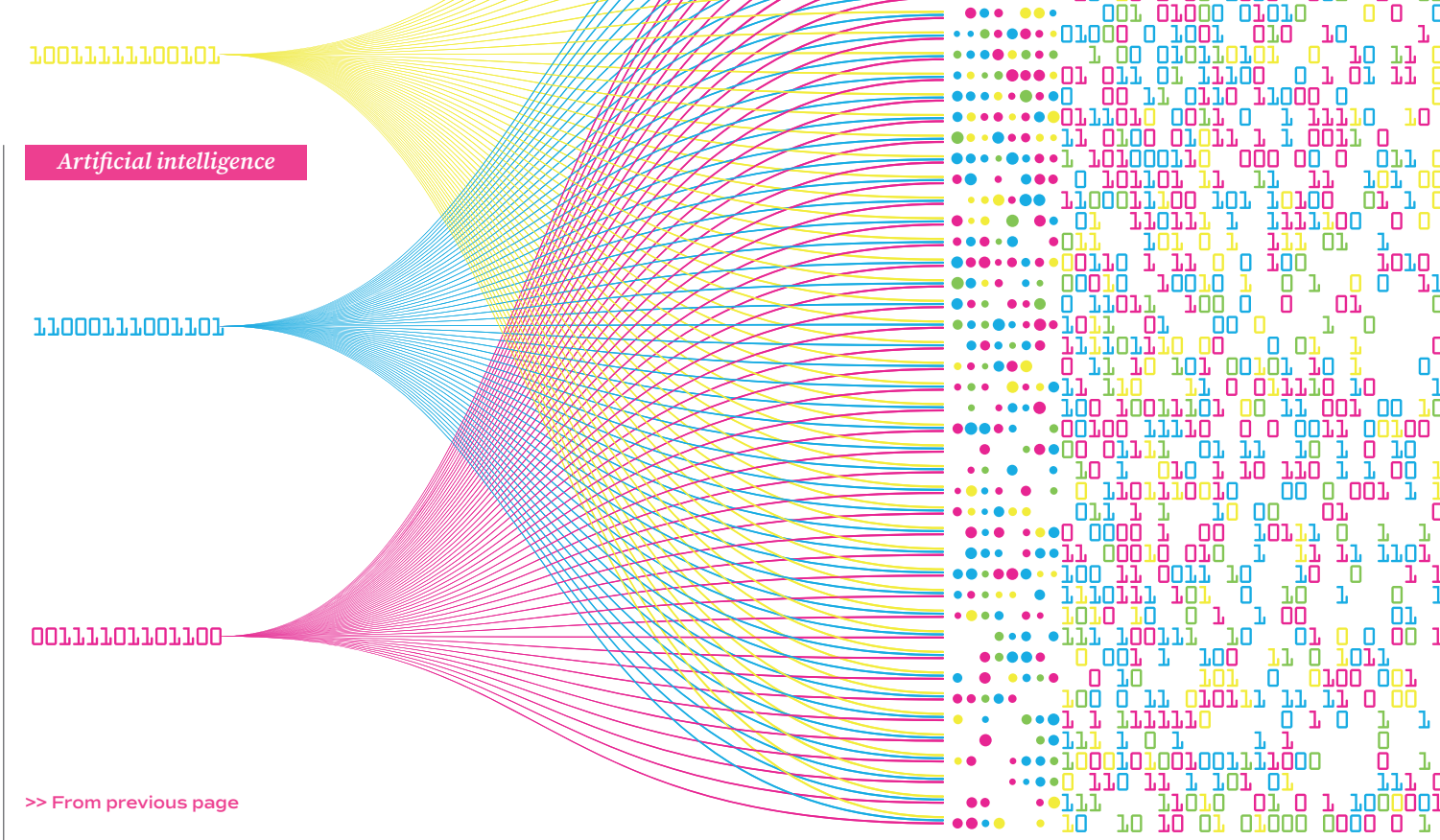
MITIGATING THE AI RISKS

Are you aware of the issues you face with generative AI, how best to take advantage of the opportunities and manage the potential dangers? Thomas Stables, Tom Sharpe, Emily Tombs and Amy Moylett of international legal firm Osborne Clarke tackle the issues for risk managers

AI is nothing new. What is new is the pace of change. This is particularly notable in a category of AI known as generative AI (GenAI). While these systems are very complex in their operation and the maths which they use, their practical functions and their implications for businesses are comparatively straightforward to understand.

