aiEthix
from Securys®

State of AI Regulation

Current and forthcoming rules, compliance guidance and risk mitigation.

22 MARCH 2024

aiethix.com

Contents

Purpose of this paper

As the use of artificial intelligence, machine learning and other automated decision-making and decision-support systems grows, so likewise do the ethical and data protection concerns. This paper is specifically focused on data protection, considering both the position being adopted by the European Union – which can be deemed to be "leading the charge" in regulating the use of AI – but also the growing impetus towards AI regulation in the US, China and elsewhere.

This paper sets out some general guidelines for the ethical and compliant use of AI and works through a number of practical examples to help elucidate the issue and present in each case a compliant approach and appropriate safeguards. This is not a technical paper – please contact Securys if you would like more detailed advice on specific uses of technology or more detail on a particular aspect of the present and proposed regulations.

As this paper is concerned with data protection, it speaks to cases where personal data – information that pertains to or can be connected to an identifiable living person ("the data subject") – are involved, or where outputs from an AI system have legal or material effects on living people.

This paper does not address specific regulatory and draft regulatory provisions or wider considerations regarding the use of AI in a law enforcement context except where this is specifically relevant for the intended commercial audience.

Illustration created using Adobe Firefly AI image generator.

Executive summary

Scope

The terms AI and machine learning cover a wide range of systems. Existing and future regulation takes a wide view, meaning that it covers not only "cutting edge" tools – such as ChatGPT or StableDiffusion – but also much more conventional decision-support systems including basic fixed decision logic and so-called "expert systems", which generally operate to defined rules rather than being based on learning from large datasets. This in turn brings much of a business's IT estate into scope.

The EU, UK[1], China and the US all have existing regulations that cover the development and deployment of AI. Of these, the Chinese rules are presently the most developed.

The EU has introduced a new regulation, the EU Act, that was approved by the European Parliament on 13th March, 2024, and will come into force in 2026; this law introduces very substantial additional requirements for the use of AI along with considerable additional financial and other penalties. The US has a number of initiatives including a proposed federal "AI bill of rights" along with FTC consultation, a proposed regulatory framework from the National Institute of Standards and Technology and a continually developing body of state law. The UK has recently published a white paper[2] setting out its proposed approach to AI regulation, which is self-described as "pro-innovation" and might reasonably be considered to be lighter-touch than the other existing and forthcoming regulations considered here.

The core purposes of all of these present and future regulations can be summarised as seeking to ensure transparency, fairness, safety and recourse.

Transparency

The regulators all agree that it is essential for individuals to know when they are interacting with an AI or are subject to machine-made decisions. These individuals must understand not only that this is the case, but also the basis for the decision-making and

[1] In the form of the UK Data Protection Act 2018, which implemented the EU's GDPR. Revised legislation is before Parliament at the time of writing.

[2] https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

Executive summary *(contd)*

their rights in respect of it. One developing area is the need for labelling of machine-generated content, something brought into sharp relief by the recent development of both “ChatGPT” and similar text-generation systems and various “deep fake” or “deep synthesis” technologies that produce real-seeming images, video and audio which may include semblances of real people as well as places and objects.

Fairness

This is perhaps the broadest – and most important – aspect of the present and forthcoming regulation. There is both a desire to avoid bias and discrimination in AI systems in terms of their outputs and also a need to ensure that the data used to train heuristic systems is properly sourced and used with the knowledge of the people to whom it relates. The regulations mandate a wide-ranging and demanding data governance and data quality programme to achieve these aims and introduce bans on some potential uses and a distinction between high- and lesser-risk AI.

Safety

Both physical and psychological harm are considered in existing EU, Chinese and proposed UK and US regulation. These considerations not only drive the prohibition of certain uses of AI but also a significant extension of requirements for quality management and security provision for machine learning systems.

Taken together the fairness and safety objectives contribute to the need for a structured and consistent application of a code of ethics governing the use of AI within the enterprise. The Chinese in particular have produced a detailed code of ethics for AI which can form the basis of an internal programme.

Recourse

Existing regulation already provides for appeal to human review of automated decisions. This is reinforced in the EU AI Act legislation with a requirement for ongoing and proactive human oversight within organisational use of AI, and by a programme of registration, conformity assessment and market surveillance for machine learning systems.

Executive summary *(contd)*

Recommendations

This paper makes a number of recommendations in the context of known AI risks alongside the various functions and documentation required by regulation. These could be summarised as:

- Effective and comprehensive governance of all aspects of AI but with particular consideration given to the ethical sourcing of training data and attention to the elimination of bias and appropriateness of datasets and implementations to geographic and social context.
- Widespread and continuous education for commissioners, developers and users of AI systems to ensure awareness of the limitations of AI capabilities and appropriate application of that understanding.
- A central focus on ensuring that application of AI is of benefit to people and society with the use of existing tools such as the data protection impact assessment to check that the interests of the enterprise are properly balanced against those of individuals both within and outside the organisation.
- Implementation of feedback loops to ensure early detection and correction of errors and bias, alongside mechanisms to ensure that human decision-makers can always override AI outputs.
- Policy and procedures making clear that use of AI is a skilled activity and that AI systems – however apparently anodyne or straightforward – should not be introduced without proper review, training and oversight.

Definition of "AI"

"Artificial Intelligence" is a widely used and widely misunderstood term. At its root it refers to humanlike "general" intelligence – the sentient computers of popular media – which is as yet still nothing but science fiction. It is important to note that no present "AI" systems can be considered to be alert to or aware of context in the same way as a person; this significantly limits their capabilities and is one of the primary drivers of risk in the use of AI systems.

At present actual AI in use will either be some form of "machine learning", which uses a variety of techniques to "teach" a system to form outputs from inputs, or a so-called "expert or rule-based system", where rules for making decisions or producing outputs from inputs are algorithmically defined by a human programmer.

While a variety of techniques for teaching machine learning systems exist, this is not especially relevant to this paper. For a detailed explanation of the four most common approaches, see e.g. Yalçın, O.G. *Machine Learning Approaches that Every Data Scientist Should Know*[3]. The important consideration from a data protection perspective is that the actual mechanism by which the trained system reaches its conclusions is often opaque – either very difficult or functionally impossible for a human being to understand or manually replicate. As a consequence, it can both be difficult to determine whether a system is acting fairly and to explain its functioning to those who are affected by the decisions it makes or supports.

One other concern with machine learning is that because the machine has itself built the connections used to derive outputs from inputs, and because as noted above these systems are not sentient and not alert to context, it is unlikely to be drawing its conclusions in the way that a human would. This can lead to non-deterministic results, where what to a human appear to be identical input data produce disparate results. For the reasons noted above it can be very difficult to determine what is causing the variance in output, which again can lead to challenges with ensuring fairness.

Expert systems are both deterministic and explicable, since their logic is explicitly defined by their programmers. Therefore, the issue of explicability is less challenging; that of fairness, however, remains but directly reflects biases in the designers of the system.

In 2023 we saw a step-change in the capability of the machine-learning approach with the advent of "generative pre-trained transformer" language and image models being

[3] https://towardsdatascience.com/4-machine-learning-approaches-that-every-data-scientist-should-know-e3a9350e-c0b9?gi=83842174710b

Definition of "AI" *(contd)*

released for public use. Transformer technology[4] was developed by Google in early 2017; it is in essence a change to the way that neural nets are able to identify and attend to features in training sets of, for example, words or pictures. While powerful it has also proved unreliable when coupled to a "generative" output function designed, not just to find patterns but to create them, and its use was mostly confined to the research departments of large tech companies while they figured out how to use it.

The release of ChatGPT by OpenAI changed all that. By training a model with over 1.75 trillion) parameters on a vast dataset – numbers that were larger than anything previously attempted – and then using human-led reinforcement learning with human feedback (RLHF) techniques to fine-tune its behaviour, OpenAI managed to produce what's now termed a "foundational" model that was uncannily good at answering natural language queries across a wide range of domains with high accuracy and low hallucination compared to previous attempts.

