has been passed, which could be enacted by the end of the year at the earliest, providing an open pathway for the government to develop AI-specific policies on an ongoing basis. A new AI development plan is also to be formulated every three years.

Hong Kong

Regulated by voluntary guidelines and generic legislations: There are no statutory laws on the use of AI, although the finance regulator has issued guidance on financial consumer protection and high-level principles for the use of AI in financial services. However, there have been calls from the data regulator for more formal legislation to be formed.

Singapore

Mostly self-regulated: The general regulatory approach is to foster AI innovation through the responsible use of AI. The financial regulator has issued AI-friendly guidelines and best practice for regulated firms, with no penalty or liability for failure to adhere to these guidelines. The data regulator has also issued guidance for the use of personal data in AI in March 2024.

Japan

Two-pronged approach: A combination of providing non-binding guidelines to support AI innovation and imposing mandatory, sector-specific restrictions on large platforms to safeguard the use of AI.

03

Specific AI laws emerge

3.1 The EU: Comprehensive regulation under the EU AI Act

The EU Artificial Intelligence Act (the "**EU AI Act**") marks an important development in the global trend towards the creation of specific AI laws. It was adopted in May 2024 and creates an entirely new regulatory regime for AI systems, with extensive extraterritorial reach.

What is an "AI system"?

"AI systems" are defined in the Act as:

> *"a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."*

That definition mirrors the definition used by the OECD, but from a technological perspective it's not entirely clear what this means in practice, and there remains considerable scope for argument as to whether or not a system operates with "autonomy". Having said that:

- There is likely to be a presumption that certain systems are "AI" as a result of the technology used in that system. For example, there is likely to be a presumption that a system using machine learning is "AI".
- The marketing of the system may be important. Clearly, if a product is described as being "Powered by AI", it may be difficult to subsequently argue it is just a "dumb" product.
- The EU Commission is expected to provide guidance on the interpretation of this term early in 2025.

Ultimately, like an elephant, the concept of AI may be difficult to describe but you will generally know it when you see it.

Tiered regulation based on risk

The obligations under the EU AI Act are tiered based on the purpose for which the AI system is intended to be used.

Some systems are banned outright, while systems classified as high-risk are subject to strict requirements. The EU AI Act also sets rules relating to general-purpose AI ("**GPAI**") models, with an additional set of obligations for the most advanced GPAI models deemed to present systemic risk. The EU AI Act imposes transparency obligations on all providers of human-facing AI systems. The table overleaf summarises those tiers.

Finally, some AI systems fall outside the scope of the EU AI Act altogether such as those used for military purposes, research and development or purely personal use. There is also a limited open source exemption.

Tiers of regulation under the EU AI Act

Tier and description	Regulation
Prohibited These are AI systems that: > Apply subliminal techniques to manipulate behaviour or cause harm > Exploit vulnerable individuals by distorting their behaviour in a manner causing harm > Carry out social scoring causing unjustified or disproportionate treatment > Predict the commission of criminal offences > Create facial recognition databases based on web scraping > Infer emotions in work or education (with exceptions) > Are used by law enforcement for biometric categorisation based on protected criteria (e.g. race or political opinions) > Are used by law enforcement for real-time remote biometric identification (with exceptions)	**Prohibited**
High-risk: Safety Component AI systems used as a safety component in planes, trains, boats, motor vehicles, drones, forestry equipment and quadricycles AI systems used as a safety component in certain medical devices, machinery, toys, watercraft, lifts, protected equipment, radio equipment, pressure equipment, cable cars and gas appliances AI systems used as safety components in critical digital infrastructure, road traffic, or in the supply of water, gas, heating or electricity[31]	**Significant regulation**

31 These systems will not be "high risk" where they do not pose a significant risk to health, safety or fundamental rights

Tier and description	Regulation
High-risk: Purposes Stand-alone AI systems used:[32] > For remote biometric identification, emotion recognition or biometric categorisation according to protected characteristics > In education to determine admission, evaluate important learning outcomes, assess the level of education the individual will receive or detect cheating > In the employment context for recruitment, performance management or task allocation > To determine access to essential public services > To determine creditworthiness (with the exception of fraud detection) > For risk assessment and pricing in relation to life and health insurance > To evaluate and classify emergency calls > For law enforcement as a polygraph, to assess the risk of a person becoming a victim, to assess the reliability of evidence, to assess the risk of offending or in an investigation > For migration, asylum and border control for various purposes > By a judicial authority as a research tool or as a means of alternative dispute resolution > To intentionally influence the outcome of an election or referendum (with exceptions)	**Significant regulation**
GPAI – Systemic risk GPAI models that create systemic risk because of their capabilities, e.g. those requiring more than 10^{25} FLOPs for training (see section below on GPAI models for more details)	**Significant regulation**
GPAI – Other Other GPAI models	**Limited obligations, focusing on documentation and copyright**
Human interaction Providers of AI systems that interact with humans or create synthetic content Deployers of systems that create deepfakes, or carry out emotion recognition or biometric categorisation	**Transparency obligations**
Other – Other AI systems	**Limited regulation, such as AI literacy requirements**

32 See footnote 31

“ The EU AI Act is arguably ”inch wide; mile deep”. The AI systems that are caught by the highest tiers of regulation will be subject to stringent regulation. However, that list is narrowly drawn and the regulation for other AI systems is light touch.

This makes the process of identifying AI systems and correctly categorising them critical to any compliance project.”

Guillaume Couneson

Inch wide; mile deep

The overall effect of this tiered approach is that – in practice – many AI systems are likely to be subject to only limited regulation.

While the list of “prohibited” and “high risk” AI systems is long, each item is narrowly drawn. The EU AI Act is also a Regulation (so directly applicable in every EU Member State) and likely applies maximum harmonisation – i.e. prevents individual Member States from creating their own AI laws within the scope of the EU AI Act.

Arguably, for many organisations this means the EU AI Act has more of a deregulatory effect by both applying very limited obligations to most AI systems and preventing national laws being created to impose extra obligations. Throughout the adoption process, this has also been touted as one of the key benefits of the EU AI Act.

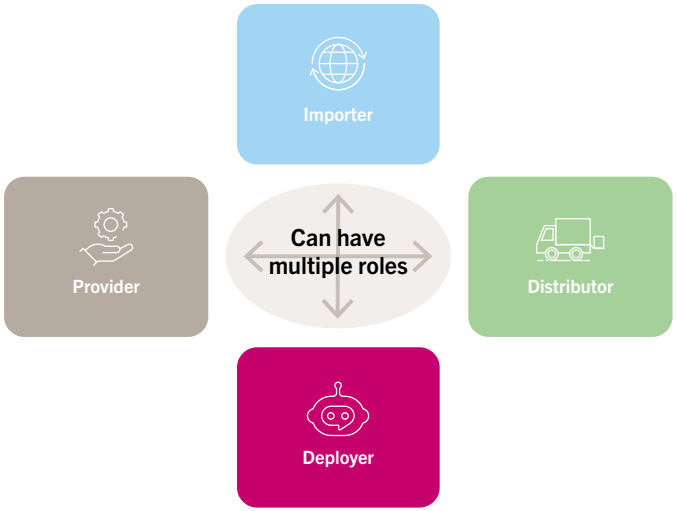
In practice, many organisations are likely to only have a handful of AI systems subject to the strongest tiers of regulation. However, determining with certainty which ones are caught is critical, as the costs of compliance for high-risk systems (and the sanctions for getting it wrong) are significant.

Onerous new obligations for “high-risk” systems

As set out above, there will be very significant new obligations for high-risk systems under the new EU AI Act.

The specific obligations vary according to your role, and the law includes the concepts of a “provider” (being the person who develops the AI system) and “deployer” (being the person using the AI system in a professional capacity).

The most burdensome obligations unsurprisingly fall on the “provider”, though deployers will also have important new obligations (see the table below). There are also separate obligations for “distributors” and “importers”.



High-risk AI systems – Obligations of providers and deployers

Providers' obligations[33]

This is the person that develops the AI system. Importantly, if a deployer, importer or distributor puts their trade mark on an AI system, substantially modifies an AI system or uses it for a high-risk purpose not foreseen by the provider, they will be deemed to be a provider themselves.

- Establish a risk management system with proper risk assessments.
- Ensure the system is trained on high quality data.
- Put in place appropriate technical documentation and record-keeping arrangements and provide instructions for use with the AI system.
- Ensure appropriate human oversight with possibilities to intervene.
- Ensure the system achieves an appropriate level of accuracy, robustness and cybersecurity.
- Put in place a quality management system for post-market monitoring and draw up an EU declaration of conformity.
- Register the system.

Deployer's obligations

This is the person using the AI system in a professional capacity.

- Ensure they use the AI system in accordance with the instructions for use.
- Apply suitable human oversight.
- Monitor the operation of the system (including its robustness and cyber security measures) and keep logs of its operation.
- Inform workers' representatives when used in the workplace.
- Ensure the input data is relevant and sufficient.
- Provide additional transparency where the AI system is used for decision making.
- Notify the provider or distributor, and regulator where the AI system is faulty and suspend use.

GPAI models

GPAI models are those models that are able to perform a wide range of tasks and can be integrated into a variety of different AI systems. Examples include GPT-4, LaMDA and Llama. Providers of all GPAI models are subject to baseline requirements to publish summaries of training material, provide use instructions and technical information for deployers, and comply with EU copyright law (see section 5 (*AI and intellectual property rights*)).

However, particularly complex or advanced GPAI models are considered to present systemic risk. This includes models that meet a computing threshold of 10^{25} floating point operations ("**FLOPs**") or are otherwise designated as presenting a systemic risk by the European Commission. Providers of systemic risk GPAI models must abide by more onerous additional obligations relating to cybersecurity protections, model evaluation and testing, risk assessment and mitigation, and incident tracking and reporting.

33 Most of the obligations in the EU AI Act are disapplied where the "high risk" AI system is used as a safety component in planes, trains, boats, motor vehicles, drones, forestry equipment and quadricycles

The territorial scope is exceptionally broad

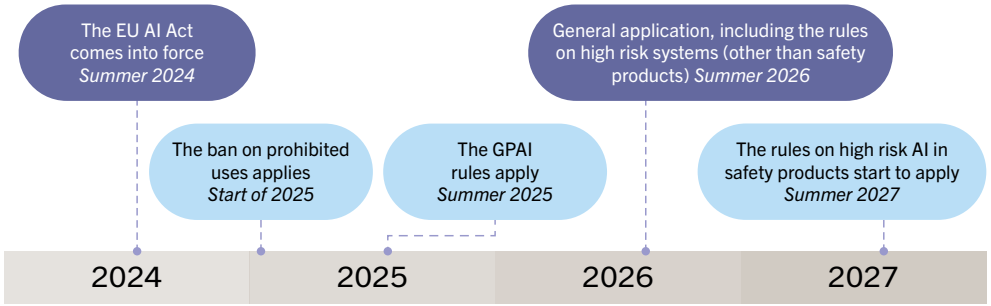
The territorial scope of the EU AI Act is exceptionally broad and so could potentially capture many international organisations with only a tangential connection to the EU. For example, the EU AI Act applies to:

- Providers of AI systems that are put into service or placed on the market in the EU.
- Deployers of AI systems established in the EU.
- Providers or deployers of AI systems where the output from the system is used in the EU.

Where the provider of a high-risk AI system is established in a third country, they must appoint an authorised representative in the EU.

Phased application over three years

The EU AI Act has been adopted and will be published in the Official Journal. Once published it will come into force 20 days later. The substantive obligations will then be phased in over three years.



What should I do now?

While the obligations under the EU AI Act will not apply immediately, it's important to start to prepare for these changes now. The scope of this work will vary from organisation to organisation, but for most there are five key steps:

- Identify the software and hardware products used within, or provided by, your organisation and determine if any qualify as an "AI system".
- For those that do, confirm if those products are caught by the very broad territorial test set out in the EU AI Act.
- For those that are potentially subject to the EU AI Act, determine which tier of regulation that product falls into, noting that (as set out above) only subsets are likely to be classified as either prohibited or high risk.
- Determine what obligations your organisation is subject to and, in relation to high-risk systems, identify if you are categorised as a provider, deployer or other.
- Put a plan in place to comply with those obligations and integrate it within your broader digital regulation compliance framework (such as under the EU Digital Package). Taking a cross-cutting approach to these new regulations is likely to be more efficient, e.g. by combining risk assessments (see section 8 (*Your AI compliance journey*)).

3.2 The UK: Self-regulation for now

The UK government has been very clear that its intention is to establish the UK as a hub for ethical AI innovation post-Brexit.

The current UK view is, broadly, that specific AI legislation would be premature[34] and an industry-led approach would be more realistic. Existing regulators (such as Ofcom, the Information Commissioner, the Competition and Markets Authority and the Financial Conduct Authority) should apply existing law in light of five core principles:

- **Safety, security and robustness:** AI should function in a secure, safe and robust way where risks are carefully managed.
- **Transparency and explainability:** Organisations developing and deploying AI should be able to communicate when and how it is used and explain a system's decision-making process in an appropriate level of detail that matches the risks posed by the use of AI.
- **Fairness:** AI should be used in a way which complies with the UK's existing laws, for example the Equality Act 2010 or the UK GDPR, and must not discriminate against individuals or create unfair commercial outcomes.
- **Accountability and governance:** Measures are needed to ensure there is appropriate oversight of the way AI is being used and clear accountability for the outcomes.
- **Contestability and redress:** Individuals need to have clear routes to dispute harmful outcomes or decisions generated by AI.[35]

Those regulators have already provided a significant amount of guidance on AI, such as the AI auditing framework developed by the Information Commissioner and the current consultations into generative AI.

The key regulators are also members of the Digital Regulatory Cooperation Forum, a body established to ensure a greater level of cooperation given the unique challenges posed by online regulation.

The Digital Regulatory Cooperation Forum has also created an AI and Digital Hub to allow innovators to obtain answers to complex queries which span the regulatory remits of the member regulators.

34 Outside of autonomous vehicles, in respect of which the Automated Vehicles Act 2024 has received royal assent as of 20 May 2024.
35 *UK unveils world leading approach to innovation in first artificial intelligence white paper to turbocharge growth* - UK Government, 29 March 2023.

“ The U.S. is fast developing a complex patchwork of AI laws, including both the application of existing federal law and fast evolving state law. In the absence of pre-emptive federal legislation the U.S. market is likely to become increasingly disjointed and difficult to navigate. ”

Ieuan Jolly

3.3 The US: A patchwork of regulation

The United States has a less developed AI-specific regulatory framework compared to jurisdictions like the European Union, though the landscape at both the federal level and the state level is quickly evolving.

Biden’s Executive Order

To date, the most significant action taken in the United States in furtherance of AI safety has been President Biden’s sweeping executive order titled “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence” (the “**AI Executive Order**”). Issued on 30 October 2023, the AI Executive Order declares an overall objective of establishing new standards for AI safety and security, protecting Americans’ privacy, advancing equity and civil rights, standing up for consumers and workers, promoting innovation and competition, advancing American leadership around the world, and more.

The AI Executive Order identifies eight foundational policy areas, across which more than 50 federal regulators have more than 100 tasks to accomplish over a relatively short period (generally over a timeframe between three months to 12 months depending on the task). Eight notable elements across these policy areas are set out in the table on the adjacent page.

The AI Executive Order is sweeping and consequential, but as an Executive Order of the President, it lacks the durability and depth of a congressionally passed statute. And on that score, there is not yet a comprehensive federal statute specifically regulating AI. In Washington, D.C., Congress is considering a raft of legislative proposals, but as of the date of publication of this handbook, a bipartisan and bicameral agreement on a comprehensive AI statute appears unlikely in the near term.