The key concept to get to grips with is that AI systems do not take decisions or have understanding in the same way that humans do. GenAI systems like ChatGPT are based on mathematical algorithms which identify patterns in data and create detailed maps of enormously vast amounts of data. These datasets enable the AI model of the patterns to become increasingly detailed and accurate, to the point that it is able to reliably predict the next steps for a given problem or question.

Continued overleaf >>



>> From previous page

When GenAI systems respond to an input or instruction, their outputs or responses are not generated by searching for the right answer amongst their massive training datasets, like a search engine checking its index of websites. Instead, GenAI uses its training to predict what is statistically most likely to be the right answer, thereby generating its response. ChatGPT is trained to produce fluent and coherent text and can have a conversation. It does not pull through answers from its training data but generates what the underlying model predicts (based on statistical probability) will be the 'correct' answer, based on the datasets that it has been trained on.

While this concept is simple to grasp, the complexity of these systems from the perspective of their technical functions has the potential to present real difficulties. Often users (or even developers) of GenAI systems are unable to explain or understand why a given answer has been reached. This is known as the 'black box' problem and has the potential to present real difficulties.

Employees will likely already have started exploring how GenAI tools can help them in their routine tasks, and the automation of more complex, knowledge-based areas of work is increasingly feasible. It is widely expected that these enormously powerful systems will proliferate to form a key part of our lives both in and out of work. As with all powerful transformative technology, it is difficult to predict at this point how GenAI will change the way we do things and where it will have the most significant impact.

THE LAWS AROUND AI

Specific law governing AI or GenAI is being rapidly developed but is not yet in place. However, in the absence of AI specific legislation, it is important to note that existing laws and regulations also apply.

> "AI SYSTEMS DO NOT TAKE DECISIONS OR HAVE UNDERSTANDING IN THE SAME WAY THAT HUMANS DO"