Other similar models have followed, some of them proprietary, many of them open source, along with similar advances in generative models for images and video as other, more specialised datasets (protein structures, robotic movement, self-driving etc).

This explosion of capabilities has made business of all stripes, many of whom had still only dimly spied a vague outline of AI somewhere on their distant horizon, sit up and take notice. This technology is suddenly here and, while not perfect, it is clearly capable – if perhaps not yet of replacing quite as many jobs as media headlines would have us believe – then certainly of changing and augmenting them. Understanding how to use the technology and benefit from it has become an imperative, and this means that having the capability to audit it, risk assess it and safely deploy it is no longer a nice to have but a key plank of corporate strategy.

[4] Attention is All You Need (research.google)

The Securys perspective

Deep learning is basically a way of finding patterns in big data, both pre-existing and potential. In a sense it's a form of search, but one that allows you to search extrapolated possibilities in a probability landscape constrained by the training data set, rather than just letting you search what's already there (like a web search does).

It can feel uncannily human, because a lot of what we humans do in culture and conversation is not dissimilar. There are key differences however: humans balance this kind of "fast" system 1 thinking (ref) with a slower system 2 set of higher order concepts that help us contextualise the search outputs and see them within wider frameworks of physical, social and cultural implication – what we call "good judgement" or, more colloquially, "common sense". At least at present, AI is not able to do this second piece. It has no common sense because it has no worldview of its own and no sense of itself as an actor in a wider environment. Attempts to build Artificial General Intelligence ("AGI") are attempts to give it that kind of sensibilitiy, something that could ultimately lead to self-consciousness. But at present, all the direction and awareness must come from us, its operators.

This means that AI remains a tool that we are deploying. This doesn't make it neutral – AIs can be dangerous, they can give a false or biased picture of data to their operators, they can give unwarranted access to dangerous knowledge, they can have big environmental impacts in terms of the resources they use, they can distort markets, they can be used for surveillance and control, they can create misinformation, they can enhance weaponry. All these are non-neutral downsides that are latent in the tech. But they are tools, not consciousnesses, and how they perform depends to a great deal on how they are built and deployed, by their users.

AI safety is therefore greatly enhanced by high quality AI audit and the application of ethics to design principles. At Securys we create and promote such principles and believe that AI built in accordance with and governed by them is very likely to prove a valuable technology that will improve human experience and capabilities for the better.

Current regulatory landscape

GDPR

The existing GDPR already governs AI in a number of ways:

Personal data must be processed fairly – this is a core principle of the GDPR and one that is reflected in the EU AI Act. Perhaps the primary concern in regulating AI is to ensure fairness – by avoiding bias, by using complete and accurate information and by offering data subjects appropriate rights of appeal and redress.

Personal data must be processed for specific and limited purposes. Many machine learning systems are trained using data originally collected for other purposes and/or once trained are deployed to process data that has been collected for other purposes. Unless the use of data for training or the subsequent processing by machine learning systems was disclosed to the data subject **at the time the data were originally collected** this further use must be disclosed; it must also be separately justified and may be subject to challenge by the data subject. In some cases, this "extension of purpose" can be unlawful.

Processing of personal data must be properly explained to the data subject ("the principle of transparency"). As noted earlier – and expanded upon below – this can be very difficult in practice when using machine learning systems. It is important to note that the explanation must be demonstrably comprehensible to the data subject, which presents additional challenges when processing data on children or on data subjects with limited education or literacy skills.

Data subjects have the right to request correction or erasure of their data, subject to a variety of constraints. Where data have been used as part of the training dataset for a machine learning model, or as ongoing inputs in reinforcement, those data may be incorporated permanently into the model itself; it can be functionally impossible to change or erase the data. This problem is not subject to simple resolution.

Personal data should only be retained for so long as is necessary; it can become challenging to demonstrate the necessity of retaining a particular data subject's information in a model if that data subject themself is no longer involved in any way in the processing being performed. Again, this can present a challenging problem given the difficulty in removing specific data from a trained model.

The GDPR also provides, in Article 22, **specific rules pertaining to "automated decision-making"**. This refers to any instance where a legal or material decision that affects a data subject is made without human involvement. It provides for an automatic right

Current regulatory landscape *(contd)*

to require human review of the decision and requires "suitable measures to safeguard the data subject's rights and freedoms". The Article also limits the use of automated decision making to specific use cases: the performance of a contract, compliance with a legal obligation or explicit consent; the use of automated decision-making that involves sensitive ("special category") data is further restricted to allow only consent or public interest processing.

> *The GDPR significantly restricts the use of automated decision-making in a commercial and employment context: you must engage with the Data Protection Team at the earliest possible opportunity if you are developing or a deploying a system that makes fully automated decisions about people.*

United States

State laws

A number of states including New York, Illinois and Maryland have introduced legislation regulating the use of **AI in employment decision-making**. This includes a requirement (in New York) for **annual auditing of so-called Automated Employment Decision Tools (AEDT)** for bias; the audit results must be published. Note that where AEDTs are regulated this is applied to decision-support systems in addition to actual automated decision making.

California, Colorado, Connecticut, Virginia, Utah, Delaware, Iowa, Indiana, Montana, Oregon, Tennessee and Texas have all passed general state privacy laws (the last seven states passed these laws in 2023; they are due to come into force 2024/2025). In general, these laws all grant residents some of the rights familiar from GDPR; specifically, they include provisions for **enhanced transparency regarding the use of AI**. The transparency provisions vary; they contain elements of existing GDPR provisions and the new proposals in the EU AI regulation. The consensus is that:

- Data subjects must be informed when AI is being used to process their data
- The algorithm and the work done to eliminate bias and ensure fairness must be explained
- Where profiling forms part of the processing, it must be justified

Current regulatory landscape *(contd)*

- Where material decisions affecting the data subject are made by AI, this must be separately identified. Some states also offer an opt-out from automated decision-making similar to the GDPR Article 22 requirement for human intervention to be made available.

State laws also introduce a requirement for a data privacy impact assessment; the Colorado law goes further in requiring a specific **AI impact assessment** which includes provisions for documentation of:

- Sources and uses of training data
- Logic and statistical methods used to create the AI model
- Evaluations of accuracy and reliability
- Evaluations of fairness and "disparate impact"
- Evidently compliance with the EU regulation would act as a superset of the present US state requirements.

Federal enforcement, guidance and legislative direction

The Equal Opportunity Employment Commission has issued **guidance on ensuring that AEDTs do not violate the Americans with Disabilities Act** by screening out disabled applicants. As with much present US federal enforcement the focus is on interpretation of existing law and associated publication of guidelines[5].

More widely but along the same lines, the Federal Trade Commission (FTC) has published guidance[6] on using **AI in compliance** with a variety of existing legislation for which the FTC has enforcement powers; importantly this includes the FTC Act which is the general basis for privacy enforcement in the US. The key provisions in the guidance documents are that companies should:

- Make sure AI is trained using data sets that are representative, and do not "miss[...] information from particular populations."
- Test AI before deployment – and periodically thereafter – to confirm it works as intended and does not create discriminatory or biased outcomes.

[5] https://www.jacksonlewis.com/sites/default/files/docs/EEOC-TechnicalAssistanceADA-AI.pdf

[6] https://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai

Current regulatory landscape *(contd)*

- Ensure AI outcomes are explainable, in case AI decisions need to be explained to consumers or regulators.
- Create accountability and governance mechanisms to document fair and responsible development, deployment, and use of AI.

The FTC has been enforcing against misuse of AI; generally, this has been based either on absence of evidence of proper consent or misuse of data for model training even when data subjects have refused consent.