Eight key points on the US AI Executive Order

1. **Safety and Security** – Developers of AI foundation models that pose a high risk to US security will be required to share their safety test results and other information with the federal government, and such foundation models will need to be tested in accordance with new standards set by the National Institute of Standards and Technology ("**NIST**"). The Department of Commerce will issue new guidance for content authentication and watermarking to label AI-generated content.
2. **Promoting Innovation and Competition** – Government agencies will pilot a new platform to provide AI researchers and students access to data, grants, and other support. New technical assistance and resources will also be made available to small developers and entrepreneurs, and the Federal Trade Commission will more aggressively police against anti-competitive behaviour by larger companies in the AI space. Visa policies will be re-evaluated to expand the ability of highly skilled immigrants with AI expertise to study, stay, and work in the US.
3. **Worker Support** – New principles and best practices will be developed to mitigate the harms and maximise the benefits of AI for workers by addressing job displacement, labour standards, workplace safety and equity, and data collection. Regulators will study and identify options for strengthening federal support for workers facing labour disruptions from AI (see Case Study: Avengers Assemble).
4. **Advancing Equity and Civil Rights** – Federal agencies have been tasked with ensuring fairness throughout the housing and criminal justice system and addressing algorithmic discrimination through training, technical assistance, and guidelines related to AI in housing, criminal sentencing, surveillance, predictive policing, and forensic analysis.
5. **Consumer, Patient, and Student Protection** – The Department of Health and Human Services will establish a new programme to receive reports of – and act to remedy – harms or unsafe healthcare practices involving AI. Federal agencies will also create new resources to support educators who are deploying AI-enabled educational tools, such as personalised tutoring.
6. **Privacy** – In addition to calling for Congress to pass bipartisan data privacy legislation, the AI Executive Order requires an evaluation of how federal agencies collect and use commercially available personal data (e.g. from data brokers) and the development of guidelines for evaluating the effectiveness of privacy-preserving techniques. New funding will assist the development of privacy-preserving research and technologies, including new techniques that use AI.
7. **Ensuring Responsible and Effective Government Use of AI** – New guidance will be forthcoming for federal agencies' use of AI, including clear standards to protect rights and safety, improve AI procurement, and strengthen AI deployment.
8. **Advancing American Leadership Abroad** – Federal agencies are tasked with expanding bilateral and multilateral collaborations on AI, including efforts to establish a robust international framework for harnessing AI's benefits and managing its risks.

Case Study: Avengers Assemble

Unhappy writers

For five months in 2023, the Writers Guild of America (WGA) went on strike and caused its nearly 12,000 members to halt work on screenplays and television shows. Several disputed issues between the WGA and the Alliance of Motion Picture and Television Producers (AMPTP) led to the labour stoppage, but one of the highest profile issues in the negotiations was studios' use of generative AI to develop creative content.

Writers worried that studios would eventually leverage generative AI tools to significantly scale back – or replace entirely – the role of human writers in the television and filmmaking process, which would in turn lead to massive layoffs or pay cuts for film and TV writers, as well as loss of creative control over the end product.

The doomsday scenario envisioned by the WGA was that as generative AI tools get smarter, faster, and cheaper, human scriptwriters may become obsolete. Under the terms of the negotiated settlement, WGA won specific assurances from AMPTP that studios can't use generative AI tools to replace human writers, but merely as a complementary resource.

Unhappy actors

The WGA stoppage overlapped with a separate four-month strike by the American screen actors' union (SAG-AFTRA). The SAG-AFTRA strike revolved around several issues, but also included concerns about AI, including disputes over protections for actors' image and likeness rights in the age of AI-generated content.

Together the WGA and SAG-AFTRA strikes were incredibly disruptive to the film and television industry in the United States, causing numerous productions to be delayed or scrapped entirely. For instance, "***Captain America: Brave New World***" is to be the next blockbuster instalment of the Avengers franchise from Marvel Studios, but its May 2024 release date has been pushed back to early 2025 in part due to the strike action.

A wider public debate

The WGA and SAG-AFTRA strikes of 2023 galvanised a public debate about the ethical and legal implications of using generative AI tools for creative purposes. But more broadly, these labour stoppages also shone a spotlight on the disruptive nature of AI to workers across all industries.

Federal agencies – existing laws still apply

In place of a comprehensive AI law, federal agencies are leveraging existing laws of general applicability, some of which have been on the books for decades. Technology-agnostic and drafted to have a flexible scope, these existing laws prohibit unfair or deceptive practices in commerce, or they prohibit discrimination against individuals in areas like credit and lending, housing, education, healthcare, hiring and firing, and other consequential decision-making.

These laws include the Fair Credit Reporting Act, Section 5 of the Federal Trade Commission Act, and the Fair Housing Act, among others. Federal regulators are coordinating efforts and have repeatedly declared their intent to pursue enforcement actions wherever regulated entities are deploying or developing AI tools in a manner that is contrary to existing law. In a coordinated policy statement, 10 different federal enforcers declared:

> *"Existing legal authorities apply to the use of automated systems and innovative new technologies just as they apply to other practices. The Consumer Financial Protection Bureau, the Department of Justice, the Equal Employment Opportunity Commission, the Federal Trade Commission, the Department of Education, the Department of Health and Human Services, the Department of Homeland Security, the Department of Housing and Urban Development, and the Department of Labor are among the federal agencies responsible for enforcing civil rights, non-discrimination, fair competition, consumer protection, and other vitally important legal protections. We take seriously our responsibility to ensure that these rapidly evolving automated systems are developed and used in a manner consistent with federal laws, and each of our agencies has previously expressed concern about potentially harmful uses of automated systems."*

Every state for itself

In the absence of a federal comprehensive AI law, state legislatures have begun attempting to fill the void by debating their own AI proposals. This follows a familiar pattern already established in the realms of data breach reporting laws (all 50 states have their own state-specific data breach laws) and comprehensive privacy laws (16 states have their own state-specific privacy laws as of the date of this AI handbook).

The state of Colorado blazed a new trail in May 2024 when its legislature passed – and its governor signed – the first comprehensive state AI law. Scheduled to take effect in February 2026, the new Colorado AI law is an anti-discrimination statute that would regulate the use of high-risk AI systems and impose new requirements on developers and deployers of such systems, such as notice, documentation and recordkeeping, disclosures, and impact assessments. The Colorado law may ultimately be amended before it takes effect. And it seems likely that other states may soon follow Colorado's lead along the path of comprehensive AI regulation.

“ In many ways, China is now the global leader in AI regulation having already implemented specific rules for generative AI and with an omnibus national AI law expected shortly. While the approach in China has some resemblance to regulation elsewhere (such as the EU AI Act), other aspects are very different, such as its strong focus on government oversight of societal good. ”

Alex Roberts

3.4 Asia – A full spectrum of regulatory responses

In step with the global picture, AI rules in Asia are in flux: diverse local culture and political landscapes lead to a patchwork of rules and standards across the region. While a light touch and voluntary approach to AI regulation is dominant across the Asia region, we see markets like Japan moving to mandatory rules. Only China is an outlier, with its prescriptive rules that show hints of influence by the EU AI Act.

Light-touch and voluntary approach

Some jurisdictions in Asia have adopted a light-touch approach, so far managing AI through existing laws and an overlay of non-binding and sector-specific guidance. For example, Singapore, which has become a pioneer in leading regulatory discussions in APAC, issued new advisory guidelines on the use of personal data in AI Recommendation and Decision Systems in March 2023, and developed a draft Model AI Governance Framework for generative AI in January 2024. The Model AI Framework, to be finalised in mid-2024, proposes a comprehensive framework calling on all key stakeholders, including policymakers, industry, the research community, and the broader public, to build a trusted and safe AI ecosystem.

In Hong Kong, there are various pieces of soft sector-specific guidance from key regulators; for example, the Hong Kong Monetary Authority and the Securities and Futures Commission have both issued guidelines on the responsible and ethical use of AI. There are also indications of a more proactive governmental approach evolving in the Asian financial centre, as the Hong Kong government first announced in the Chief Executive’s 2023 Policy Address that it will update its copyright laws to track the rapid advancements in AI technology.

Emerging regulations

There are markets in Asia, such as Australia and Japan, that are clearly progressing towards more stringent rules. For example, while AI in Australia is currently managed through a voluntary scheme along with technology-neutral laws (such as consumer protection, online safety, and privacy regulations), the federal government is reviewing the AI governance and regulatory framework and has indicated that it will use a risk-based approach to ensure safe use of AI through AI-specific legislation.

Japan, which had opted for a light-touch approach to AI regulation, recently joined the race of the EU and China to draw up specific AI legislation. Japanese Prime Minister, Fumio Kishida, introduced a generative AI regulatory framework at the OECD in Paris. Recent developments also suggest that Japan’s ruling party is pushing for comprehensive AI legislation to be ready by the end of 2024.

Legislative forerunner

Mainland China represents an outlier in Asia. Even before the EU’s AI Act grabbed the headlines late last year, China was among the first markets to establish an algorithm registry and issue some of the world’s first mandatory rules on generative AI. Since August 2023, licensing has been required for providers of generative AI offering public-facing services, covering all facets of generative AI. China also introduced rules to regulate other AI-related services and products such as algorithm recommendation and deep synthesis (or “deepfake”) services. The Chinese State Council is expected to submit a draft national AI law for review by the country’s legislature later this year.

Alongside regulatory developments, the Chinese courts have been one of the first movers in hearing AI-related disputes. Both the Beijing Internet Court's ruling on the first AI output copyright infringement case in November 2023 and the personality rights infringement case relating to AI-generated voice in April 2024 provide clarity on IP issues when deploying AI (see Case Study: Ultraman vs. AI).

Case Study: Ultraman vs. AI

In February 2024, China's Guangzhou Internet Court issued a verdict in the country's landmark case against a text-to-image AI-generated content provider for infringing the copyright of Ultraman, the iconic Japanese comic character. The court ruled that the content provider violated China's first AI regulation, the Interim Measures for the Management of Generative Artificial Intelligence Services and awarded RMB10,000 (approx. USD1,380) in damages to the claimant.

In April 2024, the Beijing Internet Court ruled on the country's first case regarding infringement of personal rights involving AI-generated voices, after a voice actor's distinctive voice was replicated without consent in various AI text-to-speech applications. The defendants were ordered to apologise and compensate the voice actor for a total loss of RMB250,000 (approx. USD34,500).

These PRC cases not only chart a course for the regulation of AI in China, but also provided food for thought for global regulators in the uncharted territory of enforcement of IP issues in AI deployment.

Harmonisation efforts

In response to multinational businesses' calls for greater consistency within Asia, there is a movement by supranational organisations towards harmonisation across the region. For example, the Association of Southeast Asian Nations (ASEAN) published the non-binding ASEAN AI Governance and Ethics Guide in February 2024, which underscores the trade block's efforts to establish common principles for member states to develop their AI frameworks in alignment with international standards. This trend towards collaboration and regulatory alignment echoes the global regulatory trend, bringing together a diverse patchwork of regulatory regimes across the globe.

04

The enduring importance of data protection laws

> *"One of the problems with GDPR is that it has become the law of everything, and that it's drawing [data protection authorities] into making an awful lot of decisions that impact societies and individuals, which go well beyond data processing."*[36]

4.1 GDPR: The law of everything

Within the EU at least, the scope of specific AI law is narrow (see further discussion the EU AI Act in section 3 (Specific AI laws emerge)). This means that for many AI systems the GDPR[37] will continue to be the key regulatory instrument. Its broad scope and flexible and technology-neutral rules mean it has a central role in the regulation of emerging technology such as AI.

Moreover, the concepts of fairness, necessity and proportionality feature heavily, aligning the GDPR closely with the principles of data ethics. The GDPR introduces new obligations that are well suited to address the challenges raised by AI. This includes duties to carry out impact assessments and restrictions on automated decisions.

The GDPR only applies to the processing of personal data, so will not apply to all AI systems. However, the concept of "personal data" is extremely broad, applying not only to inferences about individuals (even if incorrect) but also to information that does not contain any direct identifiers, but from which an individual is indirectly identifiable.[38]

Guiding principles

The obligations in the GDPR are complex and subtle. It is principle-based legislation, so, by and large, it does not provide explicit direction on what you have to do to comply with the law. Instead, the GDPR provides a series of questions to consider (such as whether the processing is "fair") that are context-sensitive, meaning they must be assessed on a case-by-case basis based on the exact details of the relevant processing.

However, in most cases the aim is to ensure individuals trust you with their personal data. The building blocks to achieve that trust can be summarised relatively briefly:

- Be transparent – Tell individuals what personal data you hold about them and how you are using it, including the fact that you are using AI technology on their personal data. This is backed by detailed disclosure obligations in the GDPR.
- Don't be creepy – Ensure you have a proper justification for using personal data and ensure that data is accurate, not excessive and not kept for longer than necessary.
- Empower individuals – To the extent possible, give individuals meaningful choices and control over how their personal data is used. This is backed up by the rights in the GDPR, including specific protection for individuals who are subject to significant automated decision-making.
- Keep personal data secure – Data security is an important issue for AI algorithms, which often use large amounts of personal data.
- Be fair and avoid discrimination – Consider the impact of the AI model on different groups of individuals and ensure that any difference in their treatment is justified. Do not use sensitive and inherently discriminatory data types, such as information on racial or ethnic origin, unless there is a very good reason to do so. See section 9 (*Ethics and AI*).

36 Helen Dixon, Irish Data Protection Commissioner. See *Two years into new EU privacy regime, questions hang over enforcement*, Politico, 25 May 2020.
37 The term GDPR refers generically to the EU General Data Protection Regulation and the UK GDPR (as retained EU law). At the time of writing, both regimes remain largely aligned in relation to the issues considered in this chapter. The UK Data Protection and Digital Information Bill would have changed some aspects of UK law, particularly in relation to automated decision-making, but did not pass prior to the 2024 general election.
38 See the CJEU's decisions in *Bundeskartellamt* (C-252/21) and *OC* (C-479/22 P) respectively.

“ Data protection authorities appear to be taking a relatively cautious approach to AI. The “techno-optimism” of the early 2000s has been replaced with a more sceptical approach hardened by both suspicion of new technology and the fundamental right to data protection in Article 8 of the EU Charter.

While specific AI regulation is starting to emerge, it does not exempt you from your obligations under the GDPR and the GDPR will continue to be a major factor in any AI compliance plan.”

Sonia Cissé

4.2 Legal basis

One notable feature of the GDPR is that any use of personal data must satisfy one of a list of statutory processing conditions, each of which is referred to as a “legal basis”. This is an exhaustive list of situations in which personal data can be processed, namely:

- Consent – This applies where the individual has given consent. Under the GDPR, consent will only be valid if there is a clear and specific request to the individual, and the individual actively agrees to that use. It is not possible to imply consent and you cannot rely on legalese buried deep within your terms and conditions. This means consent will rarely be appropriate to justify processing in an AI project.
- Necessary for performance of a contract – This applies where the processing is necessary for the performance of a contract with the individual or in order to take steps at the request of the individual prior to entering into a contract. It is not relevant where the contract is with a third party.
- Legal obligation – This applies where the processing is necessary for compliance with a legal obligation.
- Vital interests – This applies where the processing is necessary in order to protect the vital interests of the individual or of another natural person. This is typically limited to processing needed for medical emergencies.
- Public functions – This applies where the processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller.
- Legitimate interests – This applies where the processing is necessary for the purposes of the legitimate interests except where such interests are overridden by the interests of the individual.

This is an issue of significant practical importance, for example whether there is a lawful basis to train LLMs on public data. In practice, it is often necessary to rely on the so-called legitimate interests test which involves identifying what interest is being pursued, whether the processing is necessary for that purpose and where the balance lies between that interest and the rights of individuals. Some regulators expect that balancing exercise to be documented (a “legitimate interests assessment”).

4.3 Automated decisions – "The computer says no"

The GDPR also contains restrictions on automated decision-making, i.e.:

> *"a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her"* (Article 22(1), GDPR).

This includes AI-based decisions to the extent they are significant, such as deciding on loan applications or changing credit card limits.[39]

The extent to which other decisions will be caught depends on the degree of automation and discretion in the decision-making process. For example, these rules are unlikely to apply to automatic summarisation (e.g. where a computer marks a multiple-choice exam) or decision support systems (where the ultimate decision is made by a human). In contrast, they might well apply to a triaging system even if the ultimate decision is made by a human (e.g. where the system identifies potentially fraudulent insurance claims for an in-depth investigation).