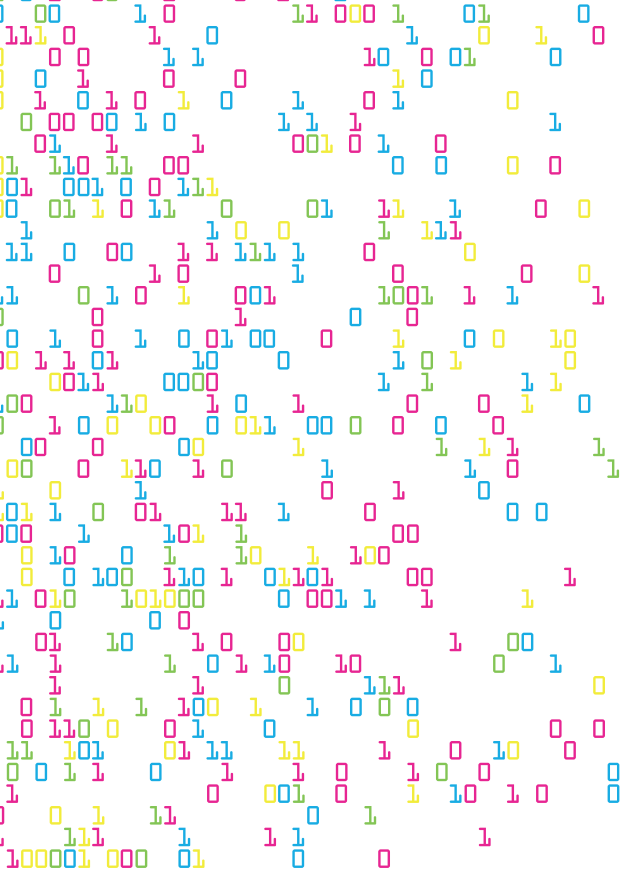
1. EU AI Act

The headline piece of AI-specific legislation around the world at the time of writing is the EU's framework for regulating AI, which focuses on risks to human health and safety and the potential for these systems to infringe on fundamental human rights. To do this, it follows the structure set by the EU's product safety framework, adjusted as needed for the technology. A tiered, risk-based approach is taken, with obligations proportionate to the risks associated with each category.

While some AI applications look likely to be banned across the EU, such as facial recognition systems in public spaces, and behavioural manipulation tools, AI which is considered high-risk will be subject to stringent compliance requirements.

Under the EU's framework, AI is broadly categorised as high-risk where it is performing a function related to health and safety or where its use could impact on fundamental rights. The latter category includes various AI applications in the fields of education, employment, financial services, infrastructure, security and various public services, as well as when AI systems are released as free-standing products, or are integrated into other products as a safety component. High risk AI systems will be subject to detailed data governance requirements, extensive technical documentation and record-keeping obligations, as well as needing to be conformity-assessed against essential requirements, registered and to carry compliance certification.

Some categories of lower risk AI will simply be subject to transparency requirements so that, for example, a consumer knows that they are talking to a chatbot or watching a deep fake video. Other forms of AI may be unregulated.



In addition, new provisions are expected to be added to deal with foundation models that perform a single function but with a wide range of potential applications, such as translation, image recognition, or the creation of images or text. The flexibility of these systems, which have emerged since the AI Act was first proposed, does not translate readily to the categorised risk-based framework. For example, a chatbot generating text could be high risk if producing disinformation, but minimal risk if writing a birthday poem.

The AI Act's 'compliance by design' approach is expected to be finalised in early 2024. It will not come into full legal effect until after a compliance period, still to be agreed but likely to be two years. Anyone seeking to use or supply AI into, or affecting, Member States and EU citizens, will need to comply with the AI Act.

As happened with personal data protection under the EU's General Data Protection Regulation, the EU AI Act is likely to become an international gold standard by setting the most stringent safety requirements. In her 2023 State of the European Union address, President von der Leyen described the AI Act as "already a blueprint for the whole world" and observed that the EU "should bring all of this work together towards minimum global standards for safe and ethical use of AI."

2. The UK's AI approach

The UK's current approach could not be more different to that of the EU. Our AI white paper, published in March 2023, proposed five high-level principles that will be informally issued by the UK government in order to guide the application of existing UK regulation by the existing UK regulators, exercising current powers within their existing jurisdictions. The five principles cover the need for:

- Safety, security and robustness.
- Appropriate transparency and explainability.
- Fairness.
- Accountability and governance.
- Contestability and redress.

For the time being, the UK Government has said that it does not intend to legislate specifically in response to AI, although this may change. However, this is not to say that AI is, or will be, unregulated in the UK.

In terms of how this practically impacts businesses, clearly AI systems and their functions will fall within the scope of many existing UK regulations (some of which are discussed below). A number of regulators, such as the Competition and Markets Authority and the Information Commissioner's Office are already engaged in understanding how AI fits within their areas of expertise and how they are going to approach enforcement.

Many regulatory frameworks are principles-based, which will make it easier to adapt them to new developments. Other regulatory areas (particularly sectoral regulation) can be more prescriptive and specific, which can mean they are less future-proof and may not be sufficiently flexible when technology or business models change. Either way, there is clearly potential for uncertainty and inconsistency while regulators refine their understanding of AI and how to apply their powers.