China

PIPL

The Personal Information Protection Law[7] grants data subjects similar rights to those found in the GDPR, albeit only with regard to the private sector. These include provisions similar to the GDPR's Article 22 governing automated decision-making, notably:

- An obligation to ensure that automated systems operate fairly and transparently, without bias and without unreasonable differential treatment.
- Rights for data subjects significantly affected by AI systems both to an explanation of algorithms and to require human review of decisions.
- Completion of a data protection impact assessment for the use of personal data in automated decision-making.

New Generation AI Ethics

The Ministry of Science and Technology published a *New Generation Artificial Intelligence Code of Ethics*[8] in 2021. The intent of the code is to ensure that AI stays under

[7] https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/

[8] http://www.most.gov.cn/kjbgz/202109/t20210926_177063.html

Current regulatory landscape *(contd)*

“meaningful human control” and it introduces 6 fundamental and 18 operational and management ethical practices (“norms”) that must be observed by developers and users of AI systems. These are summarised below, based on an English translation from the International Research Center for AI Ethics and Governance[9].

Fundamental ethical norms

1. **Enhancing the wellbeing of humankind.** This covers human rights, common values, national and regional ethical norms; also sustainable development, ecology and the public interest.
2. **Promoting fairness and justice.** Protect the legitimate rights and interests of all relevant stakeholders, promote fair sharing of the benefits of AI in the whole society, and promote social fairness and justice, and equal opportunities.
3. **Protecting privacy and security.** Respect the rights of personal information.
4. **Ensuring controllability and trustworthiness.** Ensure that AI is always under meaningful human control and that humans have the rights to choose whether to accept AI services and decision-making.
5. **Strengthening accountability.** Clarify the responsibilities of all relevant stakeholders. Establish an accountability mechanism in AI related activities.
6. **Improving ethical literacy.** Popularize knowledge related to AI ethics, objectively understand ethical issues, and do not underestimate or exaggerate ethical risks.

Norms of management

1. **Promotion of agile governance.** Respect the law of development of AI, fully understand the potential and limitations of AI, continue to optimize the governance mechanisms and methods of AI.
2. **Active practice.** Comply with AI related laws, regulations, policies and standards, actively integrate AI ethics into the entire management process.
3. **Exercise and use power correctly.** Clarify the responsibilities and power boundaries of AI-related management activities and standardize the conditions and procedures of power operations.
4. **Strengthen risk preventions.** Enhance risk awareness, carry out systematic risk monitoring and evaluations in development and operation.

[9] https://ai-ethics-and-governance.institute/2021/09/27/the-ethical-norms-for-the-new-generation-artificial-intelligence-china/

Current regulatory landscape *(contd)*

5. **Promote inclusivity and openness**. Pay full attention to the rights and demands of all stakeholders related to AI, encourage the application of diverse AI technologies to solve practical problems.

Norms of research and development

6. **Strengthen the awareness of self-discipline.** Actively integrate AI ethics into every phase of technology research and development, do not engage in AI research and development that violates ethics and morality.
7. **Improve data quality.** Strictly abide by data-related laws, standards and norms throughout the data lifecycle.
8. **Improve the completeness, timeliness, consistency, normativeness and accuracy of data.**
9. **Enhance safety, security and transparency.** Improve transparency, control, resilience and defensibility of AI systems and implement verifiable and reliable auditing and supervision.
10. **Avoid bias and discrimination.** Fully consider the diversity of demands, avoid potential data and algorithmic bias, and strive to achieve inclusivity, fairness and non-discrimination.

Norms of supply

11. Respect market rules. Abide by the regulations for market access, competition, and trading activities, actively maintain market order, and create a market environment conducive to the development of AI.
12. **Strengthen quality control.** Strengthen the quality monitoring and the evaluations on the use of AI products and services, avoid infringements on personal safety, property safety, user privacy, etc. caused by product defects.
13. **Protect the rights and interests of users**. Users should be clearly informed that AI technology is used in products and services. Users must be able to choose to use or quit the AI mode.
14. **Strengthen emergency protection.** Emergency mechanisms and loss compensation plans and measures should be investigated and formulated. Implement timely monitoring, address user feedback and make corrective actions swiftly.

Norms of use

15. **Promote good use.** Fully consider the legitimate rights and interests of stakeholders, so as to better promote economic prosperity, social progress and sustainable development.

Current regulatory landscape *(contd)*

16. **Avoid misuse and abuse.** Respect the rights of relevant entities not to use AI products or services. Avoid improper use, misuse and abuse of AI products and services.
17. **Forbid malicious use.** AI products and services that do not comply with laws, regulations, ethical norms, and standards may not be used, nor may AI be used for illegal activities or to cause a risk to public safety or the public interest.
18. **Timely and proactive feedback.** Participate in the practice of AI ethics and governance, providing prompt feedback to relevant subjects and assistance for solving problems.
19. **Improve useability.** Improve knowledge and usage skills related to AI so as to ensure the safe and efficient use of the technology.

Deep Synthesis

China's Provisions on the Administration of Deep Synthesis of Internet-based Information Services[10] - their "deep fake" law – is already in force and has detailed provisions for every stage of the digital synthesis lifecycle. Notably their provisions also apply to the creation of text as well as image and sound media. The regulations apply both to service providers and to users.

In addition to strengthening the requirement for compliance with the PIPL, the deep synthesis law expects:

- Improved transparency including a requirement for real-identity information authentication system
- Content management and labelling so that synthetic content can readily be identified
- Technical security – essentially the documentation, data governance, bias-elimination and quality management elements of the EU regulation.

This has particular impact on current trends for the deployment of chatbots in customer service, often with no immediate indication that the customer is corresponding with a machine or with seamless transition from machine to human interaction under the same apparent agent name, and for the use of generated content in marketing, such as the recently announced Levi's campaign using digital models to enhance diversity[11].

[10] http://www.cac.gov.cn/2022-12/11/c_1672221949318230.htm

[11] https://www.businessoffashion.com/articles/technology/levis-will-begin-testing-ai-generated-models/

Current regulatory landscape *(contd)*

Other jurisdictions

Other countries around the world are looking to follow the lead set by the EU in particular, as was the case in data protection with GDPR. Singapore has published a proposed governance framework for the use of generative AI[12], and Peru[13] has become the first country in Latin American to enact AI-specific legislation (Law No. 31814).

[12] nais2023.pdf (go.gov.sg)

[13] Law No. 31814

Future regulatory landscape

EU AI regulation

On Friday, December 8, 2023 – after months of intensive trilogue negotiations – the European Parliament and Council reached political agreement on the European Union's Artificial Intelligence Act ("EU AI Act"). This builds on the existing general data protection regulation (GDPR) in several ways. It was approved by the European Parliament on March 13, 2024 and will initially come into force in 2026 with the various obligations and implementations going live in a series of tranches, released at six monthly intervals, until 2028.

In what follows the relevant Article number from the regulation is in square brackets as [Ax].

Definition of AI [A3]

The EU definition of AI is extremely wide, as noted above:

a. Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

b. Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

c. Statistical approaches, Bayesian estimation, search and optimization methods.

The regulation starts by establishing a hierarchy of risk in AI systems. Some systems are defined as carrying unacceptable risk. These are prohibited [A5]:

a. Use of subliminal techniques to distort a person's behaviour in a manner that may cause them or another person psychological or physical harm

b. Exploitation of any vulnerabilities in a group of people arising from physical or mental disability or age in a manner that may cause them or another person psychological or physical harm

c. Systems operated by or on behalf of public authorities that evaluate or classify the trustworthiness of natural persons leading to detrimental or unfavourable treatment [akin to the social credit system currently in operation in China]

d. The use of real-time biometric identification systems in public spaces for the purposes of law enforcement (with some specific exceptions)

Future regulatory landscape *(contd)*

Exclusions [A2]

Products and systems in the following areas are directly not covered by the regulation; this list will be reviewed every four years. It should be noted however that the AI Act is still in the legislative process and the final text may have changes from the version extant at the time of writing. Also, although the Act does not directly apply to automated and autonomous vehicles (AVs) and their AI components, several of the Act's accountability requirements will be applied to AVs in the future through delegated acts of the Commission under the type approval frameworks for aviation and motor vehicles.