Moreover, the CJEU decided that provision of a credit score by a credit reference agency can constitute an automated decision, even though it is not the credit reference agency which makes the actual decision on whether or not to grant credit – instead it is the lender (see *Schufa* (C-634/21)). This is an evolving area of law.[40]

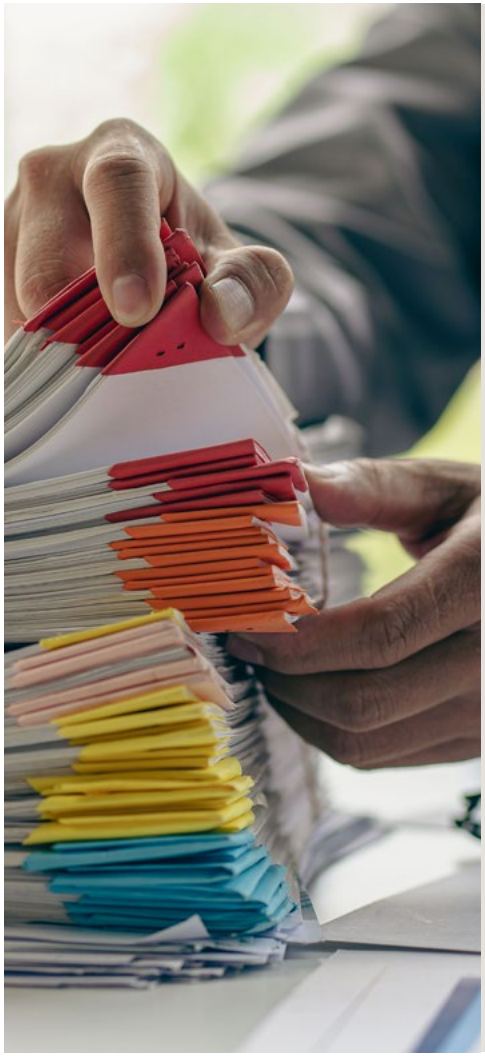
Where the rules apply, automated decision-making is only permitted in the following situations:

- Human involvement – If a human is involved in the decision-making process, it will not be a decision based solely on automated processing. However, that involvement would have to be meaningful and substantive. It must be more than just rubber-stamping the machine's decision.
- Explicit consent – Automated decision-making is permitted where the individual has provided explicit consent. While this sounds like an attractive option, as discussed above the GDPR places a very high threshold on consent and this will only be valid where the individual understands what they are signing up to and has a free choice.
- Performance of contract – Automated decision-making is also permitted where it is necessary for the performance of a contract by, or in order to enter into a contract with, an individual. An example might be carrying out credit checks on a new customer.
- Authorised by law – Finally, automated decision-making processing is permitted where it is authorised by law.

Even where significant automated decision-making is permitted, suitable safeguards must be put in place to protect the individual's interests. This means notifying the individual (see above) and giving them the right to a human evaluation of the decision and to contest the decision. Individuals also have a "right to an explanation" which entitles them to "meaningful information about the logic" involved in the decision. Where the AI taking the decision is "high risk", there will also be new transparency obligations under the EU AI Act.

Beyond the technical requirements of data protection law, there is a wider ethical question of whether it is appropriate to delegate the final decision about an individual to a machine. As Liberty submitted to a recent Select Committee hearing: "*where algorithms are used in areas that would engage human rights, they should at best be advisory*".[41]

39 These will also likely be "high risk" for the purposes of the EU AI Act.
40 See *Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR,* Reuben Binns, Michael Veale, International Data Privacy Law, Volume 11, Issue 4, November 2021.
41 See statement by Silkie Carlo of Liberty in the House of Commons Science and Technology Committee's report on *Algorithms in decision making.*



Automated CV sifting

A large employer uses an AI solution to automatically shortlist candidates for interview. This constitutes automated decision-making as the decision is made solely by automated means and significantly affects applicants.

The employer is permitted under the GDPR to make these automated decisions as they are taken with a view to entering an employment contract with the individual.[42] However, in addition to the general steps described above to ensure fair and lawful processing, the employer must:

- notify applicants that the decision not to shortlist them was taken using automated means; and
- allow the applicant to contest the decision and ask for human evaluation.

In the EU, this will also likely qualify as a "high risk" purpose under the EU AI Act so the AI system will, in parallel, be subject to strict new obligations under that law.

View from the US: Some jurisdictions, such as New York City, have specific hiring laws that also require that the employer notify the applicants of which characteristics the AI tool prioritised in its review, as well as a description of the source data and retention policy. Annual bias audits by an independent third party may also be required.

4.4 Data Protection Impact Assessments

The use of personal data for the development of AI is likely to engage a range of relatively complex issues that require a number of value judgements to be made. In most cases, this evaluation must be documented in a Data Protection Impact Assessment ("**DPIA**").[43]

Many AI projects that use personal data will trip the "high-risk" threshold for a DPIA. Guidance from the European Data Protection Board indicates that processing using new technology or using automated decision-making may need a data protection impact assessment.[44] In the UK, the Information Commissioner has issued a list of activities that will require a data protection impact assessment which specifically refers to "*Artificial intelligence, machine learning and deep learning*" as a potential trigger.[45]

42 *Guidelines on automated decision making and profiling*, Article 29 Working Party (WP251 rev 01), February 2018. Please note that the UK Information Commissioner has suggested automated decision-making cannot be used for this purpose as a human could just as easily complete the review (though this may be optimistic for employers who have a very high ratio of applications to vacancies), see *Six things to consider when using algorithms for employment decisions*, 18 December 2020.

43 Article 35, GDPR.

44 *Guidelines on Data Protection Impact Assessment and determining whether processing is "likely to result in a high risk"*, The Article 29 Working Party (WP 248 rev 01), October 2017. Endorsed by the EDPB on 25 May 2018.

45 The Information Commissioner's *Examples of processing 'likely to result in high risk'*. 22 March 2018.

The DPIA requires an assessment of the project and the risks it raises. It must include:

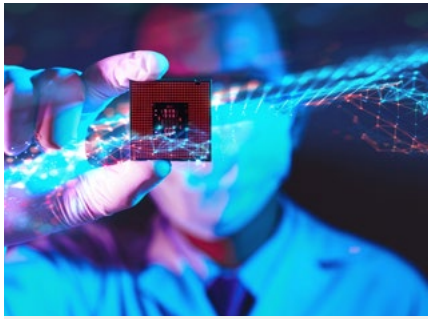
- a description of the processing, including its purposes and any legitimate interests pursued by the controller;
- an assessment of the necessity and proportionality of the processing;
- an assessment of the risks to individuals; and
- the measures taken to address those risks.

The DPIA must involve the organisation's data protection officer (if one has been appointed) and it may be necessary to consult with affected individuals or their representatives. If the DPIA reveals that there are unmitigated high risks, the controller must consult the relevant national data protection regulator about the processing; this process is likely to take several months.

The DPIA should also factor in the specific requirement in the GDPR to ensure privacy by design and default. These obligations require privacy issues to be an integral part of the design process, and that privacy options should, by default, be set to the most privacy-friendly option.

It is important to carry out a DPIA where required. One of the first steps any data protection authority will take when investigating an AI project is to ask for a copy of the relevant DPIA and the UK Information Commissioner recently took preliminary enforcement action against Snap for launching a chatbot without an adequate risk assessment.[46] Similarly, the DPIA should be carried out at the start of the project. Leaving this to the end is likely to result in costly last-minute amendments and delays.

46 *Decision of the Information Commissioner, Snap Inc. and Snap Group Limited*, 21 May 2024.



“ The GDPR takes a flexible and technology neutral approach; indeed, the core principles underpinning the GDPR are largely taken from the earlier Data Protection Directive, which was adopted in 1995.

This broad principle-based approach doesn’t tend to date and, indeed, the principles are as relevant now as they were more than two decades ago. However, it is not clear how those principles apply to any specific situation. This means there are still ‘known unknowns’ when it comes to generative AI’s compliance with the GDPR.”

Daniel Pauly

Existential questions for generative AI

The broad scope of the GDPR and its flexible and technology-neutral rules mean that it is well placed to regulate new technologies, such as generative AI. However, exactly how it applies to new technologies is not always clear and in relation to generative AI models there are some big, and potentially existential, questions about their compatibility with the GDPR. Three key questions for LLMs are set out below.

Use of public training data

LLMs typically require huge amounts of text to learn from – with that text typically coming from a number of public sources, including text from the public internet.

For example, the GPT-3 LLM was reportedly trained on 45 terabytes of compressed plaintext sourced from a variety of sources, including the CommonCrawl dataset (scrapped from the internet), book text and text from Wikipedia. That data will inevitably include personal data about individuals.

What is the legal basis for the use of these massive public datasets? The only possible legal basis is the so-called legitimate interests test. (For example, it would be impossible to get consent from every individual in the training set.) However, the legitimate interests test is highly subjective and depends on a wide range of factors, including the safeguards applied to protect the interests of individuals.

The approach of EU data protection authorities to this issue is not yet settled.[47] However, there are strong arguements that, subject to appropriate safeguards, this legal basis should apply, especially given the recognition of GPAI in the EU AI Act.

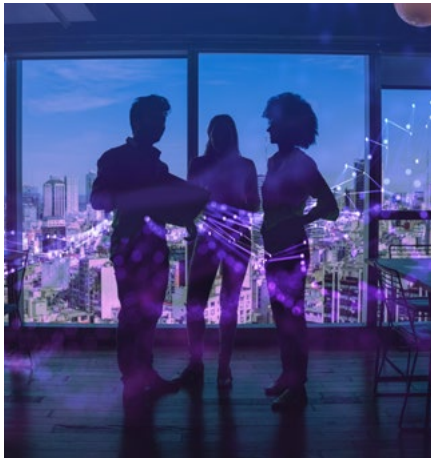
In addition, these datasets will also very likely include some special category personal data (such as the fact that Barack Obama is a member of the US Democratic Party).[48] This type of personal data can only be processed in a narrow range of situations, such as where the individual has provided explicit consent or manifestly made that information public. Given the size of training datasets, it is difficult to see how these conditions would be satisfied in relation to all of the special category personal data contained within them, or how this could be verified in practice.

This is a real conundrum. Various solutions have been suggested to this problem, such as filtering out special category personal data from any training dataset, but this is unlikely to be practicable.

> First, accurately identifying all instances of special category personal data is likely to be impossible given the flexibility of natural language and the fact that this term includes information from which special category personal data can be inferred. For example, take the statements: “*I am researching gallbladder cancer*”, “*I am running the marathon for Cancer UK*”, and “*Tony*

47 The European Data Protection Board has suggested that this condition may be satisfied where safeguards are in place, such as filtering out special category personal data (see *Report of the work undertaken by the ChatGPT Taskforce*, EDPB, 23 May 2024) whereas the Netherlands regulator has suggested that web scraping is almost always illegal and cannot be justified by commercial interests (see *AP: scraping bijna altijd illegal*, 1 May 2024).

48 The scope of special category personal data is potentially broad given that it includes both information that is “intentionally” tagged as special category personal data (such as a datafield used to specifically record an individual’s racial or ethnic origin) and information from which special category personal data can be inferred (such as the names of an individual and their spouse, which could be used to infer sexual orientation), see the EU CJEU’s decision in *OT* C-184/20.



Blair is a cancer on society". Only the last is arguably special category personal data, because it expresses a political opinion.

- Second, removing any special category personal data could result in the AI model being unrepresentative and discriminatory.

An alternative would be to take a more flexible approach to the interpretation of the law, particularly given that the providers of generative AI typically have no interest in their models learning special category personal data about specific individuals. This reflects the approach taken to search engines, which also necessarily process a large amount of special category personal data. In pre-GDPR case *GC* (C-136/17) the CJEU concluded that it was neithe desirable nor possible for search engines to filter out sensitive personal data before returning information – it would effectively prevent anyone from offering a search engine service.

Accuracy

A separate concern is the accuracy of the output from generative AI systems. The LLM simply predicts the most likely sequence of words that follow the user prompt; there is a risk that this output is statistically likely but factually incorrect. There is a question about whether these "hallucinations" are compatible with the accuracy principle in the GDPR. For example, a complaint has been made to the Austrian data protection authority that, when asked for a particular individual's date of birth, ChatGPT makes up various inaccurate answers.[49]

The answer is likely to be that the output from LLMs needs to be presented in context. In other words, the providers of these systems need to warn users they are not generating facts but rather are mere "*probabilistic output creation mechanisms*" with a "*limited level of reliability*".[50]

Individual rights

The final key challenge arises from the rights afforded to individuals under the GDPR, such as the right to access, the right to object and the right to erasure.

Compliance with these rights creates a number of challenges for LLMs. For example, how would you comply with an individual asking for access to all of their personal data contained within a training dataset?[51] What about an individual who asks for their personal data to be deleted from a trained LLM?[52]

In practice, most LLM providers deal with this through output filtering – i.e. identifying situations in which the LLM generates problematic content about a specific individual and then suppressing that content from its output.

49 See *Complaint to the Österreichische Datenschutzbehörde*, noyb, 29 April 2024.
50 See *Report of the work undertaken by the ChatGPT Taskforce*, EDPB, 23 May 2024.
51 As set out above, the training dataset for GPT-3 runs to 45 TB. If "John Smith" makes a subject access request for his personal data then, identifying every reference to "John Smith", "Mr Smith", "John" etc. in that dataset will be computationally expensive and verifying that those references are to the "John Smith" making the request (as opposed to another "John Smith") is likely to be impossible in many cases.
52 It is likely to be impossible to simply remove any references to a particular individual within a trained model. Rather, you would need to remove that individual's data from the training dataset (which is likely to be impractical, see footnote 51) and then retrain the model (which would be extremely expensive).

4.5 Anonymisation

These restrictions make avoiding any processing of personal data an attractive option. That bypasses the need to comply with data protection[53] and some confidentiality obligations, and makes working with third parties much easier.

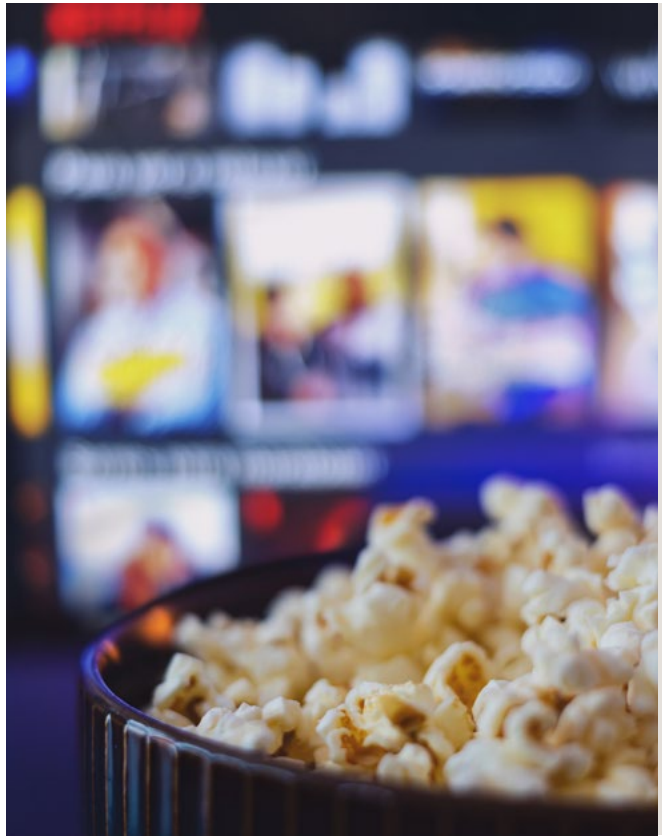
One solution is to use synthetic data. This is artificially created data that replicates the characteristics of the underlying dataset. If created carefully, the synthetic data should not relate to any individual and so its use will fall outside the GDPR.

An alternative is anonymisation, but that is hard as the risk of reidentification must be insignificant.[54] You must consider whether the relevant individual can still be identified from the dataset using:

- Direct identifiers – Simply deleting names and addresses from the data may not be sufficient to anonymise the data if it contains other direct identifiers (e.g. account numbers or telephone numbers).
- A combination of identifiers – It may be possible to identify the individual through a combination of identifiers. For example, the combination of postcode and date of birth will normally identify an individual.[55]
- Combination with other data sources – Finally, the original dataset could be combined with other data sources to identify someone. Determining whether this is the case is a very difficult exercise and there is often not a bright line test to determine when data is identifiable. The UK Information Commissioner recommends a "motivated intruder" test – i.e. considering whether a person who starts without any prior knowledge but wants to identify the persons in the original dataset could identify anyone. As a result, true anonymisation can be very hard. Statements by engineers or business people that the data is "anonymised" should be treated with caution. The Netflix example (right) demonstrates how difficult full anonymisation can be.

That said, even partial anonymisation will normally be a useful exercise and will be a powerful factor justifying this use of the information. However, anonymisation is not always a silver bullet.

53 The anonymisation process itself is a processing that must be justified under the GDPR but will normally be permitted so long as the personal data is truly anonymised.
54 See the CJEU's decision in *Breyer* (C-582/12).
55 In the UK, a postcode identifies around 15 households (though some postcodes relate to a single property) so the combination of a postcode with other information, such as date of birth, will normally identify an individual. In rare cases, postcodes alone will identify individuals.

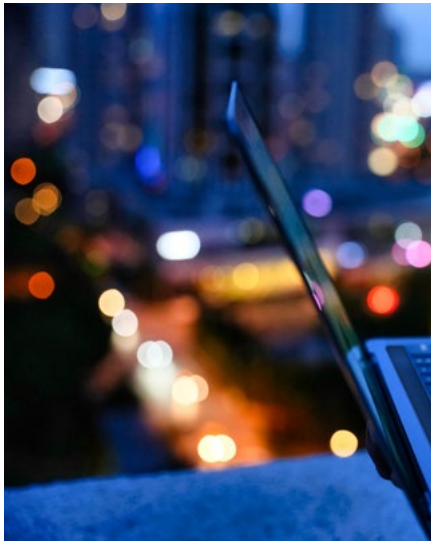


Example: Has Netflix told the world what you are watching?