3. International direction of travel

The USA and China, as well as other major jurisdictions around the world are also considering whether new regulation or guidance is needed to address the risks they see from AI and AI-generated material. The UK and the US have announced a commitment to working together on international action to ensure safety and security in relation to AI, and discussions are also ongoing between the EU and US to find a consistent approach to regulation (although the US approach is closer to the UK's strategy than the very prescriptive approach of the EU).

The UK will be hosting a global summit on AI from a safety perspective in November 2023, with other multinational initiatives also planned including the G7 Hiroshima AI process (with a summit planned for later this year) and a gathering of the OECD-backed Global Partnership on AI. The EU recently suggested setting up a new international body, like the UN's Intergovernmental Panel on Climate Change, to advise on global AI rules. Whether these initiatives will result in substantive law, new international bodies, binding commitments etc remains uncertain, particularly given the differences in approach between different jurisdictions. However, as President von der Leyen said: there is a "narrowing window of opportunity to guide this technology responsibly".

Continued overleaf >>

>> From previous page

INPUT AND OUTPUT

Even without the specific obligations which are likely to be introduced by upcoming regulations, there are existing legal and regulatory risks associated with these systems which businesses should be aware of and should be mitigating against.

GenAI systems present nuanced risks for businesses because of the way that they are trained, and the way that they are made available for use.

Risks can generally be conceptualised as falling into two buckets:

Input risks: being risks flowing from the inputs to the GenAI system, both from a pre-release perspective (such as training data) and during a business' operational use of the GenAI system.

Output risks: being risks arising from the outputs that are generated by the GenAI system – which will depend on the specific system, but might include content such as text summaries, images and code.

The nature and extent of the risks posed to a business will turn on the specific GenAI system, the datasets that it has been trained on, the anticipated use case for that system and the way in which the business plans to operationalise the GenAI system (for instance, the controls that the business will put in place).

At a very high level, GenAI systems usually involve two forms of inputs: their training data and user data.

Most generative AI systems are configured with training data during an initial pre-release stage, ultimately aimed at training the model to generate desired outputs. Typically, such datasets are very large and encompass millions of datapoints (these might be scraped from the web, or from an existing database) which are fed to the GenAI system.

In some cases, GenAI system providers will enable an enterprise customer to further configure the GenAI system with selected materials that can be very detailed and specific to the business or sector concerned, in order to 'teach' the model to respond to prompts with the context of the enterprise customer and their business.

When using a GenAI system, a user will typically input prompts to generate content. For example, in the context of a publicly accessible GenAI system, a user might input a text based prompt such as "prepare an HR policy on parental leave".

Depending on the nature of the GenAI system, a user prompt might involve imagery, text, lines of code, video content, snippets of audio and likely – as the technology evolves – a great deal more categories of user inputs.