- Civil aviation and civil aviation safety
- European rail system interoperation
- Motor vehicles (public road use)
- Motorcycles, tricycles and quadricycles
- Agricultural & forestry vehicles
- Marine equipment

High-risk systems [A6]

The regulation then goes on to define "high-risk" systems. Much of the additional regulatory prescription refers only to these high-risk systems.

It is important at this point to note that the bulk of the regulation is directed at "providers" of AI systems. This is intended to capture the manufacturers and marketers of such systems with the expectation that users of the systems can then rely on the conformance and compliance of the provider. However, a user can become a provider, and hence subject to the full force of the regulation, if it either modifies a system – as, for instance, by doing its own model training, or develops an AI system for internal use.

High-risk systems are defined as follows, noting that the regulation provides for this list to be updated at any time:

a. Any system where an AI system acts as a safety component of a product that is subject to EU harmonisation legislation and associated third-party conformity assessment.

b. Biometric identification and categorisation of natural persons

Future regulatory landscape *(contd)*

c. Management and operation of critical infrastructure (safety components)

d. Education and vocational training (admission and assessment)

e. Employment, workers management and access to self-employment (recruitment, performance assessment, promotion and termination, task allocation)

f. Access to and enjoyment of essential private services and public services and benefits (eligibility, credit scoring, emergency services prioritisation)

g. Law enforcement (risk assessment – offending/re-offending, polygraph equivalents, deep fake detection, assessment of reliability of evidence, crime prediction based on profiling, traits or past behaviour of individuals or groups, crime analytics)

h. Migration, asylum and border control management (polygraph equivalents, risk assessment, document verification, asylum/visa/residence application assessment)

i. Administration of justice and democratic processes (research and interpretation of facts, application of law)

Obligations on providers of high-risk systems

High risk systems are subject to the following – extensive – requirements, all of which must be document and capable of demonstration to a supervisory authority. The provider (which may, as noted, be the user) is required to self-certify that all of these requirements are met. Where the provider is not the user, there remains a responsibility as part of supplier selection and assessment to ensure that this self-certification has been performed and that the user is satisfied with the underlying compliance of the provider.

a. A **risk management system** must be implemented; this is a wide-ranging requirement that specifies a broad and precautionary view of risk and includes both testing and continuous assessment. [A9]

b. All high-risk systems must be supported by an evidenced **data governance framework**. Those that involve model training must also meet specific criteria for the development of the training, validation and test data sets. These requirements are intended to minimise the risk of inherent bias, to ensure the relevance of the data to context – including e.g. geographic, behavioural or functional – and to assure compliance with existing constraints on data collection and processing. [A10]

c. Systems must have accurate and complete **technical documentation** that is prepared before the system is put into service; the documentation must be maintained. [A11],[A18]

Future regulatory landscape *(contd)*

d. Detailed logs are required as part of a commitment to accurate **record keeping**; these logs must not only record use and data sources but also identification of natural persons whose data has been processed and the information necessary to evidence mitigation of bias and fairness risks. [A12]

e. In support of the wider **transparency** obligation discussed above, the system must be accompanied by detailed user documentation which allows users to understand the capabilities and limitations of the system, any circumstances which may lead to risks to health and safety or fundamental rights and specifications of datasets, validation and testing. This documentation must be updated and include change logs. There are also requirements for appropriate contact information and for details of the human oversight required by the next point. [A13]

f. High-risk systems must be subject to **human oversight**. The human overseers are expected to be able to fully understand the capacities and limitations of the system and monitor its operation for anomalies, dysfunctions and unexpected performance; to remain aware of the possible tendency of automatically relying or over-relying on the output (discussed in more detail below); to be able correctly to interpret the systems output; to be able in any particular situation to decide not to use the system or to override, disregard or reverse its output; to be able to intervene in the operation of the system or interrupt it through a “stop button” or similar procedure. [A14]

g. There is an enhanced requirement for both **accuracy** and **cybersecurity**; this is aimed at ensuring both that the data and underlying decision-making model are resilient to error and failure and that the system is protected from outside interference or manipulation; this includes a specific expectation of input sanitisation to avoid exploitation of model behaviour or model flaws. [A15]

There are additional organisational requirements placed on the providers of high-risk AI systems (noting again that user may well be considered the provider). In addition to compliance with the list above, these are:

Operation of a **quality management system**. Although ISO 9001 is not specifically cited, it’s clear that this would serve as a demonstration of compliance, and the listed requirements in the draft regulation in the relevant Article 17 conform closely to those in ISO 9001. It is worth noting that certification to ISO 27001 and 27701 would also provide good evidence of compliance with many of the other requirements. [A17]

Submission of a **conformity assessment**. The specific modality for this submission varies depending on the nature, use and context of the AI system but will, for most applications, likely be a self-certification procedure – either by the provider or by the user if it is considered the provider. It is important to note that operating a high-risk AI system that does not have an EU declaration of conformity will be unlawful once the regulation is in force. [A19]

Future regulatory landscape *(contd)*

High-risk AI systems must be **registered** with the EU – a database for this purpose is defined in the regulation – before being put into service. [A51]

The logs and other records required under point (d) above must be **retained**; the retention period will vary – it may be legally defined, or it may be left to the judgement of the provider as "appropriate in the light of the intended purpose" – but note also the provision for retention of documentation covered later. [A20]

Providers must inform distributors – who in turn must inform users – of any **corrective actions** that may be required on detection of a non-conformity in a system. The regulation requires that this is done immediately. Where the user is considered to be the provider, evidently all of the responsibility will fall on the internal organisation. [A21]

If **a system is considered to present a risk** as defined in the Regulation on Market Surveillance and Compliance of Products[14] the provider is required to **notify the relevant competent authorities** and any certifying body, providing details of the risk and associated corrective actions. The text suggests that this will arise when a system that has been in operation (hence the reference to an existing certification) develops a fault or is discovered to have a bias or other unplanned behaviour that is considered to constitute a risk – which means the **quality control process needs to include a continuous risk assessment programme** in order to identify such notifiable circumstances. [A22]

Providers must co-operate with national competent authorities – noting here and above that the AI Act does not envisage a one-stop shop approach unlike the GDPR – when required, **to provide information and documentation in demonstration of compliance**. This includes providing access to the logs mentioned earlier. [A23]

There is also a provision **extending all of the duties** placed on providers of high-risk systems **to manufacturers** whose products integrate a supplied high-risk AI system and are included in a list of products subject to harmonised EU safety legislation. In summary the list covers: machinery; toys; recreational watercraft; lifts; PPE for use in explosive atmospheres; radio equipment; pressure equipment; cableways; PPE in general; appliances powered by gaseous fuels; medical devices; in-vitro diagnostic devices. [A24]

[14] EU 2019/1020 Article 3.19: 'product presenting a risk' means a product having the potential to affect adversely health and safety of persons in general, health and safety in the workplace, protection of consumers, the environment, public security and other public interests, protected by the applicable Union harmonisation legislation, to a degree which goes beyond that considered reasonable and acceptable in relation to its intended purpose or under the normal or reasonably foreseeable conditions of use of the product concerned, including the duration of use and, where applicable, its putting into service, installation and maintenance requirements;

Future regulatory landscape *(contd)*

Non-EU providers must have EU representation; this requirement is similar in character to the requirement for representation under GDPR Article 27. [A25]

Importers and **Distributors** of high-risk systems are **responsible for ensuring conformity**. This will be of critical importance for any business established in the EU which selects and implements a high-risk AI system using a product that is not directly offered in the EU and therefore has not been subject to the conformity process. This will have the effect of transferring all of the responsibility for compliance to the business even if the normal threshold for being considered the provider (see next paragraph) is not met. [A26], [A27]

The responsibilities of the provider transfer to the user (or importer, or distributor) if they:

a. **Market** or implement a high-risk system under their name or trademark

b. **Modify the intended purpose** of a high-risk system already placed on the market or implemented

c. **Make a substantial modification** to a high-risk system

It is not clear how the "intended purpose" of a high-risk system will be determined; in general, the regulation is not clear on the modalities governing general purpose artificial intelligence systems. Even ChatGPT can be employed in "high-risk" ways but is unlikely to be regarded as directly subject to the high-risk provisions in its standard form. [A28]

The provider must **retain all documentation required by the regulation for 10 years** following a system being placed on the market or put into service and make it available as needed to competent authorities.