As part of an annual competition, Netflix released the rankings of 500,000 customers for 100 million films as part of a prize competition to help create a better film recommendation algorithm. The information was pseudonymised (the customers' names were replaced with a key) and "noise" was added to the ratings by slightly increasing or decreasing those ratings.

At first glance, this appears more than enough to protect its customers' privacy. However, researchers found that the combination of ratings formed a distinctive "fingerprint" that could be matched to movie ratings in the public IMDb database (i.e. the films a person likes, and hates, can be a unique and identifying factor). This allowed some customers to be identified.[56]

56 *Robust De-anonymization of Large Sparse Datasets*, Arvind Narayanan and Vitaly Shmatikov, 2006.



Data protection in the US

The United States does not have a single comprehensive data protection law. Rather, it has a patchwork quilt of narrowly tailored laws at the federal level (e.g. to protect health privacy or financial privacy or under-13 children's online privacy) and a growing number of general data protection laws at the state level ushered in by the California Consumer Privacy Act in 2020.

It is difficult to provide a comprehensive summary of this complex patchwork of legislation but, in broad terms, like the GDPR, the state data protection laws define "personal data" quite broadly and create distinct compliance obligations for data processors and data controllers. For instance, risk assessments must be conducted prior to certain high-risk processing activities, which may include the use of automated decision-making tools.

Each state law generally affords residents of that state with rights to request access, correction, or erasure of their personal data, as well as the right to opt out of the "sale" of their personal data or the processing of their personal data for targeted advertising. These laws also generally give individuals the right to opt out of automated decision-making that has significant effects, such as automated decisions relating to housing, education, credit lending, healthcare, or hiring.

The risk assessment and automated decision-making provisions of state data protection laws, along with general requirements for transparency and fairness in processing, means that even in the absence of AI-specific legislation, developers and deployers of AI tools have baseline compliance obligations. Moreover, a number of existing federal anti-discrimination statutes – some of which have been law for decades – would not directly regulate AI tools, but would indirectly prohibit developers and deployers from using these tools for discriminatory purposes in relation to housing, education, credit lending, healthcare, hiring, voting, or other fundamental areas.

For more information, see our detailed overview of US data protection laws: www.linklaters.com/en/insights/data-protected/data-protected---us

“ The privacy laws in Asia are being transformed. In the last couple of years several key jurisdictions, including Australia, the PRC, India, Indonesia, Japan, Thailand, and Vietnam have either updated their existing privacy laws or have introduced entirely new comprehensive data protection laws.

Data protection regulators across Asia are likely to be considering AI just as carefully as their counterparties in the EU.”

Adrian Fisher

Data protection in Asia

In recent years, Asia has seen a growing number of data protection laws which have emerged in response to data privacy challenges brought by emerging technologies and the need for interoperability with the EU to facilitate cross-boundary data flows.

For example, on the one hand, China, Thailand and Indonesia enacted their first comprehensive privacy laws in 2021 and 2022, followed by Vietnam and India in 2023, while on the other hand, jurisdictions with mature privacy regimes, including Singapore, South Korea, Japan and Australia, amended their privacy regimes to align more closely with the GDPR.

These new and refreshed privacy regimes impose stricter data protection standards across APAC with extra-territorial application and distinct compliance obligations for data processors. Businesses are required to be transparent and fair in processing data subjects’ personal data and to improve their internal governance to be accountable to data subjects. In most markets, data subjects are also given the right to data portability, alongside rights to request access or correct or erase errors. Some laws require data controllers using automated decision-making tools to allow data subjects to object to automated decisions where the results may significantly impact individuals’ rights or interests.

Despite the onset of AI-specific legislations recently, these privacy regimes continue to prescribe more comprehensive, technology-neutral rules that AI developers must observe in their personal data management practices.

05

AI and intellectual property rights

With AI use growing exponentially, governments worldwide are scrambling to find the right balance between granting AI developers sufficient freedoms to effectively train and use AI systems, while also protecting the rights of content creators and consumers. Despite the global nature of AI development and deployment, there is very little consensus internationally as to where the balance should lie.

5.1 Inputs – AI training data

Generative AI requires vast amounts of high-quality data to learn. Data is typically sourced from the internet (directly or indirectly), often without permission, and inevitably in a manner that creates copies. The legal treatment of this sort of copying, and whether it may infringe copyright or other IP rights, varies by jurisdiction.

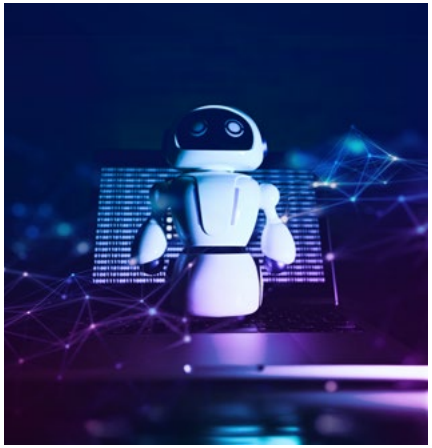
- EU: In the EU, using text and data mining (TDM) methods for commercial purposes on lawfully accessed works is permitted, provided that the rights holder has not opted out "*in an appropriate manner, such as machine-readable means*" (e.g. using a robots.txt file).

  To support that opt-out option, the EU's recently adopted AI Act (see section 3 *(Specific AI laws emerge)*) requires that providers of general-purpose AI models provide detailed summaries of the content used to train those models – a provision clearly designed to increase transparency so that content creators can more effectively police use of their content by AI developers.

  Moreover, the AI Act also requires that any provider placing a general-purpose AI model on the EU market must put in place a policy to comply with EU copyright law, including by respecting any rights holders' opt-outs. The EU therefore appears to be aiming to export its IP laws to all those who provide general-purpose AI models in the EU, regardless of where the model is trained – a significant expansion of a right that is inherently territorial.

- UK: In the UK, TDM is not currently permitted, save for non-commercial purposes or with the consent of the relevant rights holders. A proposal to permit commercial TDM (with no opt-out) was axed following a forceful backlash from the creative industries. Subsequently, the UK Intellectual Property Office failed in its attempts to broker a deal between creative industry stakeholders and AI companies regarding a voluntary code of conduct for AI and IP. The UK government has not ruled out further legislative interventions, but no further proposals have been made to date.

- US: In the US, the fact-specific "fair use" defence to copyright infringement *may* permit TDM in some cases where the developers are able to show that their use of copyrighted material is transformative. However, this AI-friendly outcome is far from certain for most applications of generative AI: there are various lawsuits currently working their way through US courts (see below section on AI and IP litigation) alleging that training AI models on publicly accessible works constitutes copyright infringement. The resolution of these cases should bring some clarity to this difficult area.



“The intersection of AI and IP raises thought-provoking questions. As AI systems generate content in the style of specific individuals, such as creating new Beatles songs or poems reminiscent of Maya Angelou, the issue of ownership becomes crucial. Who owns the copyright for AI-generated works? And how do we protect the creative process itself.”

Pauline Debré

5.2 Outputs – AI-generated content

The outputs of AI models raise a range of issues relating to IP.

> Ownership of outputs: Across much of the world, the outputs of AI models acting autonomously and without human input do not attract IP rights.

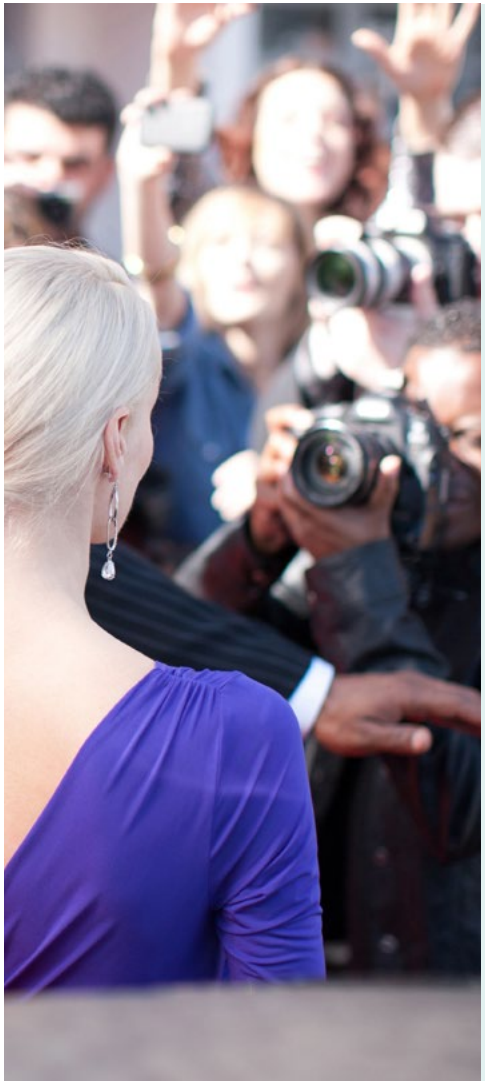
Patent applications that name an AI model as the sole inventor are generally not permitted.

Case study: I, Robot – the rejection of the AI inventor

The law has developed rapidly in this area in recent years due, in large part, to applications made at various intellectual property offices worldwide by or on behalf of Dr Stephen Thaler, the creator of “DABUS” – an AI “creativity” machine capable, according to Dr Thaler, of inventing autonomously. In each case, Dr Thaler’s patent application named DABUS as the sole inventor of the relevant invention. Patent applications made by him in the UK, Australia, at the European Patent Office, and in Germany, India, Korea, New Zealand, Taiwan and the US were all rejected, mainly on the basis that a natural person must be named as an inventor on a patent application, and because there was no legal basis on which he had derived title to the invention. Dr Thaler did manage to obtain a granted patent naming DABUS as the inventor in South Africa, but it is worth noting that the South African intellectual property office does not conduct substantive examination of patent applications and this decision remains controversial.

Where AI outputs comprise written or artistic works, the position varies internationally. In the UK (and certain other jurisdictions including Ireland, New Zealand and India), computer-generated works can attract copyright protection even where there is no identifiable human author. However, in most other jurisdictions, including most civil law European jurisdictions and the US, human authorship remains central to copyright protection, and so copyright is not available for computer-generated works unless there has also been sufficient human input or intervention to support a copyright claim. The US Copyright Office has recently made it clear that merely entering text prompts into an AI image generator, however numerous or prescriptive, will not be sufficient to establish human authorship over the resulting image.

To the extent that IP *does* subsist in the outputs of generative AI models, the statutory framework typically provides that the owner is the person who makes the arrangements for the work to be created – usually the end user. However, it is worth noting that regardless of the default legal position in any jurisdiction, the T&Cs for use of generative AI systems invariably assign to the end user any IP rights in outputs they generate, while also acknowledging that outputs may not be unique across users. This sets up a clear and functional system of broad use rights for all users but will not prevent third parties using the same tools to achieve the same result.



Case Study: Vocal cloning – The ELVIS Act

Vocal cloning is a concerning issue for celebrities, like Scarlett Johansson, who took legal action against OpenAI in May 2024 after hearing a demo of its ChatGPT system which had a voice that sounded "*eerily similar*" to her own.[57] Johansson's position was strengthened by the fact that prior to launching the demo, OpenAI had reached out to her twice seeking to strike a deal for her involvement with the project (although the CEO of OpenAI has denied that the voice was intended to resemble Johansson). Following the receipt of letters from Johansson's attorney, OpenAI has removed the relevant voice "*out of respect for Ms. Johansson*".

While certain existing actions based on rights of publicity and unfair competition may already provide a basis for preventing unauthorised vocal cloning, some US states are now choosing to legislate specifically to protect against AI deepfakes. For example, Tennessee has just amended its laws to introduce the Ensuring Likeness Voice and Image Security (ELVIS) Act – which extends an individual's right of publicity to include their voice and protects against any "*readily identifiable*" unauthorised commercial use of it or any imitation. A new federal initiative along the same lines known as the "No AI FRAUD Act" has also been put before Congress, but has not gained much traction to date.

> **Infringement of third party IP rights:** AI outputs may infringe third party intellectual property and related rights. For example, literary or artistic outputs from generative AI models may not comprise entirely new content but instead reproduce the whole or substantial parts of existing copyright works. Existing IP laws are generally considered fit for purpose to address these issues. However, certain novel uses of AI are prompting new legal solutions. For example, AI can now be used to convincingly imitate a specific person's voice – commonly known as vocal cloning or vocal deepfakes (see Case Study: Vocal cloning – The ELVIS Act).

5.3 AI and IP litigation

As use of AI increases, so does the amount of litigation relating to AI and IP. Many of the most prominent AI developers, including OpenAI, Stability AI, Microsoft and Meta, are facing litigation alleging that their AI operations infringe (or have infringed) third party IP. These proceedings raise fundamental questions as to whether the training and deployment of AI tools are permissible under the current legal framework.

Most of these claims have been brought in the US. The New York Times has sued OpenAI and Microsoft in New York for copyright infringement and DMCA violations, among other things, in respect of use of its content to train ChatGPT, Bing Chat and Microsoft 365 CoPilot. Other high profile US proceedings include various class actions brought by authors against AI developers in respect of their use of written content to train large language models.

57 *Scarlett Johansson's OpenAI feud rekindles fear in Hollywood*, Reuters, 23 May 2024.

“ At its core, the current IP debate is a clash between the creative and tech industries. Creators quite reasonably want to be compensated by those benefiting from their intellectual creativity but equally tech companies want a workable framework to allow them to obtain the data to train new AI models. ”

Paul Joseph

AI image generators are also under scrutiny. In one of the earliest and most high-profile claims to date, stock image provider Getty Images has commenced proceedings against Stability AI in California, alleging that its images have been used to train Stable Diffusion, Stability AI’s image generator, thereby infringing its copyrights; that some of Stable Diffusion’s outputs infringe Getty’s copyrights; and that Stable Diffusion’s outputs recreate and infringe Getty’s trade marks. All of these proceedings are ongoing.

Outside the US, Getty Images has brought a parallel claim against Stability AI in the English courts, and a photographer has brought a claim against LAION (a German non-profit that develops datasets as machine learning resources) which will be heard in the Hamburg Regional Court.

We expect that IP disputes of this type will continue to abound until greater clarity is achieved (whether by legislation or litigation) regarding the legal landscapes in which AI developers operate.

5.4 Trends in AI licensing

Against the backdrop of all this legal action, a new trend is emerging – licensing of content for AI development. Licensing is a key avenue to balance the interests of AI developers and content creators and to avoid or resolve disputes.

Earlier in 2024, OpenAI reportedly struck a deal with News Corp, the media giant which owns the Wall Street Journal, the New York Post, the Times and other news assets. Media sources report that this multi-year agreement gives OpenAI access to current and archived articles and is worth over US$250 million. This is just one of a series of licensing deals agreed between OpenAI and a string of media outlets that are more eager to sign than to sue – including Axel Springer and the Financial Times. Other AI developers are reportedly following suit, seeking licensing deals with content creators of various types.

For the time being, these deals are being heavily negotiated and there is no one-size-fits-all solution. For press publishers, while granting rights over archived materials may be a no-brainer, granting ongoing access to new content raises existential questions: will this push people away from publisher-controlled platforms and into the welcoming arms of AI?

Those entering licensing deals would be well advised not to undervalue their datasets; to include appropriate limitations on how and where the data can be used; to address ownership and use of AI outputs as well as inputs; to limit their liability in respect of any AI outputs that contravene any laws or otherwise give rise to any legal actions; and to carefully consider the circumstances in which they can exit any licensing arrangements, particularly in the light of the uncertain regulatory environment.

“ The current surge in AI-related litigation is likely to carry on for some time while existing issues are addressed and new issues surface. Some of these cases will get all the way to court, though in many cases it is likely they will be settled through new AI licensing arrangements. ”

Shruti Chopra

5.5 Output indemnities – Putting your money where your mouth is

Another interesting emerging trend in the AI and IP space is the practice of AI developers granting IP indemnities to their (commercial) end users. Most of the big players in the AI development space, including OpenAI, Microsoft, Amazon and Google, have provided comfort to their paying customers in respect of the IP litigation risk by offering an indemnity in relation to costs arising from claims of third party copyright infringement.

These indemnities are, on the face of it, drafted fairly and in terms that broadly mirror those typically offered by commercial software providers outside the AI space. However, end users should not be fooled into thinking that these indemnities provide blanket protection against third party IP claims. On the contrary, they are subject to a range of exclusions and conditions that should be carefully considered, negotiated, and navigated in practice.