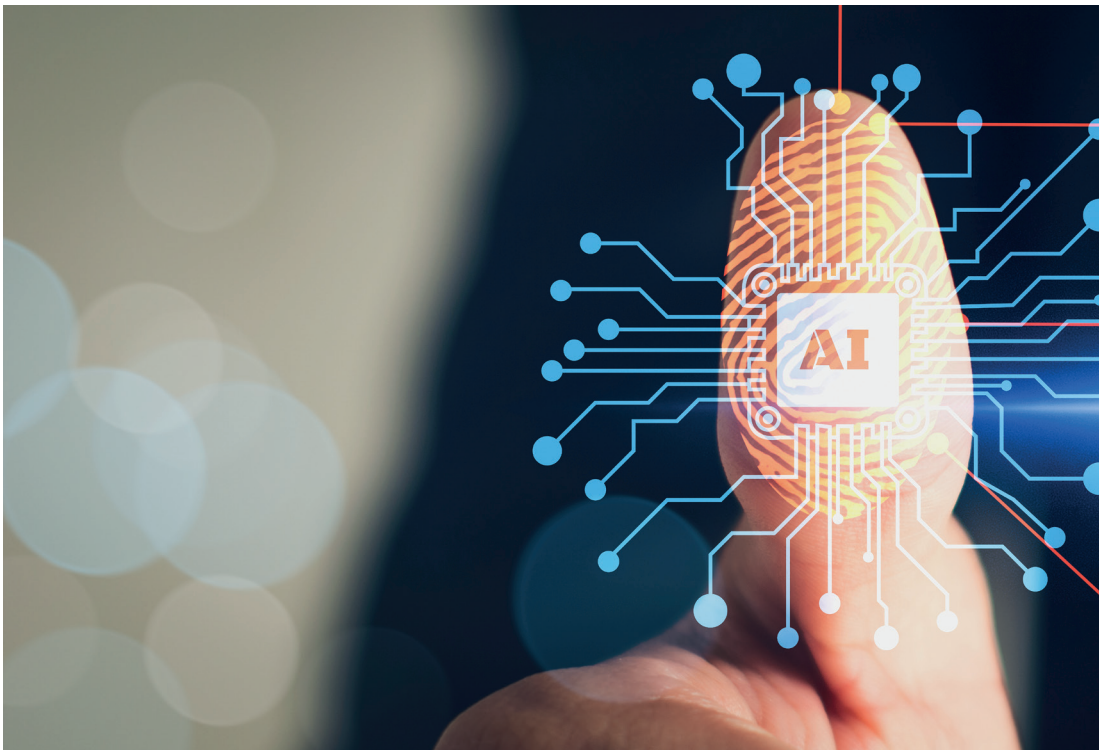
Some GenAI providers will configure GenAI systems to 'learn' from user prompts and other feedback (such as user engagement data and the context of the environment in which the GenAI system is deployed) after the GenAI system has been released for public use.

A GenAI system is only as good as its training data – garbage in, garbage out. A crucial aspect of evaluating the suitability of an AI tool is therefore understanding the quality and composition of its training data. The upcoming EU AI Act will create data governance obligations that emphasise the importance of relevant and representative training data, accounting for specific contexts in which the AI tool will operate.

> "THE UK WILL BE HOSTING A GLOBAL SUMMIT ON AI FROM A SAFETY PERSPECTIVE IN NOVEMBER 2023, WITH OTHER MULTINATIONAL INITIATIVES ALSO PLANNED"

THE LEGAL RISKS

Specific legal risks can flow from training data that has been poorly curated, including bias and discrimination. If the underlying



training data is skewed, for example, against a particular social, racial or cultural profile, then outputs generated may be similarly skewed. Bias may result in both legal risks and reputational risks for your business.

a) Bias and discrimination risk

An AI system's knowledge is confined to the information within its training data. If this data is skewed towards (or against) a particular social, racial or cultural profile, for example, the AI's outputs may reflect these biases to the detriment of an individual or group of individuals with a legally protected characteristic or, more generally, may be offensive or harmful. Such biases and offensive content can lead to illegal discrimination under equalities legislation (e.g. in the UK, the Equality Act 2010) and consumer protection law infringements, posing legal and reputational risks.

It is important to note that organisations deploying algorithms need not intend to discriminate for their actions to be unlawful; indirect discrimination is particularly relevant in the context of AI. For instance, if a data set is used to train and an AI system causes it to show adverts for high-paying jobs more often to men than to women, it can be viewed as a "provision, criterion or practice" that puts women at a disadvantage under the Equality Act. In this case, a woman who might have been interested in applying for one of those high-paying positions but didn't because she never saw the relevant advertisements, may potentially be within the protection of the Equality Act.

While it may be possible to defend indirect discrimination claims on the basis that apparently illegal discrimination is in fact objectively justified, understanding the way that AI systems are making decisions will be essential to support a defence of objective justification; however, one main difficulty with AI is knowledge. The black box nature of AI decision-making means that human insight into the reasoning behind specific decisions is often limited. To avoid risks and potential liabilities, businesses will need to conduct procurement due diligence and monitor the operational functioning of an AI system in order to ensure any discrimination or bias is picked up, and to unravel where responsibility may lie (with those who formulated the algorithm, those who supplied the training data sets, or the user business).