Providers are required to establish a **monitoring regime** to collect performance data from users and **continuously assess continued compliance**. The documentation for the regime must be available to regulators. [A61]

Providers are also required to **report any serious incidents or malfunctions** to regulators. Users are similarly required to report incidents and malfunctions to providers (see later). There is a 15 day maximum reporting timeframe. [A62]

Future regulatory landscape *(contd)*

Obligations on users of high-risk systems [A29]

Users of high-risk systems are also subject to certain provisions. In summary these are:

- To use the system only in accordance with its **instructions of use**
- To **monitor input data** for relevance to the intended purpose
- To **monitor operations for risks** as identified in the Market Surveillance regulation (see above). If such risks are identified, to notify the provider and suspend use of the system.
- To **monitor operations for malfunctions and incidents**, reporting these to the provider and, if the provider cannot be reached, directly to the competent authority.
- To **maintain automatically generated logs** and retain them for an appropriate period; additional constraints apply to credit institutions regarding the maintenance of logs.
- To carry out a **Data Protection Impact Assessment**.

The transparency obligation [A52]

All users of AI systems that interact with natural persons (whether high-risk or not) have additional transparency requirements beyond what is already covered in Articles 12-14 of the GDPR. Notably:

It must be made obvious that the data subject is interacting with an AI system; this is evidently relevant for example when using chatbots as part of employee or customer support.

Where emotion recognition or biometric categorisation is in use it must be separately notified to the user. As well as potential implications in wellness and safety systems, this would also affect biometric elements of authorisation and access control, such as keystroke pattern recognition or facial recognition.

AI-generated video, image and audio content that resembles living persons, objects or places must be labelled as such. The regulation specifically calls out “deep fakes”, but this applies to any AI-generated media, so will evidently include training material, architectural simulations and so forth.

Future regulatory landscape *(contd)*

Sandboxes [A53-54]

There is provision for regulators to establish "sandboxes" where new AI systems can be developed without being subject to the full force of the regulation – although the powers of the regulator are not reduced by the sandbox. This includes additional exemptions for systems developed in the public interest.

Small-scale providers [A55]

Small-scale providers are to be offered additional support including training and reduced certification fees, noting that "small-scale" is not defined in the Act.

Codes of conduct [A69]

The European Commission is tasked with encouraging providers and provider organisations to draw up codes of conduct that will facilitate voluntary application to non-high-risk systems of some or all of the provisions for high-risk systems in the regulation. The codes of conduct are also intended to include provisions for sustainability, accessibility, stakeholder participation in design and development team diversity, to be applied (voluntarily) to all AI systems.

Authorities and Conformity [A30-A49], [A56-60]

This paper does not address the provisions in the draft regulation establishing the European Artificial Intelligence Board, the competent authorities and the certification bodies; nor does it deal with the process for declaring conformity and obtaining certification. Readers are advised to refer directly to the text of the draft AI Act if they require further information in this regard..

Enforcement and penalties [A63-68], [A70]

The regulation provides for ongoing market surveillance; the responsibility for executing this is devolved to different regulators according to the purpose of the systems and/or the sector in which the provider or user operates. Data protection regulators and other bodies tasked with protecting fundamental rights will also have access to the

Future regulatory landscape *(contd)*

documentation required under the regulation. If those bodies are concerned that a system may not be compliant, they can request that the market surveillance authority conduct testing of the system in question.

Market surveillance authorities have powers to suspend the use of a system and to require corrective action if they consider the system to be non-compliant or to present a risk to rights and freedoms. There is also a mechanism for co-ordination of suspension and correction activities across the Union. Regulators also have the power at national and EU level to prohibit the use or sale of systems where corrective actions have not been performed or are not considered sufficient.

Administrative fines can be levied at various levels for compliance failure:

- Use of systems for prohibited purposes or failure to operate proper data governance of high-risk systems carries a fine of €30m or 6% of global turnover, whichever is the higher.
- Other penalties are €20m/4% for non-compliance with other aspects of the regulation and €10m/2% for providing incorrect, incomplete or misleading information.

The regulation does not specify that these fines are mutually exclusive. They are separate to any penalties that might be applied under GDPR for, e.g., breaches of Article 9 (controlling the use of "Special Category" or sensitive data) or Article 22 (controlling automated decision making with regard to natural persons).

Proposed UK legislation

The UK's approach as set out in the "A pro-innovation approach to AI regulation" white paper[15] is intended to be iterative; it is therefore less prescriptive and less detailed than the EU regulation. The paper sets out five key principles for AI:

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

[15] A pro-innovation approach to AI regulation - GOV.UK (www.gov.uk)

Future regulatory landscape *(contd)*

While these are similar to those present in existing and proposed regulation elsewhere, the UK has chosen initially not to place them on a statutory footing. Notably the white paper introduces no additional regulatory powers or fines, relying on existing regulators to use the principles for guidance within their present regulatory framework. The paper does anticipate adopting the principles in statue at some future date.

The UK definition of AI in the white paper is considerably narrower than that used in the EU Act – it considers only heuristic ("adaptive" in the text) models that have some autonomous decision-making capability. This would exclude, therefore, the rules-based and statistical systems that are covered by the EU AI Act.

The white paper follows the EU in considering AI in context; in place of the EU's risk-tiering, the UK looks at use cases and follows a sectoral approach that is consistent with AI regulation being an extended function of existing regulators. However, as a result implementation of the proposed approach will be heavily dependent on cross-regulatory co-operation; this makes it difficult at this point to predict the specific impact of the approach to sectors, such as healthcare or financial services, where a number of sector-specific regulators may be involved as well as the data protection regulator, the Information Commissioner's Office (ICO).

Government will provide some new central regulatory functions:

- A market surveillance model similar to that proposed by the EU but without the detail on how information is to be gathered; the paper does not, for instance, place the same reporting duties on industry as are present in the EU AI Act, but states instead that it will "gather evidence and feedback from a range of sources and actors in the ecosystem".
- The paper does identify the need for cross-regulatory guidance in order to manage conflicting regulatory priorities; however, without a statutory basis for the principles it is not clear how this will be managed at ministerial level when different regulators are accountable to different departments.
- A cross-sectoral AI risk register is to be compiled. Again it is not clear how the information to inform this register will be obtained or how risk will be assessed consistently across sectors.
- There will as in the EU Act be provision for sandboxes and testbeds; these will be delivered in a multi-regulator format. The paper is not clear about how access to sandboxes will be determined; there will be an initial pilot within a single sector that has multiple regulators, that sector to be identified. The Digital Regulatory Co-operation Forum is also working in this area[16].