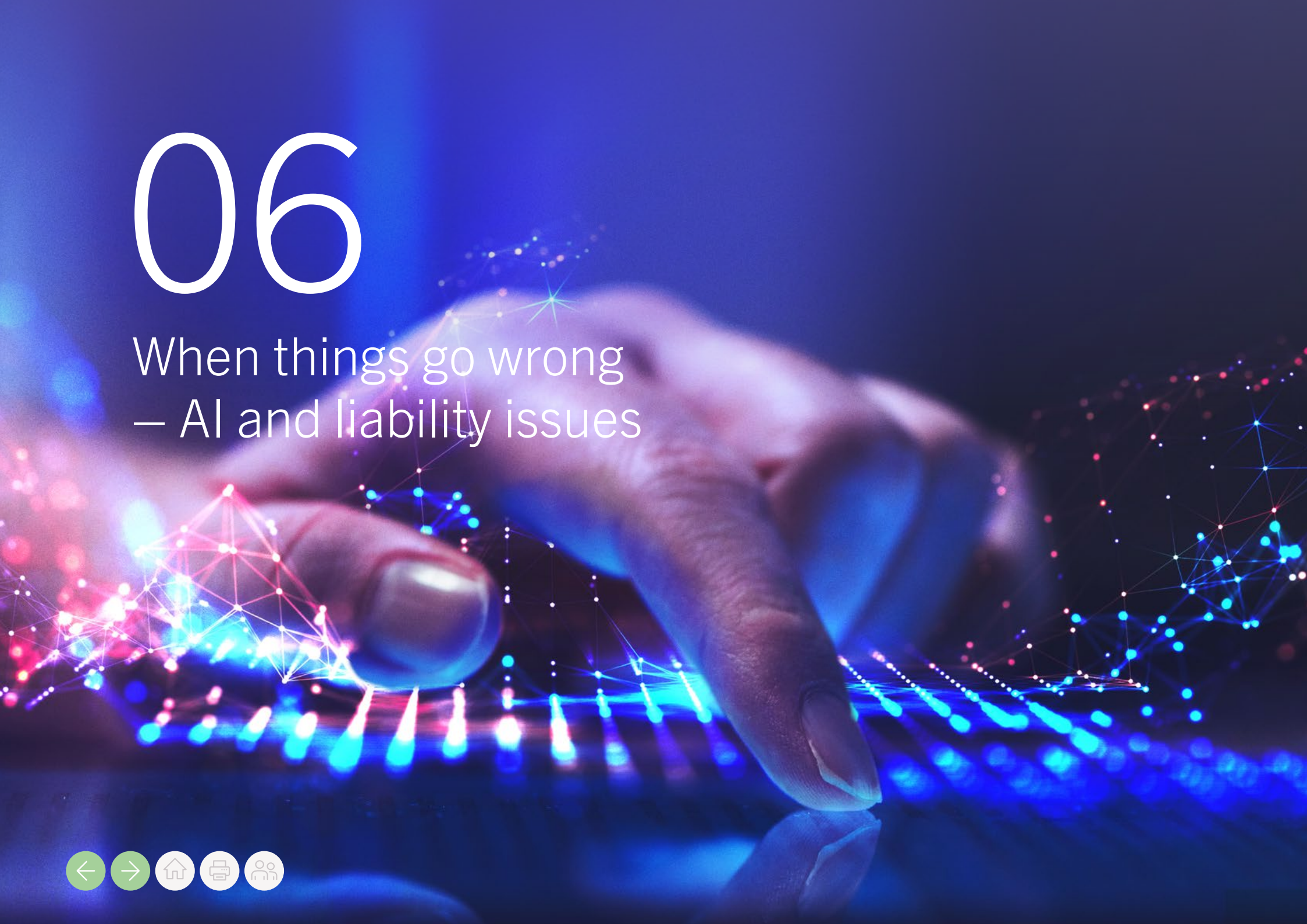
Perhaps most importantly, most of the indemnities offered by major players exclude indemnity protection in circumstances where end users knowingly prompted the infringing response or should have known the response was likely to infringe. Users of generative AI therefore still need to ensure that they have robust internal policies in place to ensure that sufficient third party rights clearance is conducted prior to any use of AI-generated content.

The aim of an AI system is to be intelligent – to analyse, decide, and potentially act, with a degree of independence from its maker.

This is a concern. The algorithm at the heart of the AI system may be opaque and, unlike a human, it has no common-sense safety valve. Delegating decisions to a machine which you do not control or even understand raises interesting issues.

Even though some jurisdictions are starting to revolutionise their approaches to AI regulation, existing laws and evolving regulation in adjacent areas already throw up key issues that need to be treated carefully by any lawyer involved in advising on AI.

We consider the key issues that we are thinking about with our clients below. However, the fast-moving and ever-changing nature of the AI landscape means that the list of issues is likely to evolve and change rapidly over time. You will need to put in place suitable practical measures to supervise the operation of the AI system (see section 8 *(Your AI compliance journey)*) to mitigate these liabilities and meet your regulatory responsibilities.



06

When things go wrong – AI and liability issues

“ The changes to product liability in the EU will significantly increase the litigation risk for AI systems placed on the market in the EU. It will also raise difficult new legal problems, such as exactly when a complex and unpredictable AI system is to be treated as defective.

We expect this to be an area of significant interest for claimant law firms, particularly given the new EU rules on class actions.”

Rupert Bellinghausen

6.1 Product liability

Strict liability could arise under product liability laws. Businesses supplying products to consumers have strict liability obligations to ensure their safety.[58] The term “product” includes “*all movables… even though incorporated into another movable or into an immovable*”. This term would not cover a business’s internal use of AI tools (as they are not provided to a consumer) or web-based access to an AI tool (as there is no product). However, product liability would be relevant if the AI were embedded in a product sold to a consumer.

The EU is intending to provide more protection through the proposed Product Liability Directive. This is driven by the challenges the digital economy and AI impose on the directive’s decade-old definitions and concepts, so the new rules will equally apply to all products, from garden chairs to medical devices. This will be accompanied by a targeted harmonisation of national liability rules under the AI Liability Directive (discussed below). The proposals are far more claimant-friendly than expected. Should they be implemented, the European product liability regime will change drastically.

At the EU level, these risks are aggravated by the prospect of class actions by consumers under the national laws implementing the EU Collective Redress Directive. Companies may well be exposed to major litigation risks by the introduction of these new procedural mechanisms alongside new claimant law firms and the increasing availability of litigation funding. This is a reputational issue, just as much as a legal issue. Social media allows consumers and claimant law firms to target companies with public accusations without any form of fact checking.

58 Primary liability falls on the manufacturer, “own brander” or importer, but distributors can also have liability in more limited circumstances. See the Product Liability Directive 85/374.

6.2 Tortious liability

If you deploy an AI system, you should also consider your potential liability in tort.

In the UK, if there is no tangible damage, the claim is likely to be for pure economic loss. The position here is less clear and will depend on the answers to the following questions: *Does a duty of care arise? Who has a duty of care? A duty of care to do what? What is the impact of any disclaimer?* This, of course, assumes that liability is assessed using traditional duty of care concepts. Given the potential for independent and autonomous action, there have been suggestions that the "owner" of an AI should be subject to more stringent and strict liability for its actions, either through the development of the law of tort (such as an extension to the laws of vicarious liability or the concepts in *Rylands v Fletcher*) or statutory intervention. However, contract lawyers are already seeking to clarify the impact of tortious liabilities via contract.

In the EU, while the Product Liability Directive (discussed above) is the centrepiece of product liability legislation, its application is limited to: (a) claims against the "manufacturer" for damage caused by defective products; (b) material losses due to loss of life, damage to health or property and data loss; and (c) claims made by private individuals.

In addition to these rules, national fault-based liability regimes exist as a "second pillar". To assist claimants harmed by malfunctioning AI systems, the new AI Liability Directive provides a targeted reform of national fault-based liability regimes. It will apply to claims against any person arising from fault in an AI system which caused the damage; any type of damage covered under national law (including discrimination or breach of fundamental rights such as privacy); and claims made by any natural or legal person.

The proposal introduces two key changes:

- the disclosure of information about AI systems; and
- a rebuttable presumption of a causal link between the fault of the defendant and the output produced by the AI system.

Although the proposal is limited in scope, these are far-reaching and important changes. Moreover, the proposal also provides for an evaluation of the directive and envisages further instruments (e.g. no-fault liability rules) should the evaluation deem additional measures necessary.

Artificial Unintelligence – when AI misses the mark

AI can be an incredibly useful tool, but it can also leave businesses with egg on their faces when things go wrong, as these stories illustrate:

High-flying fibs - Air Canada operates an AI-powered customer service chatbot, which became the subject of a lawsuit for "hallucinating" the airline's bereavement discount policy. The chatbot falsely told a customer that he could book a full-fare flight for his grandmother's funeral and apply for a discount after the purchase. Air Canada refused the discount, but a tribunal required it to pay damages to the customer, holding that Air Canada was responsible for the information the chatbot provided.

Swearing software - Like Air Canada, parcel delivery firm DPD introduced an AI-powered chatbot to help answer simple questions from customers. The chatbot made headlines when a user posted on social media about prompting the chatbot to swear and criticise DPD. In one response, the chatbot wrote that "*DPD is the worst delivery firm in the world*". The post went viral, and DPD was forced to disable the bot in light of the error.

Right AI, wrong context - When a mass shooting occurred at Michigan State University, officials at Vanderbilt University emailed students, reminding them to "*take care of each other*". However, students noticed that the email also contained a disclaimer that the text was a "*paraphrase from OpenAI's ChatGPT AI language model*". The university faced a backlash from students for using AI to draft an email about a mass shooting, with one of the email's signatories calling it "*poor judgement*" and senior staff stepping back from their roles following the outcry.

6.3 Contractual liability

Where you provide an AI system or service to a third party under contract, there is a risk of contractual liability if the system fails to perform. However, that liability can be regulated in two important ways.

First, and most importantly, the contract can define the basis on which you provide the system. For example, this might: (a) impose an "output duty" to ensure the output of the system meets specified standards, such as percentage accuracy; (b) impose a lesser "supervision duty" to take reasonable care developing or deploying the system; or (c) make it clear that the system is provided "as is" and that use is entirely at the counterparty's risk. Where you are dealing with a consumer, you would need to ensure that your terms are consistent with the statutory implied terms that digital services and content are of satisfactory quality and fit for purpose.[59]

Secondly, the contract can exclude or limit your liability. These protections would, however, be subject to the normal statutory controls.[60] See section 7 for more detail on contracting for AI systems.

6.4 Intellectual Property Claims

IP claims are likely to be a central concern for anyone seeking to use AI in their business model; for a discussion of key issues in this area, see section 5 (*AI and intellectual property rights*).

59 See Directive (EU) 2019/771 in the EU and the Consumer Rights Act 2015 in the UK.
60 See Council Directive 93/13/EEC on unfair terms in consumer contracts in the EU or, in the UK, the Unfair Contract Terms Act 1977 (business-to-business contracts) and the Consumer Rights Act 2015 (business-to-consumer).

6.5 Cyber threats

Used as an offensive tool, AI has the ability to allow cyber-criminals to launch attacks that can outsmart and outpace many of the rule-based tools currently used to detect cyber attacks. AI can now be used to create deepfakes capable of passing some biometric tests. On the flipside, AI can also be harnessed to analyse huge amounts of data and improve cyber threat hunting and cyber incident response.

The security of any AI system itself will be essential. A breach could have serious consequences including:

- Uncontrolled behaviour – The security breach could allow the hacker to take control of the system. The most visceral example is someone hijacking a medical device, which could result in personal injury or death. However, it is easy to imagine other situations in which an out-of-control AI could cause serious damage (see Case Study: Armed to the teeth – guarding against specific risks).
- Unauthorised disclosure – If the security breach results in personal data being compromised, that could be a breach of data protection legislation. There are mandatory breach reporting obligations in the EU and these are increasingly seen in the United States and in Asia.

6.6 ESG: friend or foe – or both?

AI – and generative AI in particular – has a key role to play in the net zero transition and the wider drive for improved sustainability. For example, it can help organisations make sense of vast quantities of ESG data to help them comply with increasing sustainability disclosure requirements, help compare a company's climate/sustainability performance against its peers, and help the financial sector make better-informed investment decisions.

And it is already proving to be a game-changer in climate monitoring and modelling and in increasing operational efficiencies from smart cities to electricity grids – to name just a few examples.

However, what AI gives with one (ESG) arm, it can also take with the other. Much has been said about the vast quantities of energy needed to power AI systems. For companies (not just tech companies) under pressure to reduce their carbon footprints and meet their climate commitments, this can prove a serious challenge, including feeding into the debate about if and when to use carbon offsets and the risk of potential liability if a company misses its climate targets or does not make accurate or sufficient disclosures about lack of progress towards meeting its publicly-disclosed climate targets.

The use of AI to help comply with ESG disclosure requirements also has its challenges – for example, using AI modelling to fill in gaps in information (for example, gaps in Scope 3 emissions data from a company's value chain). The risks include liability for inaccuracies or lack of transparency when using AI for ESG data collection and analysis, with regulators like the European Securities and Markets Authority confirming that where AI is used in sorting data, for example assessing the ESG credentials of investment data, proper management oversight of the AI system is required for regulatory compliance.[61] Another ESG regulatory regime that is set to have a very significant impact on providers and users of AI is the EU's new environmental and human rights due diligence rules under the incoming CSDDD/CS3D – which will require companies to shift from talking about what they are doing to actually doing (and telling the world at large what they are doing). This is no mean feat for businesses with complex global supply and value chains.

61 ESMA public statement on AI and investment services, 30 May 2024.

6.7 Anti-competitive pricing bots

You should also address the risk that an algorithm might result in anti-competitive behaviour. There are four areas of concern.[62]

The first and least controversial is the **messenger** scenario; where the technology is intended to monitor or implement a cartel – i.e. it is a tool to execute a human intention to act anti-competitively. One example is two poster sellers who agreed not to undercut each other's prices on Amazon's UK website. That agreement was implemented using automated repricing software.[63]

The second concern arises where more than one business is relying on the same pricing algorithm, a so-called (inadvertent) **hub and spoke** arrangement. In *Eturas*,[64] the administrator of an online travel booking system sent out a notice to travel agents informing them of a restriction on discount rates which had been built into the system. The Court of Justice decided that those travel agents could be liable if, knowing of this message, they failed to distance themselves from it. Neither this, nor the first scenario, necessarily involve AI or stretch the boundaries of competition law.

In the United States, the Department of Justice (DOJ) has recently filed a series of statements of interest supporting private antitrust cases trying to move the law on the use of common pricing software as *per se* price fixing. Most recently, the DOJ supported plaintiffs challenging a casino's use of common pricing software for casino hotel rooms, arguing that this was presumptive evidence of a tacit agreement even where pricing recommendations generated by the software are non-binding and there was no evidence of communications between the parties.[65]

The DOJ has also reportedly opened a criminal investigation into real estate software developer RealPage, after filing a statement of interest supporting private claims that it facilitated collusion between real estate brokers by using their competitively sensitive information to generate specific pricing recommendations intended to inflate housing prices.[66]

The third, **predictable agent**, scenario is more interesting. This arises where a number of parties across an industry unilaterally deploy their own AI systems based on fast, predictive and similar analytics, each of which integrates competitors' reactions drawn from data collected from past experience of price variations. This is described as "tacit collusion on steroids",[67] leading to greater transparency between market players and enabling easy detection and punishment of price variations, which in turn could lead to pricing levels that are sustained at higher levels than would otherwise exist in a competitive market. This collusion might take place without any intention by, or possibly knowledge of, the parties.

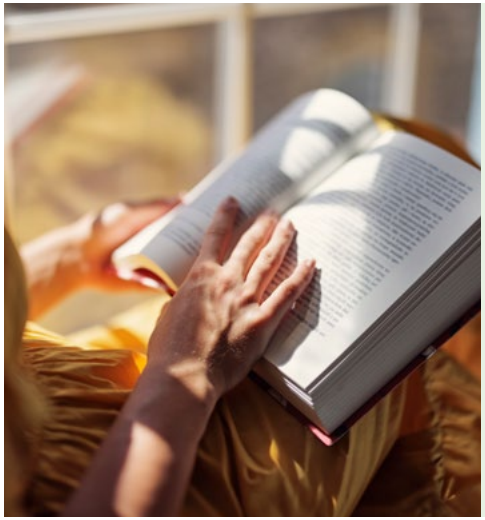
Whether the parties are liable for this anti-competitive behaviour is an open issue. Competition law does not necessarily prohibit the use by a non-dominant undertaking of pricing algorithms that act independently, even if that results in tacit collusion that can lead to higher prices.