> "ORGANISATIONS DEPLOYING ALGORITHMS NEED NOT INTEND TO DISCRIMINATE FOR THEIR ACTIONS TO BE UNLAWFUL; INDIRECT DISCRIMINATION IS PARTICULARLY RELEVANT IN THE CONTEXT OF AI"

b) Intellectual property (inputs)

In addition, training data may be subject to third party intellectual property rights. Databases used to train generative AI are known to include significant quantities of web-scraped content, gathered by automated systems copying content from across the internet. A significant proportion of this data may be protected by copyright. If such copyright works are used to train a GenAI system without appropriate licences, then there is

Continued overleaf >>

>> From previous page

a real risk of copyright infringement. In the UK, exceptions to copyright for text and data mining currently do not apply to commercial uses of such data. In the EU, exceptions are more generous but copyright holders are able to opt out of them (and often include such a provision in their website terms and conditions to prevent content from being scraped).

This is an area where litigation is on the rise, with both individuals and businesses taking action, albeit typically against the providers of GenAI systems, rather than GenAI system users. It is also an area where policymakers, wanting to promote the growth of AI, are trying to find a path between protecting the interests of copyright holders and making sure that suitable training data is readily available to power AI systems.

And, what should you bear in mind in relation to user data?

Depending on how a GenAI system is used, there is potential for sensitive or confidential data, submitted by the user, to be transmitted by the model to the GenAI provider. In practice, once confidential information is misused by a recipient of that information, there are limited options for redress available to the aggrieved discloser of that information – it is not possible to 'reverse' a disclosure of information. Where users log in and their activity is stored against their account, inputs may be saved cumulatively, creating a further layer of risk if inputs are harmless individually but seen collectively reveal confidential information. Disclosure of confidential information can breach contractual, regulatory or ethical obligations. It could also undermine the protection and value of commercial trade secrets, etc.

In some cases – particularly where a GenAI system is licensed on an enterprise basis – GenAI system providers may offer contractual assurances that a user's inputs are not retained. Also, it can be open to an enterprise user of a GenAI system to configure a model so that user prompts are not used for the purposes of training the GenAI system. Understanding what happens to inputted questions, documents or other content should be explored as part of procurement due diligence. Ultimately, as is the case for any third party system with access to confidential materials, a business should reach its own level of comfort in respect of the security controls offered by the GenAI system provider in the context of the specific use-case.

c) Data protection risks

When inputs and user data have the potential to include personal data, then applicable data protection legislation needs to be complied with.

Unfortunately there are difficulties with this legislation when it comes to 'black box' AI systems. This is because the legislation requires businesses to know the purposes for which the personal data is to be used, and to explain this to the people (including employees and applicants) whose data they are using. It presumes that all IT systems are 'white box' – in essence that when you commission and install them, you know what they will be used for and what outputs they will generate. But GenAI systems, mostly developed since the GDPR was enacted, are able to find non-obvious correlations and can generate unexpected results.

The EU and UK GDPR requirements on automated decision making (ADM) cover, in essence, any automated decision which creates legal or similarly significant effects on an individual (e.g. using psychometric testing to automatically filter out candidates). The ADM provisions require, in most cases, that ADM within their scope should be explicitly authorised by direct consent – legitimate interests as a lawful processing ground is not enough. This provision seeks to ensure that there is meaningful human involvement in ADM processes but may be difficult to meet the context of AI systems.

The ICO has issued extensive guidance around the use of personal data in AI tools. A number of data protection authorities in the EU have also actively scrutinised GDPR compliance of publicly available and widely used GenAI systems. For example, concerns about GDPR compliance led to the temporary suspension of the use of ChatGPT in Italy while the Italian regulator made related inquiries of OpenAI. A number of other data protection authorities have scrutinised ChatGPT and similar GenAI systems such as Google's Bard. We have seen increasing vigilance from data protection authorities in ensuring that AI systems handle personal data in a manner that aligns with stringent data protection laws.