[16] https://www.gov.uk/government/publications/projects-selected-for-the-regulators-pioneer-fund/projects-selected-for-the-regulators-pioneer-fund-2022#project-led-by-the-information-commissioners-office

Future regulatory landscape *(contd)*

- The paper envisages centralised promotion of education and awareness. This appears from the paper to be seen as a function of central government, presumably under the aegis of the Department for Science, Innovation and Technology, rather than an extension of the ICO's existing education and awareness responsibilities.
- There will be a horizon-scanning function delivered both by individual regulators and by central government through convened committees and conferences. This will be central to the effectiveness of the iterative approach.
- The paper recognises the need for interoperability with international frameworks; however it is not clear how this will work in practice. Notably the direction of travel outside the UK is towards statutory and detailed AI regulation, usually by extending the powers of the existing data protection regulator. **A key commercial consideration is whether the UK's regulatory regime will be superseded even for UK-based international enterprises by the need to comply with the EU, Chinese or US laws**.

The paper explicitly states that the UK does not at present intend to address accountability within the AI supply chain – in direct contrast to the EU's provider and user distinction and explicit registration, conformity and reporting requirements. The responsibility is instead devolved onto industry, with technical standards used to inform how supply chain risk may be managed.

Specific mention is made of "foundation models" – generalised AI based on very large volumes of training data such as large language models – and the specific need for additional surveillance and consideration of regulatory interaction. The UK has a Foundation Model Taskforce that was established in the 2023 Integrated Review Refresh[17]; it is responsible for this monitoring.

AI Safety Summits

In November 2023 the UK government hosted a landmark AI Safety Summit at Bletchley Park. The event brought together international officials, leading AI companies, and experts to discuss the risks and mitigation strategies for advanced AI development. The participants agreed "the Bletchley Declaration", an agreement to pursue an international management framework for cutting-edge artificial intelligence ("frontier AI") that

[17] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1145586/11857435_NS_IR_Refresh_2023_Supply_AllPages_Revision_7_WEB_PDF.pdf - see §19.xiv

Future regulatory landscape *(contd)*

balances safety and development, to be achieved through prudent negotiation and inclusive, mutually beneficial approaches.

Also announced was the establishment of a new UK-based AI Safety Institute, and a proposal for an International Panel on Artificial Intelligence Safety (independent from political interference) that would be able to inform policymakers and the public.

The AI Safety Summit is to be hosted every 6 months by a different country with the next summit scheduled for the first half of 2024 by Korea (mini virtual summit) and the one after that (second half of 2024) by France (full in-person summit).

These summits are likely to help frame concerns and standards around the future direction of the most powerful foundation models and general purpose AI.

Additional US regulatory developments

FTC

At the close of 2023 the FTC was also seeking comment[18] on proposed regulation (through the FTC's rule-making powers) of automated decision-making (within a document predominantly dealing with surveillance). This asks:

- Whether rules should require companies to take "specific steps to prevent algorithmic errors," and what kind of error rates are generally prevalent in AI.
- Whether companies should have to certify that AI they use meets accuracy, validity, and reliability standards. If so, who should set the standards – the FTC, industry, or companies' own published policies?
- Whether rulemaking should prohibit or limit companies from developing or using AI whose outcomes are "unfair or deceptive" under Section 5 of the FTC Act, and if so, whether the prohibition should be economy-wide or only apply in certain sectors.
- What kind of transparency companies provide to consumers about AI they use.

[18] https://www.ftc.gov/system/files/ftc_gov/pdf/commercial_surveillance_and_data_security_anpr.pdf

Future regulatory landscape *(contd)*

While not presently carrying any regulatory force, this consultation is strongly indicative of likely future rulemaking by the FTC. As before, compliance with the EU regulation would exceed what is proposed by the FTC.

White House blueprint

Further indication of the direction of travel in terms of US AI regulation comes from the publication by the administration administration in October 2022 of a proposal for an AI Bill of Rights[19] that again contains many of the same elements as the EU regulation. Whether such a bill will gain legislative force remains to be seen, but the similarity to the EU position, albeit without the same developed view of certification and regulatory enforcement, is instructive.

US Executive Order

On October 30, 2023, the Biden Administration released Executive Order (E.O.) 14110 on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. It establishes a government-wide effort to guide responsible artificial intelligence (AI) development and deployment through federal agency leadership, regulation of industry, and engagement with international partners. The E.O. directs over 50 federal entities to engage in more than 100 specific actions to implement the guidance set forth across eight overarching policy areas: Safety and Security; Innovation and Competition; Worker Support; Consideration of AI bias and civil rights; Consumer protection; Privacy; Federal use of AI; and International Leadership.[20]

NIST AI Risk Management Framwework

The National Institute for Science and Technology (NIST), the standard-setter for IT and cybersecurity in the US, has also published a draft risk management framework for AI[21]. While not a binding piece of regulation, failure to follow NIST standards without

[19] https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[20] https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

[21] https://www.nist.gov/itl/ai-risk-management-framework

Future regulatory landscape *(contd)*

demonstrating alternative controls leaves businesses exposed both to litigation and to FTC enforcement. The framework adopts a quality management methodology ("Map, Measure, Manage") that is analogous to what is proposed by the EU regulation; it seeks similar outcomes:

- Valid & Reliable: AI is accurate, able to perform as required over time, and robust under changing conditions.
- Safe: AI does not cause physical or psychological harm, or endanger human life, health, or property.
- Fair & Nonbiased: Bias in results is managed at the systemic, computational, and human levels.
- Explainable & Interpretable: The AI's operations can be represented in simplified format to others, and outputs from AI can be meaningfully interpreted in their intended context.
- Transparent & Accountable: Appropriate information about AI is available to individuals, and actors responsible for AI risks and outcomes can be held accountable.

Other frameworks

OECD Framework for the Classification of AI Systems

A tool developed by the Organisation for Economic Co-operation and Development (OECD) to help policymakers, regulators, legislators, and others understand and analyse different types of AI systems in line with the OECD AI Principles, which promote responsible and ethical AI development and deployment. The framework looks at systems along five key dimensions: People & Planet, Economic Context, Data & Input, AI Model, and Task & Output. Each dimension has its own set of properties and attributes that help assess the policy implications of the specific AI system under consideration.

Assessment List for Trustworthy Artificial Intelligence (ALTAI)

A self-assessment checklist based on the seven key requirements of Trustworthy AI outlined in the Ethics Guidelines for Trustworthy AI developed by the European Commission's High-Level Expert Group on AI (AI HLEG). The key requirements

Future regulatory landscape *(contd)*

are: Human Agency and Oversight; Technical Robustness and Safety; Privacy and Data Governance; Transparency; Diversity; Environmental and Societal Well-being; Accountability. ALTAI is a self-assessment tool, so its effectiveness depends on the honesty and objectivity of the user.

Practical Considerations

Preamble

This paper is necessarily focused on mitigating risks, especially regulatory risks, arising from the use of machine learning and AI. However, the intention is not to discourage the adoption of machine learning tools where these can add value without risking harm to individuals or groups. Many opportunities exist for positive deployment of AI, especially in safety systems where consistent and rapid responses are a key component of reducing risks of physical harm. For instance, while in the examples below this paper highlights risks arising from prototype self-driving systems on public highways, there is good evidence that the more controlled environments of industrial sites are suitable use cases for automation[22].

Similarly, while there are risks associated with AI generated text and image content – as highlighted by the Chinese regulation of deep synthesis – there are also significant productivity gains to be found provided that users understand the limitations. As a general rule, therefore, perhaps the key mitigation for all AI risks is adequate training and awareness-raising to ensure that both commissioners, developers and users of machine learning systems are properly aware of the risks, regulations and constraints, as well as the benefits and opportunities.

[22] https://www.roboticsbusinessreview.com/opinion/the-levandowski-pardon-and-the-pursuit-of-autonomous-vehicle-technology/

Practical Considerations *(contd)*

Over-estimation of capability risk

An overarching danger is the tendency to exaggerate the capabilities of machine learning models; take for instance the case where a Google engineer claimed sentience for their model[23], or the extensive hype surrounding the capabilities of ChatGPT[24]. This leads to users and other involved persons mistakenly treating machine learning models as sentient general AIs. When this happens, users often fail to account for the limitations of machine learning perception and sensitivity to context.