62 See *Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy (2016)*, Maurice E. Stucke and Ariel Ezrachi.
63 *Online seller admits breaking competition law*, Competition and Markets Authority, July 2016.
64 *"Eturas" UAB and Others v Lietuvos Respublikos konkurencijos taryba* (C-74/14).
65 Statement of Interest, Cornish-Adebiyi v. Caesars Entertainment, No. 1:23-cv-02536-KMW-EAP (D.N.J. Mar. 28, 2024).
66 Statement of Interest, In re RealPage Rental Software Antitrust Litigation (NO. II), No. 3:23-MD-3071 (M.D. Tenn. Nov. 15, 2023); DOJ escalates price-fixing probe on housing market, Politico, Mar. 20, 2024.
67 See footnote 62.

However, these algorithms are coming under greater scrutiny and EU Competition Commissioner Vestager has said that "*pricing algorithms need to be built in a way that doesn't allow them to collude*".[68] In particular, AI tools are not really intelligent and do not appear by magic. They are domain-specific tools trained for particular tasks, typically through the use of data and a means to evaluate the success of the system's decision. It will be difficult to blame the computer if the training data or success criteria predispose the AI system to anti-competitive behaviour. Indeed, Commissioner Vestager advocates a positive obligation to avoid these effects, so-called "*antitrust compliance by design*".

The fourth situation is the **digital eye**. An all-seeing and fully intelligent AI is able to survey the market and extend tacit collusion beyond oligopolistic markets to non-price factors. This type of AI envisages users of the algorithm being able to tell it to "make me money" and, through a process of "learning by doing", the algorithm reaches an optimal solution for achieving this aim.

This is a developing area of law and one to be watched carefully: "*companies can't escape responsibility for collusion by hiding behind a computer program*".[69] Indeed, in the APAC region, the regulatory approach to addressing anti-competitive practices involving AI and data algorithms is also emerging. India's Competition Commission has examined an algorithmic collusion case (hub and spoke) among ride-hailing platform drivers, but did not ultimately establish collusion.[70] China has incorporated algorithm-related provisions into its monopoly regulations, demonstrating recognition of the competition law implications.[71] Singapore's competition watchdog has warned businesses about algorithm-related risks following a market study on e-commerce platforms.[72]

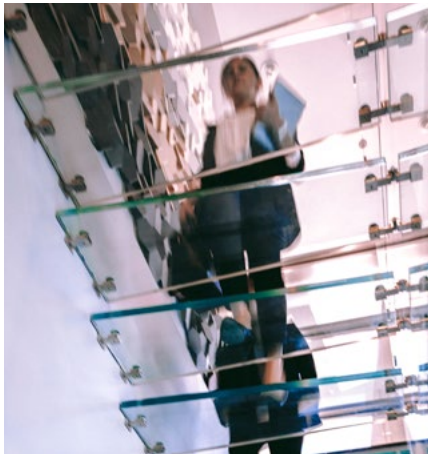


The $23 million textbook

Two companies were selling a textbook called *The Making of a Fly*. One of those sellers used an algorithm which matched its rival's price. The other used an algorithm which always set a price 27% higher than the first.

As a result, the price spiralled upwards. By the time someone noticed what was going on the book was being offered at $23,698,655.93. *The Making of a Fly* is currently available for £16.95 from all good booksellers.

68 *Bundeskartellamt 18th Conference on Competition*, Berlin, 16 March 2017.
69 Commissioner Vestager, *Bundeskartellamt 18th Conference on Competition*, Berlin, 16 March 2017.
70 Case No. 37 of 2018 (CCI).
71 Article 8 and Article 13 on the Provisions on Prohibiting Monopolistic Agreements.
72 CCCS Market Study on E-commerce Platforms Recommends Update to Competition Guidelines, 10 Sept 2020.



“ Use of AI in the financial sector has accelerated rapidly, alongside other digitalisation efforts.

Financial regulators have made clear that AI is not just a “technical issue” to be delegated to AI developers and enthusiasts. Instead, those in C-suite and board-level roles must understand the technology and take responsibility for the big issues and implications. ”

Josh Klayman

Competition considerations: could your collaboration trigger a merger control investigation?

Some collaboration models, e.g. taking an equity stake or entering into a joint venture, might trigger merger control investigations by competition authorities. In some jurisdictions, including the UK, even a minority investment could be reviewed under merger control rules if the acquirer has material influence or de facto control over the target.

The UK’s Competition and Markets Authority has been closely monitoring the impact of partnerships and strategic agreements which could result in the weakening of competition in the development or use of Foundational Models. This has included commencing investigations into Microsoft’s investments in OpenAI, Inflection and Mistral, and Amazon’s investment into Anthropic (although it has concluded that Microsoft/Mistral does not constitute a ‘relevant merger situation’).

6.8 Financial services

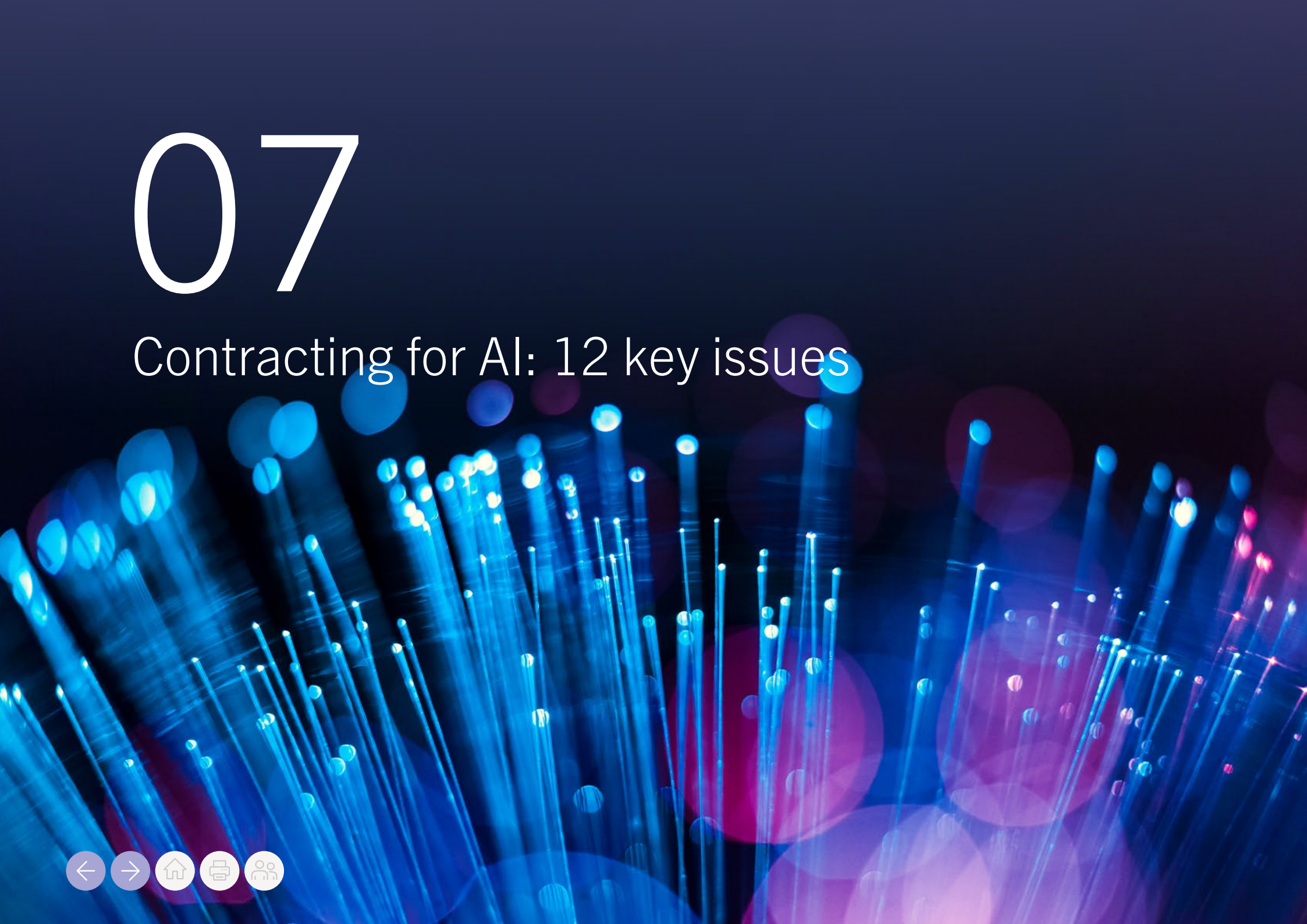
Financial services firms must ensure that their approach to AI reflects the additional regulatory requirements placed upon them.

For further information on AI in the context of financial services regulation, please see our separate report, AI in financial services 3.0.[73]

73 https://www.linklaters.com/en/insights/thought-leadership/fintech/artificial-intelligence-in-financial-services

07

Contracting for AI: 12 key issues



“ In our experience, contracting for AI is quite different from normal technology procurement. One key difference is speed – the pressure to get on the AI train means that the whole process, including the wider compliance assessment, can be significantly compressed. ”

Georgina Kon

With AI markets growing at an exponential rate and AI systems offering ever more exciting opportunities to monetise data, and boost accuracy, productivity and efficiency, it’s no surprise that businesses across all sectors are rapidly investing in AI collaborations, development and procurement.

At a practical level, even before AI compliance programmes have had a chance to bed down fully – and the risks of new types of AI are well-understood – legal teams are often faced with the task of contracting at speed for AI.

While there is no one-size-fits-all for AI contracts, there are some key issues it is worth diligencing, and potentially treating, contractually. Many of these have been discussed in section 6 (*When things go wrong – AI and liability issues*) and are considered in section 8 (*Your AI compliance journey*).

This process starts with the use case. When novel technologies are hot in the market and in the press, it may seem sensible to grab every opportunity to use them. However, many AI projects fail simply because the parties involved have no clear or sufficiently detailed idea of what the use case is. It is important to be clear about benefits the AI is intended to bring and that relevant stakeholders verify the validity and business sense of the use case.

Once that is established, there are 12 key issues to consider.

This assessment needs to reflect the current regulatory landscape. Jurisdictions like the EU are enhancing existing legislation or introducing new rules to cater more specifically for liability for AI systems as well as introducing new transparency and documentation requirements for some AI systems. It will be important for lawyers to understand and treat the impact of this within liability regimes, which already look fairly different to those for other types of technology contracts, see section 3 (*Specific AI laws emerge*).

1.

Has appropriate vendor/product due diligence been performed? How credible are the vendor and any key subcontractors? How commonly used is the product? Is there evidence of credible claims being filed, or regulatory interest, in relation to the vendor or product? Can the vendor comply with your organisation's supplier code of conduct?

2.

Have you validated any dependencies? Often, AI tools/projects will require specific inputs from your organisation. Do you know what these are (e.g. infrastructure, design input, licences, training, collaboration, etc.) and are you confident that your organisation can deliver them and has costed them? If not, can the vendor offer alternative solutions?

3.

Have identified risks been accepted/treated? Your business may have been obliged to conduct a number of risk assessments or diligence – whether for the purposes of AI, data protection, product liability, competition, consumer, ESG, employment, health & safety, online safety and sectoral laws or otherwise. Are there any contractual warranties, obligations, restrictions or indemnities you ought to put in place to better identify or to treat these risks? For what time periods is it appropriate for these to remain in place, and what detailed testing must occur prior to "go live"? If you are contracting for a high-risk AI system, who is responsible for implementing and documenting a risk management system to identify and mitigate risks appropriately?

4.

Who bears the risk and cost of legal compliance? As most lawyers know, it is very easy to write into a contract that everything must comply with applicable laws, but this can be a tricky request in practice. As well as considering the normal issues arising in relation to compliance with law obligations (including on change of law), consider which party should be responsible for compliance with evolving elements of AI regulation and what the cost impact is. What formal role(s), if any, does each counterparty take under AI, product liability and data protection legislation, for example (e.g. manufacturer, provider, deployer, controller)? Are product liability indemnities appropriate? Should either party be able to terminate if new legislation makes your project commercially unattractive?

5.

Are there appropriate assurances around data governance, ownership and quality? Do you know what datasets have been/will be used and what data governance rules will apply? Risk tolerances in respect of data ownership and quality may differ between organisations and in relation to different AI products. For example, do you understand your "house" positions on the use of web-scraped content (which may include inaccurate, biased and/or proprietary data)? Is there a need for data annotation, labelling, cleaning, enrichment and aggregation, for example? Who is responsible for ensuring (to the extent appropriate) that datasets are representative, statistically accurate, non-discriminatory, etc.?

6.

Is data security properly addressed? Ultimately, this is a question that should be answered hand-in-hand with your information security team, who will no doubt have performed diligence to check that they are comfortable with the security of the solution in practice (e.g. cloud-based or not, reliability of upgrades/updates, etc.). Will your data be visible to third parties and can it be used by them (and can any harm be reverse engineered or stopped quickly)? It may be appropriate to require data security metrics and other verification rights, as well as (for non-standardised solutions) to stipulate compliance with specific security standards and policies.

7.

Do you have adequate protections for the usability of outputs? Sometimes, AI contracts are written in such a way as to undermine customer protections by commingling the concept of inputs and outputs, or by making vendor obligations so obscure that it would be difficult to prove a breach of contract. You will also need to ensure that any "requirements specifications", confidentiality and knowledge/technology transfer provisions (and through-life contract management habits) support any planned use of outputs.

8.

Have you considered output verification? AI systems may react unpredictably or unusually to certain input combinations, including producing hallucinations, repeating biases, etc. and this may not be picked up in the testing phase. Therefore, traditional warranties around accuracy are unlikely to be accepted by a supplier. Is human review of output required and are human-machine interfaces supplied? Who is responsible for any human review and who bears the cost?

9.

Can you comply with transparency and related requirements? Under both AI and data protection regulation, as well as sectoral rules, you may have transparency and explainability obligations, including to individuals and to regulators. You may also have obligations to treat customers fairly or to put in place quality management systems, including under sectoral rules. Should the vendor be contractually obliged to provide information to assist you in complying with these obligations and, if so, how frequently? Does the vendor need to keep system logs to assist with traceability of the AI's functioning throughout its lifecycle?

10.

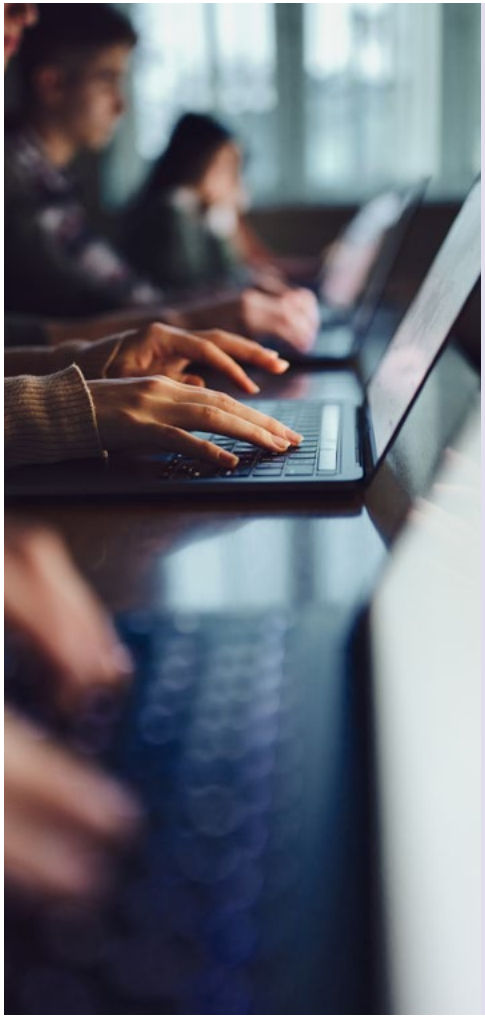
Do you need extended hypercare and training? Depending on what kind of arrangement is being entered into, consider whether you need technical documentation, written instructions for use, heightened support from a vendor during any snagging period after an initial "go live" and access to specific training (e.g. for generative AI, do you need access to a prompt library, guidance/training on prompts?). Will regular updates be needed?

11.

Who owns what? What rights do you need in the AI system, the amendments to that system, the input data and the output data? It is important to consider both positive rights (what do you want to do?) and negative rights (what do you want to prevent others from doing?). It is also worth remembering that broad licences are sometimes capable of achieving your aims without needing ownership of all IP.

12.

Can you exit safely? What support will you need to transition away from the vendor and/or the relevant AI system on exit (e.g. data, tools, knowhow, access to skilled people)? Can you extract input data and/or outputs in a portable and easily readable format? Will/can the vendor destroy your datasets on exit? Will you have the capability in-house to support the exit and transition to the "to be"?