THE OUTPUT RISKS

By 'output' we are referring to the generated material that a GenAI system will produce in response to a prompt.

a) Transparency

Fundamentally, as with other tools or machines, a lack of understanding of how they work does not negate responsibility for their outputs or effects. In the UK, regulators are expected to be able to obtain sufficient information about an AI system in order to perform their functions. Transparency of how the system operates and more specifically why a certain output has been generated is therefore not usually considered to be an absolute requirement, but depends on the context.

There may be contractual requirements affecting business which relate to the transparency or explainability of outputs. The EU AI Act will create overarching transparency requirements for high-risk categories of AI, and this is also required under data protection law, including in relation to automated decision-making involving personal data. The ICO has issued guidance developed with the Alan Turing Institute about explaining AI that processes personal data.

Transparency could also relate to ensuring that stakeholders inside and outside a business know that GenAI systems, or their outputs, are in use. This will be stipulated by the EU AI Act, with the regulatory trend around the world being to ensure that people are aware of when they are engaging or interacting with AI.

b) Intellectual property (outputs)

It remains unclear the extent to which a user of a GenAI system owns IP in generated content, and the risk of infringing a third party's IP because of the manner in which the GenAI system is trained.

Continued overleaf >>

> "THERE IS A RISK THAT OUTPUT MAY INFRINGE A THIRD PARTY'S COPYRIGHT PROTECTED WORK. THIS RISK MAINLY ARISES BECAUSE OF THE WAY THE GENAI SYSTEM WAS TRAINED"

>> From previous page

That said, in some cases, IP ownership will not be of chief concern to your business when it comes to GenAI outputs – for instance, if you only intend to use a GenAI system to create relatively generic text for use by your internal personnel, then owning that material may not be of central importance. In other cases – such as where generated images are incorporated into works for your clients – IP ownership may be viewed as crucial.

Generally, the legal position is not clear cut: in the UK and the EU, where content is created by a human with assistance from a GenAI system, the human may own copyright in the output, but much will turn on the nature of the work and how that work is generated. Where you are using a GenAI system to generate creative content, the risk that IP rights might not be secured is primarily where the AI is being used to replace human authors or inventors, rather than as an assistive precursory tool.

In terms of IP infringement risks, there is a risk that output may infringe a third party's copyright protected work. This risk mainly arises because of the way the GenAI system was trained. There is a possibility that the generated output of a GenAI system may be very similar to copyright works, for example if you ask DALL-E to produce David Bowie, the generated images are often clearly heavily inspired by Ziggy Stardust. The risks here are very context specific, and generally turn on the extent to which there is a human in the loop to have oversight of, review – and amend – outputs prior to incorporation in public facing materials.

c) Hallucinations

Users of GenAI systems might sometimes be presented with outputs which are entirely fictitious or full of factual inaccuracies but nevertheless are the right kind of answer – an answer which is statistically likely to be a good response to an input or prompt. A simple solution to this problem is to ensure that appropriately knowledgeable people provide oversight and sense-check outputs, rather than effectively allowing the system to run unsupervised.

The more important the accuracy of an output is, the more important it is to use the AI system as a tool and not a self-standing solution.

The risks presented by inaccuracies or hallucinations are myriad. Where outputs have an impact on business customers, there may be a need for contractual provisions that create (or exclude) obligations in relation to the accuracy or quality of the services or products being provided. If output inaccuracies could impact on product safety, compliance, or quality, there may be a breach of product regulations, or infringement of consumer protection laws.