> *We need only remember the speed with which Microsoft's incorporation of ChatGPT technology into a function of its Bing search engine devolved into chaos to see an example of this risk*[25].

Recommendation

Ensure that proposals for use of AI accurately reflect real capabilities, and that limitations are properly communicated to all stakeholders and users.

[23] https://edition.cnn.com/2022/07/23/business/google-ai-engineer-fired-sentient/index.html#:~:text=Blake%20Lemoine%2C%20a%20software%20engineer,thousands%20of%20messages%20with%20it.

[24] What is going on with ChatGPT? | Arwa Mahdawi | The Guardian

[25] https://www.popularmechanics.com/technology/robots/a43017405/microsoft-bing-ai-chatbot-problems/

Practical Considerations *(contd)*

Case failure

Machine learning models are dependent on both training data and nature of inputs. Because they are not actually sentient, they are not "aware" of their limitations. This can lead to case failure, where a condition that is apparent to a human observer is not detected by the machine learning model.

> *Tesla's "self-driving" feature relies on machine vision processing. There have been several cases, one unfortunately fatal*[26]*, where the vision system could not adequately identify a white object as a physical obstruction – it being theorised that the model falsely identified the white area as a cloud or light sky instead – where this would have been obvious to a human.*

Recommendation

Designers should ensure, especially in safety-critical systems, that the compass of machine decision-making is properly understood and identify and mitigate the possible consequences of case failure. See also Zerili, Knott, Maclaurin & Gavaghan, *Algorithmic Decision-making and the Control Problem, Minds and Machines 29, 2019*[27]

[26] https://arstechnica.com/cars/2019/05/feds-autopilot-was-active-during-deadly-march-tesla-crash/

[27] https://link.springer.com/article/10.1007/s11023-019-09513-7

Practical Considerations *(contd)*

Reliance risk

Machine learning models can deliver tremendous efficiency gains, both from simple speed of operation and from tireless performance. The danger is that business models are constructed that do not allow for human intervention without impossible reduction in productivity. This in turn can lead to serious consequences for natural people such as e.g. when automated detection of suspicious financial activity leads to transaction or account blocking but the appeal process – operated by humans – is unable to address complaints within a reasonable timeframe.

> *This has been a problem for new-entrant financial services firms in particular, who have used machine learning extensively in detection and prevention of fraud and money-laundering. Failure to support this function with sufficient numbers of trained staff meant that customers who were subject to a false-positive result and subsequent account or transaction block were unable to have the decision rectified in an acceptable time*[28].

Recommendation

Ensure that all processes which can have a material effect on natural people have mechanisms for timely review and intervention by human operators, and that such operators are recruited and trained in sufficient numbers. This is also a key regulatory requirement.

[28] https://www.altfi.com/article/8953_why-account-closures-are-a-bigger-problem-for-neobanks-compared-to-incumbents

Practical Considerations *(contd)*

Rigidity risk

Machine learning models do not exercise judgement or provide for contextual variation from learned rules. Human rule- and law-making has historically relied on a degree of flexibility – whether arising from individual judgment or from limitations in enforcement capability – to compensate for the impossibility of writing defined laws or rules that take account of all circumstances. Machine learning models because of their speed and scalability can offer "perfect" enforcement of rules and laws, but this is unlikely to deliver the best social or organisational outcome.

> *Examples here include automated censorship of social media posts, which recently led to an amusing and much publicised case where the British Trust for Ornithology lost access to its Twitter feed following a post about woodcocks[29]. More seriously, the use of machine learning for e.g. traffic enforcement in place of human officers can lead both to situations where mitigating circumstances are ignored and to a resultant loss of public support for traffic rules.*

Recommendation

Where AI systems are to be used to enforce rules, study the consequences of complete and inflexible enforcement before committing to be sure that potential negative social effects do not outweigh the benefits.

[29] https://www.bbc.co.uk/news/uk-england-norfolk-64451977

Practical Considerations *(contd)*

Training failures

At present this is the single greatest area of concern in machine learning. Where "AI" is based on some form of deep learning such as general adversarial networks, the quality of that training largely conditions the success of the model. A range of potential failures can be identified from historical examples:

Training data sourcing

Where training data involves personal data, the data must be sourced ethically and in compliance with relevant data protection laws. This is often harder to do than it might appear. Notably in GDPR the fact that information is, or appears to be, in the public domain does not in itself provide a justification for processing the data; by the same token just because data are available within an organisation does not mean that their use in training a machine learning model is compatible with the purposes for which the data where originally collected.

Clearview, a vendor of facial recognition software, has been fined substantial sums by several different European data protection authorities including the UK, France and Italy, for making use of facial images obtained without the consent of the data subject[30].

Recommendation

Do not use data from existing systems to train new machine learning systems unless the data subjects concerned have been consulted and either informed or consented depending on the assessment of the lawful basis concerned.

Training data bias (aka proxy risk)

Much training data, whether collected from real sources or synthesised, can incorporate inherent biases. Precisely because machine learning models are not sentient, they

[30] https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/05/ico-fines-facial-recognition-database-company-clearview-ai-inc/

Practical Considerations *(contd)*

will replicate this bias in the model. It can be extremely difficult – and in some cases impossible – to remove this bias from the model once trained in. In many cases, attributes that are highly correlated with demographic features, known as proxies, can contribute to algorithmic discrimination.

> *Amazon was an early adopter of AEDT, building a CV-screening system that sought to identify the best engineers from submitted CVs. The system was trained on the CVs of existing successful engineers. Unfortunately, the vast majority of those engineers were male and the resulting model was extremely adept at identifying and selecting for male CVs even when obvious gender identifiers were removed. Amazon was unable to remove this bias from the model and was forced to retire the system*[31].

Recommendation

Use rigorous data science to ensure that training data does not contain bias and/or to create synthetic training data that properly reflects the desired diversity of outcome. Test systems for bias, including seeking to identify potential proxies that may have been incorporated into the learned model.

Outcome biases

By the same token, machine learning model training works by having the trainer (directly or by dataset or condition selection) teaching the machine the different between right and wrong answers. The machine is, therefore, seeking to replicate the trainer's definition of success by design. Any biases in the trainer that are not detected and corrected before training will therefore be replicated in the model. This can be exacerbated when a feedback loop is established but the human in the loop continues to exhibit bias in confirming results provided by the model.

> *An exhaustive study of this problem (and others mentioned here) can be found in Garvie, Clare, A Forensic Without the Science: Face Recognition in U.S. Criminal Investigations, Center on Privacy & Technology at Georgetown Law (2022)*[32].

[31] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[32] https://mcusercontent.com/672aa4fbde73b1a49df5cf61f/files/2c2dd6de-d325-335d-5d4e-84066159df71/Forensic_Without_the_Science_Face_Recognition_in_U.S._Criminal_Investigations.pdf

Practical Considerations *(contd)*

Recommendation

Ensure that operators and trainers receive training to determine their own conscious and unconscious biases and that feedback loops are properly monitored – including by the inclusion of known incorrect outputs in order to test outcome bias and feedback loop quality.

Prompts and hinting

There is an additional risk arising from the need for most current machine learning systems to supplement their direct learning with "prompting" or "hinting" – operator-provided overrides that seek to reinforce desired outputs or suppress undesirable ones. This is particularly common in natural language interfaces such as chatbots (including e.g. ChatGPT) but is also seen when attempts are made to compensate for perceived bias in a trained model. The issue is that the hints themselves can embed operator bias, act to restrict open access to information or prevent a continuous learning model from developing properly in response to new input data. By definition, a prompt that states "the following content is offensive" embeds the operator's personal perception into the model, where it will be treated as fact.