Case study: Armed to the teeth – guarding against specific risks

In February 2024, a Swiss newspaper[74] published a troubling story of a distributed denial-of-service (DDoS) attack that brought down a website causing millions of dollars of damage… and which had been committed by cybercriminals hacking into 3 million smart toothbrushes to create a botnet. Happily, within 24 hours the story was confirmed to be a hypothetical scenario rather than a factual account.

Nevertheless, the connectivity that AI systems both rely upon and support often gives rise to genuine concerns about increased **cyber risk**, and businesses must consider whether their use case merits the introduction of additional contractual security controls to mitigate this. Equally, the numbers involved in this fictional scenario illustrate the growing risk of **class actions** that could arise where AI goes wrong, emphasising the importance of liability and risk allocation in AI contracts.

Other key risk considerations in the context of AI procurement include:

- Arrangements involving algorithmic pricing and data analytics could inadvertently lead to **competition law** concerns, and so these must be carefully contractually structured to ensure they remain legally compliant.
- Where AI is used in relation to tangible products, responsibility and risk must be clearly allocated in the contract in case of any future **product liability** claims.
- If AI is introduced in an **employment** context, the provisions of the contract must reflect employees' legal rights, relating to issues like fairness, privacy and automated decision-making.
- Using AI in the context of financial markets creates risks including discriminatory practices, unpredictable errors and market dislocation, risking non-compliance with **financial regulation**. Once again, contractual arrangements must be clear on the division of roles, responsibilities and liability between the parties in this scenario.

  See section 6 (*When things go wrong – AI and liability issues*) for a more detailed analysis of contentious AI issues.

74 Article published by Aargauer Zeitung on 07 February 2024.

08

Your AI compliance journey

The emergence of AI creates new compliance and governance challenges. There are four reasons for this.

First, the technology is changing fast and organisations need to react quickly to benefit from it and avoid being left behind in this new technological revolution.

Second, the regulatory environment is also changing rapidly as more and more jurisdictions adopt specific AI law (see section 3 (*Specific AI laws emerge*)).

The third factor arises from the inherently unpredictable nature of AI. Historically, organisations have built their compliance programmes around "people" and "things". Most organisations have significant experience trying to regulate the behaviour of the former, whilst the latter is normally fully under an organisation's control. AI is a hybrid between the two and requires new solutions.

The final factor is the pervasive nature of AI which extends across an organisation's functions and products, making it impossible to confine risk management and compliance to the organisation's technology or procurement teams alone.

This all means that your compliance framework must be both flexible (so it can be applied consistently across different functions and as the technology changes) and specific (such that it is workable for individual use cases). The AI compliance programme must also be integrated with an organisation's existing compliance and risk frameworks.

Roadmap to compliance

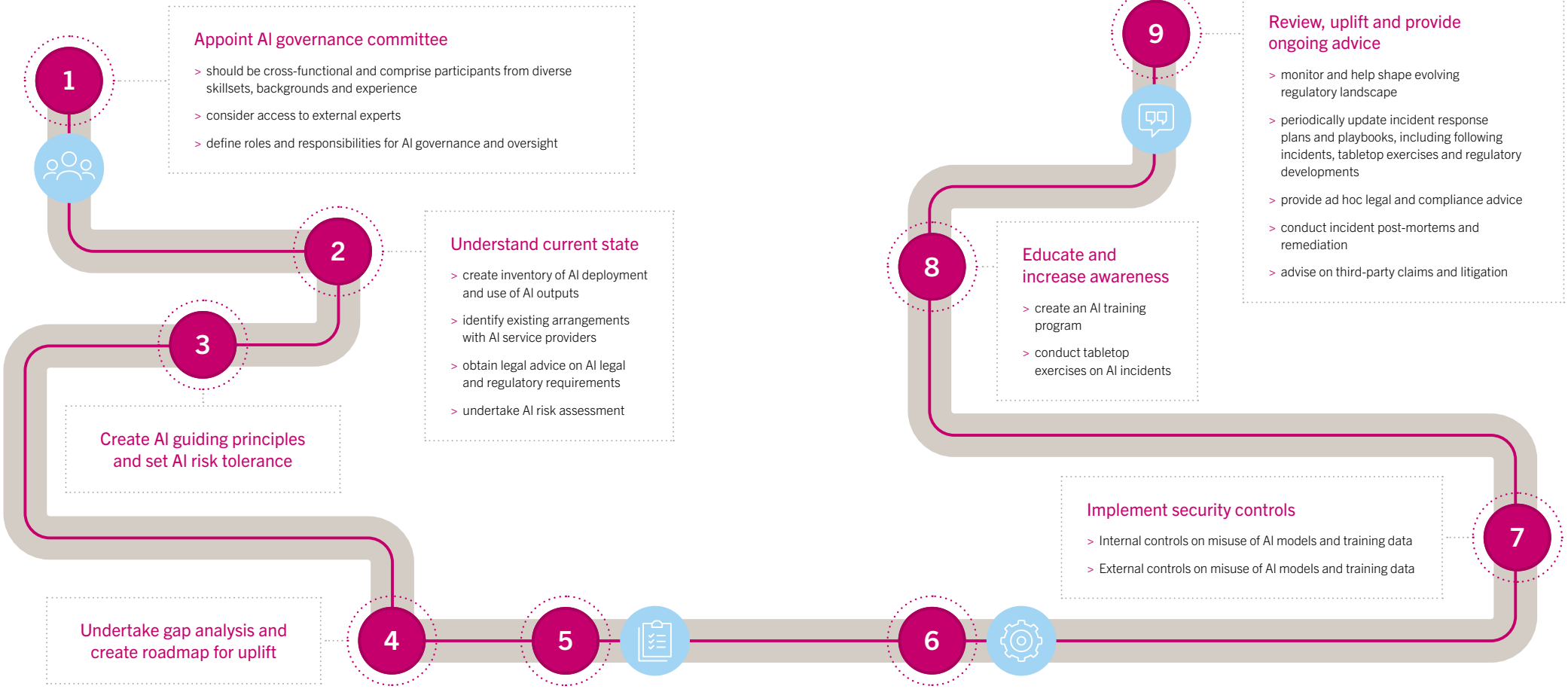
A common challenge is having a compliance plan flexible enough to deal with the rapidly evolving and diverse range of rules relevant to the use of AI around the world. Given regional differences and varying compliance risk levels – like those in the EU related to deploying certain products – organisations should review their AI systems and map their use of AI across to those different rules.

This requires careful balancing. The aim is to not spend excessive time on compliance at the risk of stalling the organisation's ability to leverage these tools. A harmonised, top-down strategy is crucial. This requires board and management to clearly outline the business's goals in relation to the responsible deployment of AI, ensuring AI security and allowing for geographic variances in both AI deployment and regulation.

Conversely, it is also vital for feedback to flow from the different business lines of an organisation up to management, providing an evaluation of the effectiveness of the AI compliance programme that enables dynamic adjustments to the programme where needed. This process must be continuous, keeping pace with the rapid changes in both regulations and technology advancements.

This section provides an overview of a 'compliance roadmap', with some key aspects of a compliance programme, including AI governance and risk management strategies, that are integral to compliant adoption of AI.

Roadmap – setting the direction



“ Any governance structure should not be focused just on the risks of AI but also on the opportunities it provides. This means pushing the business to create a strategic plan that addresses the threats and opportunities of this new technology.

This should recognise that the response may not always deliver an immediate payoff. It may instead follow a J-curve as the initial investment in redesigning business process takes time to deliver a return.”

Harriet Ellis

Start with good governance

A first step towards AI compliance is to establish a dedicated AI governance committee, or expand the remit of an existing committee or management body to encompass AI oversight. This provides a centralised point of accountability for AI decision-making and monitors AI initiatives to ensure that they align with the broader ethical and governance standards of the organisation.

The committee should be cross-functional and comprise participants from diverse skillsets, backgrounds and areas of experience. It is likely that the committee will, in particular, need to include or have access to AI specialists who have a comprehensive understanding of the AI technology under consideration.

The committee should govern with reference to an overarching AI strategy, which should promote responsible use of AI, help identify and proactively manage the risks and ensure AI is used in a way which is fit for purpose and proportionate to the risks.

The responsibilities of an AI governance committee may include:

> creation of AI strategy or guiding principles and setting of a defined AI risk tolerance;

> evaluation of new AI technologies, the assessment of associated risks, and the development of internal policies and procedures to govern the deployment of AI; and

> ongoing monitoring and review of AI systems post-deployment, ensuring that they continue to operate within the defined ethical and compliance frameworks. The committee should have the authority to enforce modifications or even decommission AI systems that fail to meet the required standards.

Inventorise and assess AI

Once governance is in place, the next step is to identify and record the current and proposed AI systems and prepare assessments of those systems. This may involve:

> creating an inventory of AI deployment and use of AI outputs, including which business units are using AI within the organisation;

> identifying existing arrangements with AI service providers;

> obtaining legal advice on AI legal and regulatory requirements; and

> designing and carrying out AI risk assessments, including in relation to any relevant supply chains.

The inventory of AI will help provide an initial view of higher risk use cases where you may wish to focus your compliance efforts and understand the compliance steps that are needed in relation to specific AI implementations.

How to carry out risk assessments

The risk assessment and mitigation cycle will typically follow these four steps:

“ The regulatory environment now requires more and more work to assess and document risk before launching a new product or service. For example, under the GDPR you might well have to carry out a DPIA, LIA and TIA.[75] We are now increasingly seeing Algorithmic Impact Assessments (AIA) being added to that list.

While there are some similarities between these assessments, AIAs are potentially quite different exercises as they can engage a much broader range of fundamental rights and are more closely rooted in the underlying technology. ”

Greg Palmer

How to approach your risk assessment process

- The first step is to identify your compliance footprint and ensure that you are assessing your AI system with reference to the regulations that govern, or are expected to govern, your AI system in various different parts of the world where you operate or where your system will be provided or used.
- The good news is that the principles underpinning AI regulations and proposed regulations overlap to a large extent. You might want to use these similarities to inform the scope of your AI system risk assessment template.
- You'll want to leverage your existing compliance processes where possible. If you are already familiar with similar assessments like DPIAs under the GDPR, you might start by considering the following question: “Should your AI risk assessment be combined with your privacy DPIA?”
- There is clearly going to be value in preparing a single risk assessment that meets both your data protection and AI law demands. By assessing both sets of regulations together in a single document, you won't have separate teams looking at the same issues separately, potentially reaching inconsistent conclusions. Revisiting your risk assessment when your system changes will be, practically, much easier.
- However, there are other factors to consider. Although many AI systems will touch personal data, not all of them will. Whether your template should cater for both regulatory regimes will be informed by how your organisation will use AI. You will also want to think about the risk of having to disclose your risk assessment at some point down the line to your regulators, or in connection with litigation. Are you comfortable sharing all aspects of your AI risk assessment with your data protection regulator when they are only asking to understand the impact of your system on personal data?
- You will also want to think about how your assessment will be completed in practice. Who within the business will “own” the risk assessment? You might choose to structure this so that the business line will be responsible for completing the assessment, with your compliance team providing support. However, the business team may be able to answer factful and technical questions but compliance questions may be alien to it. For example, ‘what key considerations were taken to ensure design fairness’? Your risk assessment template can be supplemented with clear guidance for your business team to explain how to answer this type of question.
- What is the cadence for reviewing and, where needed, updating the (completed) risk assessment, bearing in mind that a risk management system should be a continuous process, even after the system goes live? There are also interesting issues around whether all questions in your template are needed if your AI system is in R&D mode, and so you are ‘in the sandbox’ and not using live data or carrying out testing in real world conditions.

75 Data protection impact assessment, legitimate interest assessment and transfer impact assessment, respectively.

How to identify and manage AI risk

As part of your organisation's AI compliance project, you will want to revisit the process for managing risks that arise from the hidden use of AI in your business and in your supply chain. This means understanding your supplier AI engagement policies and other processes.

Some questions to consider are:

- Who within your organisation will be responsible for identifying (a) your use of AI and (b) the use of AI by suppliers of services to you?
- Are you going to use a 'carrot' method to incentivise your staff to report AI use to you? You could do this, for example, by providing centralised funding, support or infrastructure for AI.
- Or will you use a 'stick' method, such as restricting unapproved AI use in your policies and then monitoring for any use of AI on your systems and then blocking that AI use if it is not centrally approved?
- How will you require your suppliers to tell you about their use of AI? You might choose to incorporate AI-related questions in your supplier due diligence process (both for any new suppliers you choose to onboard and for existing suppliers when you carry out ongoing due diligence on them).

If you've identified that a supplier is using an AI tool to provide services to you, consider the risks that arise from that use and any mitigants you need to put in place.

They could include risks around confidential information or IP infringement if you are using output from the tool. Are your current vendor terms sufficient or do you need to revisit your contractual and other controls? See section 7 for more discussion on contracting for AI systems.

Create or uplift internal policy documents and plans

Your AI compliance programme will benefit from developing AI-specific procedures and policy documents, to provide guidance to teams grappling with how best to identify, assess and manage AI risk. Those procedures and policies are likely to include, for example, appropriate use guidelines for staff, model risk management guidelines for developers, AI incident response plans, risk assessment templates, guidelines for procurement teams, updated supplier due diligence questionnaires, and other templates for capturing technical information relating to AI, so as to assist with transparency or explainability requirements.

You should consider how those policies and procedures fit within your existing compliance structure, such as in relation to acceptable use policies for IT, data protection, non-discrimination and accountability. For instance, an AI non-discrimination policy might require regular reviews of AI systems to ensure they do not perpetuate bias or inequality. Your assessment of AI may also need to be incorporated in your data protection impact assessment process as well as other data protection assessments, including legitimate interest and transfer impact assessments.

Processes should also be updated to allow for transparency in relation to AI usage. This might include updating privacy notices, processes to address customer enquiries about use of AI, processes to evaluate AI systems' alignment with operational and ethical standards, and an AI operational monitoring framework for early issue detection and escalation, as well as regular board engagement on AI.

Organisations will want to maintain a comprehensive AI audit trail, encompassing AI system specifications, data handling procedures, training protocols and performance records. These records are essential for auditing AI systems, understanding their decision-making pathways and outcomes, and providing transparent information to clients and regulators about their use.

Finally, a process for regular review of AI testing feedback will be needed, as well as incident responses and regulatory developments, to ensure the organisation's AI systems continue to operate within evolving legal frameworks and ethical boundaries.

Revisit security controls

Organisations will need to establish a secure environment for AI use that protects against unlawful or unauthorised access to, or potential misuse of, AI systems, including manipulation of those systems and access to data processed by them.

There will likely be other AI-specific cybersecurity risks that are not addressed by current cybersecurity controls; for example, the ability of adversaries to extract training data from large language models or novel risks such as prompt injection. Organisations will want to take particular care where any AI product is public-facing (as compared to "in-house" processes that are not externally exposed), given the increased cybersecurity exposure and reputational risks that a public-facing AI product may give rise to.

Educate your staff

For the compliance programme to be effective, staff will need to be aware of its requirements. Users of AI need to understand its capabilities and limitations, as well as the ethical considerations of its application. In particular, staff should be alive to the tendency of current generative AI models to "hallucinate" and the need to apply their own judgement to, and critically assess, the output from those systems.

Moreover, AI education programmes should cover the importance of data quality and the potential biases that can arise from inadequate datasets, as well as how to spot such issues in practice – see section 9 *(Ethics and AI)*, for more detail.

By equipping employees with knowledge on how to use AI appropriately, how to ask the right questions and, in particular, how to assess AI outputs critically, organisations can start to mitigate known AI risks and foster an environment of compliance and ethical AI use.

Overall approach to AI compliance

The key to a successful compliance project for AI will be its ability to be dynamic and flexible. As AI technology and its applications continue to evolve, the use of AI within your business will grow, and regulation of AI will develop. An agile approach to compliance is needed to help mitigate AI-associated risks, including regulatory sanctions and reputational exposure. The compliance programme will need to cover governance, risk assessments, policies, security measures, education and training, and regular reviews, not only of the company's compliance, but also of evolving regulations and technologies.

By integrating your AI compliance programme within your existing compliance and risk frameworks, while making adjustments to take into account the evolving nature of both AI technology and regulation, you can ensure that your business takes full advantage of the technology while upholding its compliance obligations.

09

Ethics and AI

9.1 Your values

Any business using AI should be mindful of the wider ethical implications of using the technology. This means taking a broad and forward-looking view of the likely implications of AI on your business, your employees, the environment and the communities you interact with.