THE SOCIAL RISK

A key aspect of the risk presented by AI is its impact on a company's workforce. When AI is introduced and automation becomes part of a business' productivity, it often necessitates reskilling, upskilling, or even restructuring of the workforce. AI adoption might lead to the automation of certain tasks that were previously handled by employees. In such cases, reskilling becomes imperative. Employees may need to acquire new skills to take on different roles within the organisation. However, if suitable alternative positions are unavailable, it might result in redundancies. It's important to remember that workforce restructuring is, of course, subject to legal requirements, including process and consultation. Failing to meet these obligations risks employee disputes.

As AI tools become integrated into business operations, companies must expand or create policies related to the acceptable use of technology and internet resources to incorporate these new elements. These policies should provide clear guidance to staff on the risks associated with the use of AI, including the importance of not inputting confidential or client information into publicly available AI tools. Additionally, such policies should emphasise the importance of verifying the accuracy of AI-generated output before using it. The policies should make clear that non-observance can lead to disciplinary actions being taken against employees.

THE GOVERNANCE RISK

Regardless of incoming regulatory obligations, there is plenty that businesses developing or implementing AI systems can do to prepare themselves to respond to legislative changes and minimise their risks now and in the future. At a high level, businesses need to have effective governance processes and policies set up from the outset, with audit and oversight of their work on AI systems, and work being done with AI as well. When made public, clear and responsible policies can drive trust in a business and its products or services. Internally, it is critical that businesses ensure that their AI tools, their use-cases and their outputs, are aligned with their values and do not expose them to risk.

Particularly in the UK, effective governance systems are likely to be a key risk mitigation measure, as well as offering ethical benefits. In the absence of specific legislation in relation to AI, risks will arise within existing legal frameworks, and businesses will need to draw on their existing understanding of the regulation that applies to their business and develop appropriate risk mitigation for their use of AI.

Effective due diligence both when procuring AI systems and when using them on an ongoing basis will also make businesses more resilient to regulatory change, as there will be clear processes to be followed in order to make the necessary adjustments, and records of decision making.

Under the EU AI Act, categories of high-risk AI will require conformity assessments against essential requirements. These systems will similarly benefit from the improved integrity and trustworthiness which comes from an effective governance system.

OSBORNE CLARKE'S RECOMMENDATIONS

Governance systems

- Undertake due diligence before making GenAI systems available for use, considering the business context, use cases, and regulatory and legal issues.
- Develop an overarching governance framework for your use of AI, taking into account your approach to ethical issues, reputation and risk management. This might include:
  - Developing a set of priorities and benchmarks against which AI tools can be audited and establishing audit processes for AI systems being used; and
  - Implementing ongoing due diligence processes to ensure activities are subject to sufficient governance and oversight, including the impact that your use of AI has on your physical and digital supply chains.
- Develop policies, procedures and training to address the risks of third party IP infringement, and to protect your own IP where this is a significant asset class for your business.

Operational and employee guidance

- Communicate internally to understand what AI systems are being used within the organisation (both AI that has been specifically developed or procured, as well as tools which staff may be accessing independently in the context of their work), and establish clear expectations in relation to the use of GenAI, potentially by amending an existing IT and communications systems policy, introducing a new one or making a statement to employees about it.
- Work with your internal teams to ensure that there is meaningful "human oversight" in respect of materials that incorporate outputs created by GenAI, prior to external publication.
- Consider the cost-benefit of employees using generative AI to perform tasks such as writing routine letters and emails, generating simple reports, and creating presentations, for example, against the potential loss in developmental opportunities for employees performing such tasks themselves.

Contractual assurances

- Ensure AI models are trained on diverse and representative datasets to reduce biases inherited from historical data. Demand transparency from AI vendors regarding their algorithms, data sources and decision-making processes.
- Consider the contractual (and other) assurances that you provide your customers and other third parties, and the impact of your use of AI. Terms of service and marketing materials may need to be updated to ensure transparency about the use ofAI in products and services, and in particular regarding IP ownership in the context of materials created using a GenAI system.
- Adapt data protection compliance provisions in contracts and terms to reflect whether you are processing personal data via AI systems, including securing a lawful basis for using personal data as training data where this is needed.