9.2 Are AI ethics and compliance compatible?

There is an emerging corpus of laws and regulations that apply to an organisation's implementation of AI. Achieving compliance will be a key component of every organisation's ethical approach to technology adoption. However, the lack of geographic and thematic consistency between these overlapping sets of rules means that adopting a narrow focus on requirements in a particular legal area or part of the world when dealing with a globally implemented technology can expose a multinational organisation to censure or criticism elsewhere.

Some examples of these potential clashes include:

- Addressing the potential environmental and workforce impacts of wide adoption of AI while delivering the significant customer benefits and efficiencies.
- Using AI chatbots to achieve accessibility goals without increasing the risk of liability for mis-selling.
- Ensuring that the data used to train AI models is sufficiently comprehensive in terms of scope and volume to provide fair and robust outputs, while at the same time complying with data minimisation rules.
- Seeking to provide customers with adequate transparency and explainability regarding the operation of AI within your organisation, while at the same time avoiding creating vulnerabilities that can be exploited by bad actors.

We have already seen these considerations causing significant issues for major businesses, including chatbots selling cars for $1[76], providing expensive incorrect advice to customers[77] and even swearing at customers and criticising their own company[78] (see *Artificial Unintelligence – when AI misses the mark*). With the expansion of AI being such a high-profile issue, the reputational (and therefore commercial) implications of being viewed as using "AI for Bad" as opposed to "AI for Good" are considerable, and likely to hit businesses a lot faster than the legal liabilities that may follow.

76 *A Chevy Dealership Added an AI Chatbot to Its Site. Then All Hell Broke Loose*, Business Insider, 18 December 2023.
77 *Airline held liable for its chatbot giving passenger bad advice - what this means for travellers*, BBC, 23 February 2024.
78 *DPD error caused chatbot to swear at customer*, BBC, 19 January 2024.

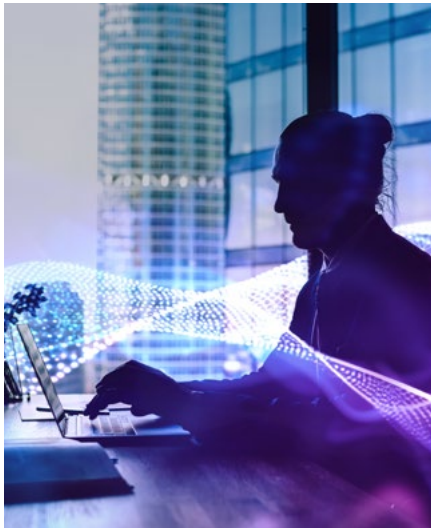
9.3 Striking the balance

When faced with the sort of inconsistencies described above, organisations need to develop a process for identifying and resolving the resulting conflicts, and this necessitates a risk-based approach. A first step is to identify the key ethical considerations for the adoption of AI within your organisation and assessing alignment with your current ethical standards. There are already several international initiatives that have sought to distil the core requirements for "AI for Good"[79] at a level above the more detailed and divergent national rules that are being developed[80].

A review of these international frameworks identifies a core set of fairly consistent principles that can be drawn upon to develop an organisation's own ethical approach. The emphasis of the principles depends on the remit of the relevant body, with some (such as the Bletchley Declaration) focused on public safety issues in areas such as health, transport, energy and the environment, while others are more directed at the steps that companies should take within their own ecosystem. The most relevant principles will depend on the role of each organisation, but some of the key principles include:

- Human-centricity - this increasingly used term influences many of the other principles and emphasises the need for AI developments to prioritise the protection of human rights and development over other technological or commercial imperatives.
- Fairness – working to eliminate bias and discrimination and enhance inclusivity in the design and operation of AI.
- Accountability – ensuring that the organisation remains sufficiently in control of its activities through monitoring, governance and human intervention, and does not seek to disclaim responsibility for the outputs of the technology that it deploys.
- Transparency and explainability – closely aligned with accountability, the organisation should ensure that it is able to explain appropriately to staff, customers, management and other stakeholders how AI is being used and what impacts it is having.
- Sustainable development – use of AI should positively support the development of society and the planet.
- Robustness, security and safety – AI's autonomy and the potential scale of its processing capability require new approaches to relevant aspects of technology, data and operational risk management.

79 Examples include the standards issued by the OECD, G7, ICC and the Bletchley Declaration, while other organisations are supporting "AI for Good" initiatives (e.g. World Bank Group).
80 *Global Competition for AI Regulation, or a Framework for AI Diplomacy?*, The Diplomat, 7 November 2023.



Ethics: the view from the US

In the United States, federal agencies pledge to vigorously use their collective authorities to combat discrimination, bias, and unfair acts in the use of AI that fall within their respective domains. For instance, the Federal Trade Commission (FTC), which focuses on consumer protection, has taken AI-related actions with respect to deceptive or unfair business practices and with respect to unfair methods of competition. The FTC was also one of four federal agencies that in April 2023 issued a "Joint Statement on Enforcement Efforts against Discrimination and Bias in Automated Systems" that disavowed the notion that, in the absence of AI-specific legislation, such technology and systems might be immune from legal challenge.

9.4 Identify affected stakeholders

Examples of the ethical risks posed by AI are already appearing in a number of different areas, affecting different stakeholder groups.

Employees

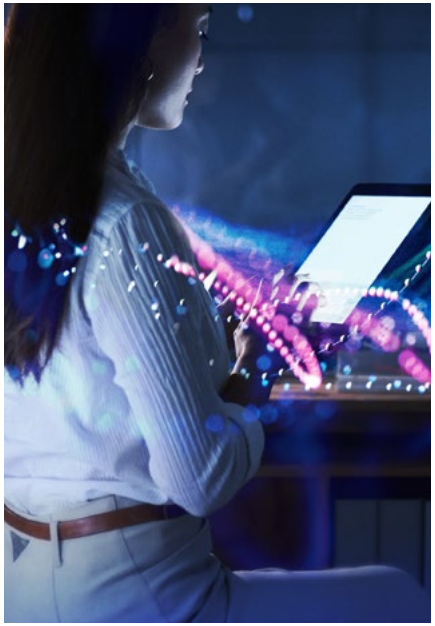
One of the most immediate impacts is upon staff, whose roles are being disrupted by the introduction of AI. The effects are being felt throughout the employment lifecycle, from emergent bias in recruitment processes[81] to role erosion and concerns regarding intrusive surveillance and unfair assessment processes[82].

This has led to various unions undertaking high-profile strikes against the use of AI (see Case Study: Avengers Assemble) and negotiating with big tech companies to mitigate its effects[83]. To sustain employee engagement, organisations will need to invest heavily in training and development to ensure staff continue to have fulfilling and effective roles in conjunction with AI, and for roles which are particularly impacted in the short term it may be necessary to create new rules around how employees can use the time that has been saved in their routines.

Customers

As well as the negative commercial impacts of the ungoverned use of AI chatbots highlighted above, a lack of sufficiently rigorous design and testing is leading to inappropriate customer outcomes in a range of areas. Issues directly affecting customer safety are quick to hit the headlines, with autonomous vehicles frequently in the news[84], and now generative AI is also demonstrating its ability to cause customer detriment, for example in relation to deepfakes[85], where celebrity endorsement via social media offers significant commercial opportunities for bad actors, and alarming advice provided to customers[86].

81 *Ethics and discrimination in artificial intelligence-enabled recruitment practices*, Nature, 13 September 2023.
82 *Employers using AI to monitor workers has negative impact on employees*, CNBC, 8 September 2023.
83 *Workers Are Fighting Back (and Winning) Against AI Overlords*, The Daily Beast, 12 December 2023.
84 *GM's Cruise recalling 950 driverless cars after pedestrian dragged in crash*, Reuters, 8 November 2023.
85 *Beware Of The Viral "MrBeast 10,000 IPhone 15 Pro For $2" Scam*, Malware Tips Forum, 30 October 2023.
86 *Supermarket's AI meal planner recommends recipe for chlorine gas*, The Daily Beast, 10 August 2023.



Public figures

With generative AI's training being largely based on information obtained from trawling the internet, allowing your AI to make public statements is fraught with difficulty without stringent human moderation. Recent examples include X mistakenly analysing tweets regarding a basketball game to accuse a famous player of a vandalism spree[87] and AI-based translation software causing diplomatic incidents[88].

As can be seen from these examples, a lack of sufficient human oversight and control over the deployment of AI can lead to significant ethical and reputational consequences. However, many of these would not have been caught by a typical, compliance-based risk framework, indicating that a different approach is required to managing these risks.

AI and bias

Even where AI technology is being deployed for protective reasons, its results can cause issues. For example, Patrick Grother, a computer scientist at the National Institute of Standards and Technology, when presenting his institute's first major test of the technologies at a recent summit, said that "*Despite rapid advances in accuracy, algorithms estimating a person's age by assessing facial features still show biases against women and people with darker skin tones.*"[89]

AI will also reflect biases inherent in the data it was trained on – for example, Amazon had to recall an AI recruitment tool which had been trained to assess candidates based on historical patterns in applications the company had received over the previous 10 years, because the tool was found to be biased against women since more CVs had come from men in that period.[90]

Not only would these biases negatively impact the people and businesses engaging with the tool, risking negative outcomes and reputational damage, it may also breach the obligation to process data fairly under data protection law (see section 4 *(The enduring importance of data protection laws)*). Therefore it will be vital to be alive to the risk of bias in your AI system, and managing this by developing good ethical governance as detailed (right), and using verification and testing of the system at appropriate points.

87 *Twitter Grok AI made Klay Thompson into story on brick-vandalism spree*, USA Today, 17 April 2024.
88 *Facebook Apologizes for Vulgar Translation of Chinese Leader's Name*, The New York Times, 18 January 2020.
89 Global Age Assurance Standards Summit, Age Check Certification Scheme, Manchester, 8-12 April 2024.
90 *Insight - Amazon scraps secret AI recruiting tool that showed bias against women*, Reuters, 11 October 2018.

“ Artificial intelligence is a powerful but emotive tool. You need to step back and ask yourself; ‘Are we doing the right thing?’, ‘Can I explain the purpose of this project effectively to the various internal and external stakeholders?’.

A project that does not sufficiently consider the concerns your customers or employees, or follow your business values, will be problematic regardless of the technical legal analysis.”

Julian Cunningham-Day

9.5 Managing ethical risks associated with AI – internal governance

Developing an overarching set of principles for the ethical deployment of AI is an important first step in ensuring that the associated risks are given sufficient prominence in risk assessment and approval processes for new AI use cases. Once the principles have been established, they can be used to apply an AI ethics lens to existing governance frameworks within the organisation (e.g. codes of conduct, risk policies and audit programmes) to consider where revisions are required. This approach will assist the organisation in ensuring that ethical issues do not slip between the cracks of those existing frameworks.

Organisations are often surprised at how much change is required to effectively identify and manage these risks. This is in part due to the fact that existing frameworks rely to a large extent on human agency and judgement, and an assumption that important decisions and changes will always occur with sufficient human oversight. Widespread adoption of AI requires an organisation to revisit these assumptions and introduce additional controls. This requires new skills in areas such as data ethics and model risk management and a more holistic approach to managing risk between the various stakeholders, e.g. in HR, technology, operations and legal and compliance. Useful templates for this risk-driven approach can be obtained from standards bodies such as NIST.

Contacts

Technology, digital and data


Julian Cunningham-Day
Partner, London
+44207456 4048
julian.cunningham-day@linklaters.com


Sonia Cissé
Partner, Paris
+33156435729
sonia.cisse@linklaters.com


Tanguy Van Overstraeten
Partner, Brussels
+3225019405
tanguy.van_overstraeten@linklaters.com


Adrian Fisher
Partner, Singapore
+6566925856
adrian.fisher@linklaters.com


Marly Didizian
Partner, London
+442074563258
marly.didizian@linklaters.com


Georgina Kon
Partner, London
+44207456 5532
georgina.kon@linklaters.com


Greg Palmer
Partner, London
+442074562925
greg.palmer@linklaters.com


Albert Yuen
Counsel, Hong Kong SAR
+85229015068
albert.yuen@linklaters.com


Alaister Johnson
Partner, Singapore
+6563215229
alaister.johnson@linklaters.com


Rich Jones
Partner, London
+442074563316
rich.jones@linklaters.com


Guillaume Couneson
Partner, Brussels
+3225019305
guillaume.couneson@linklaters.com


Alex Roberts
Partner, Shanghai
+862128911842
alex.roberts@linklaters.com


Nick Roudev
Counsel, Dubai
+97143695807
nick.roudev@linklaters.com


Daniel Pauly
Partner, Frankfurt
+496971003570
daniel.pauly@linklaters.com


Ieuan Jolly
Partner, New York
+12129039574
ieuan.jolly@linklaters.com


Richard Cumbley
Partner, London
+44207456 4681
richard.cumbley@linklaters.com


Ceyhun Pehlivan
Counsel, Madrid
+34913996182
ceyhun.pehlivan@linklaters.com


Michael Leicht
Partner, Frankfurt
+496971003463
michael.leicht@linklaters.com

Intellectual Property


Julia Schönbohm
Global Head of TMT/IP
+496971003 138
julia.schoenbohm@linklaters.com


Bolko Ehlgen
Partner, Frankfurt
+496971003174
bolko.ehlgen@linklaters.com


Yohan Liyanage
Partner, London
+44207456 4623
yohan.liyanage@linklaters.com


Shruti Chopra
Counsel, New York
+12129039448
shruti.chopra@linklaters.com


Jennifer Heley
Counsel, London
+44207456 5477
jennifer.heley@linklaters.com


Paul Joseph
Partner, London
+442074562411
paul.joseph@linklaters.com


Pauline Debré
Partner, Paris
+3315643 5796
pauline.debre@linklaters.com


Katie Coltart
Partner, London
+442074563681
katie.coltart@linklaters.com


Gaëlle Bourout
Counsel, Paris
+3315643 5620
gaelle.bourout@linklaters.com


Atif Wolfgang Bhatti
Counsel, Frankfurt
+496971003610
atif.bhatti@linklaters.com

AI Litigation


Doug Davison
Partner, Washington, D.C.
+12026549244
doug.davison@linklaters.com


Ben Packer
Partner, London
+442074562774
ben.packer@linklaters.com


Andrew Poulton
Partner, London
+442074563301
andrew.poulton@linklaters.com


Ellen Qidi Zhang
Partner (Zhao Sheng), Shanghai
+861065350 468
qidi.zhang@linklaterszs.com


Kerstin Wilhelm
Partner, Munich
+498941808 506
kerstin.wilhelm@linklaters.com


Harriet Ellis
Partner, London
+442074565515
harriet.ellis@linklaters.com


David Thomas
Partner, London
+442074563059
david.e.thomas@linklaters.com


Satindar Dogra
Partner, London
+442074564316
satindar.dogra@linklaters.com


Jelita Pandjaitan
Partner, Singapore
+656692 5881
jelita.pandjaitan@linklaters.com

Product Liability

Rupert Bellinghausen
Partner, Frankfurt
+496971003470
rupert.bellinghausen@linklaters.com

Harriet Ellis
Partner, London
+44207456 5515
harriet.ellis@linklaters.com

James Warnot
Partner, New York
+19177440971
james.warnot@linklaters.com

Employment

Sinead Casey
Partner, London
+442074562723
sinead.casey@linklaters.com

Laure de Panafieu
Partner, Singapore
+6566925791
laure.de_panafieu@linklaters.com

Marie Sinniger
Counsel, Luxembourg
+35226088451
marie.sinniger@linklaters.com

Andrew Gaines
Partner, New York
+12129039478
andrew.gaines@linklaters.com

Matthew Devey
Partner, Frankfurt
+496971003 209
matthew.devey@linklaters.com

Competition and Consumer

Verity Egerton-Doyle
Partner, London
+442074563389
verity.egerton-doyle@linklaters.com

Marcus Pollard
Partner, Hong Kong SAR
+85229015121
marcus.pollard@linklaters.com

John Eichlin
Partner, New York
+12129039231
john.eichlin@linklaters.com

Annamaria Mangiaracina
Partner, Brussels
+32479666668
annamaria.mangiaracina@linklaters.com

Fay Zhou
Partner, Beijing
+861065350 686
fay.zhou@linklaters.com

Will Leslie
Partner, Brussels
+3225019047
william.leslie@linklaters.com

Finance

Martyn Hopper
Partner, London
+442074565126
martyn.hopper@linklaters.com

Josh Klayman
Senior Counsel, New York
+12129039047
joshua.klayman@linklaters.com

Peiying Chua Heikes
Partner, Singapore
+6566925869
peiying.chua@linklaters.com

ESG

Rachel Barrett
Partner, London
+442074565414
rachel.barrett@linklaters.com

Vanessa Havard-Williams
Consultant, London
+44207456 4280
vanessa.havard-williams@linklaters.com

linklaters.com/ai

Edited by Peter Church and Ria Moody.