> *The COMPAS algorithm (referred to under Transparency risks below) could be prompted to ignore sensitive characteristics like race and gender. However, separate analysis of offender data sets indicates that female recidivism is less likely, so prompting the system to ignore gender would in fact result in unfair outcomes for women*[33].

Recommendation

If prompts are used to control for bias in a model, establish an independent review process and apply rigorous statistical analysis to ensure that the prompts are actually improving fairness rather than simply introducing alternative bias.

[33] https://arxiv.org/abs/1701.08230

Practical Considerations *(contd)*

Correlative error

Deep-learning machine learning models work by finding matching patterns between inputs and outputs. However, one of the problems with transparency (see later) is that understanding and documenting the specific patterns identified by the machine can be effectively impossible. As a consequence there is a danger of assuming that, simply because a model can reliably produce the desired outputs from a set of inputs, it is doing so in an analogous fashion to a human process. This is not the case and, accordingly, can lead to unpredictable and undesirable outputs when presented with new inputs.

> *A system trained to detect pneumonia using X-rays worked well in a single hospital but failed when sent images from other hospitals. It turned out that the system had included metal marker tokens only used in the original hospital in its input data and correlative matching.*[34]

Recommendation

Make sure that correlations are understood. Remove elements from datasets to ensure that these are not part of the model's logic and test early with data from other sources – noting though always the constraints on testing with live personal data.

[34] https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683

Practical Considerations *(contd)*

Query ambiguity, error or bias

Machine learning systems designed to interrogate data answer the question asked. Allied to the over-estimation of capability risk mentioned earlier is the risk that the operator assumes an understanding of context or of intent by the machine that is not present. This can result in operator bias or assumption driving incorrect outputs; in reinforcement of existing assumptions – a confirmation bias problem – or in actual failure, in the sense that the question answered is not the one the operator thought they were asking.

This is also a general observation in statistics and data analytics. In AI in particular the sudden explosion in use of ChatGPT has led to policies banning the publishing of material generated by the program because so many of the answers are incorrect, often as a result of ambiguity or lack of context in the question.

> *A sample from the explanation for such a ban from Stackoverflow: "The primary problem is that while the answers which ChatGPT produces have a high rate of being incorrect, they typically look like they might be good and the answers are very easy to produce. There are also many people trying out ChatGPT to create answers, without the expertise or willingness to verify that the answer is correct prior to posting. Because such answers are so easy to produce, a large number of people are posting a lot of answers. The volume of these answers (thousands) and the fact that the answers often require a detailed read by someone with at least some subject matter expertise in order to determine that the answer is actually bad has effectively swamped our volunteer-based quality curation infrastructure."*[35]

Recommendation

Consider use of AI to be a skilled task and ensure that everyone responsible for setting the problems to be solved and questions to be answered is both trained in prompt engineering and has the knowledge to identify incorrect results and positively influence the feedback loop.

[35] https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned

Practical Considerations *(contd)*

Cyber risks

As recognised in the proposed EU AI Act, use of machine learning models introduces additional cyber risk. This arises in three ways:

- Most "AI" models – both learning models and static algorithms - rely on large datasets, often combining data from multiple sources. Compromising these datasets could cause significant harm to data subjects; in particular it may permit reidentification of other deidentified or pseudonymised data. Protecting these datasets both at rest and under processing adds to the burden on organisational information security.
- Where machine learning is being used to automate processing there is heightened risk arising from any possible compromise of the system. Especially where the decision-making criteria are intrinsically opaque, as in most deep-learning models, it can be difficult to tell that a system has been compromised; the result can be undetected fraud, data alteration or data exfiltration.
- Machine learning systems are frequently built on external code bases, including open source, and will often use a variety of flexible compute resource. The external code, the repositories from which it is obtained and the compute resource have all historically been targets for malicious actors; users of such code may not have the internal resources to check it for exploits. In addition machine learning remains a cutting-edge technology where changes are being made at great speed and, in some cases, with limited oversight or focus on security – much of the work is experimental and was not intended by its authors to be used with live data (particularly sensitive personal data). As a consequence, there is a greater risk of zero-day exploits arising from code defects which may also be exploited by malicious actors or result in accidental breaches.

> *Both British Airways and Ticketmaster suffered large-scale data breaches – and were fined millions of pounds – as a consequence of insufficiently secure deployment of third-party NLP machine learning models ("chatbots").*[36]

[36] https://www.shlegal.com/news/an-analysis-of-the-monetary-penalty-notice-issued-by-the-information-commissioner-s-office-to-ticket-master-uk-limited-dated-13-november-2020

Practical Considerations *(contd)*

Recommendation

Ensure that information security teams are properly skilled in assessing cyber-risks from machine learning, and avoid the temptation to deploy poorly understood technology

Transparency risks

All "AI" models are complex. Deep-learning models, especially convoluted networks, are to some extent unknowably so once trained. Explaining the logic by which outputs arise from inputs in language that can be understood by data subjects is unlikely to be possible. This creates significant challenges in transparency, something that is explicitly recognised in the EU regulation and in other emerging AI legislation.

> *Northpointe produces a system, COMPAS, that is widely used in the US to determine parole for criminal offenders. The system has been repeatedly accused of exhibiting systemic racism but the vendor has not provided insight into the operation of the model and the system's users are unable to examine the decision-making logic themselves*[37].

Recommendation

Ensure that vendors are able to provide satisfactory documentation and explanation of algorithmic functions and learning methodologies. For internal development avoid the use of learning techniques and models that become "black boxes" where it is not possible for a human to elucidate the decision-making process.

[37] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Practical Considerations *(contd)*

Oversight and governance

Not only is emerging regulation clear on the need for enhanced oversight of machine learning, it's also essential as part of demonstrating compliance with existing rules including the GDPR's Article 22 and is central to the conceit of the EU AI regulation, the Chinese AI ethics code and the US blueprint for an AI bill of rights. This includes the need to offer escalation to a human decision-maker – at least for any machine-made decision that has a material effect on the data subject – but more broadly there is an overriding requirement to show fairness towards data subjects. Given the potential issues highlighted in this document it's clear that the risk of acting unfairly – either through undetected bias or by over-reliance on decision making without proper context – is significantly heightened by the use of machine learning. Naturally machine learning can also be used to improve fairness – a model that can be demonstrated not to have bias is likely to act fairly more consistently than humans given the same information.

> *Deliveroo's Italian operation was fined €2.5m in 2021 for implementing an automated scheduling system that was considered to treat delivery riders unfairly and for failing to provide human oversight and an appropriate intervention and appeals process.*[38]

Recommendation

Ensure that processes supported by AI have proper governance and always provide for timely appeal and human intervention. Ensure that all automated decision making is subject to ongoing review to ensure fairness and that the deployment of such systems has provably positive benefits for data subjects.

[38] https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9685994

Practical Considerations *(contd)*

Rights limitations and leakage risk

This is one of the most challenging areas in ML. Where personal data have been used to train a model, an abstracted representation of that data is embedded within the system. If a data subject asks for a copy of their data, or – more difficult still – for that data to be corrected or deleted, it can be impossible for the model operator to comply, both because of the transparency concerns cited earlier (the data cannot easily be located) and because of the impossibility of removing specific data from or manually adjusting a trained model. While there are some protections in law against being required to comply with the exercise of these rights where it is technically unfeasible, this must be set in the context of the original basis for processing. If this basis is, for example, the organisation's legitimate interest it can be very difficult to demonstrate that this interest outweighs the data subject's rights; if the model prevents the exercise of those rights, then it is conceivable that an objection to the use of the data could lead to enforcement requiring the entire model to be shut down.

A 2017 paper, Villaronga, Kieseberg & Li, *Humans forget, machines remember: Artificial intelligence and the Right to Be Forgotten, Computer Law & Security Review, v34*[39] suggests persuasively that for some types of machine learning it may be impossible to delete personal data that has been incorporated into the model.

Recommendation

There is no easy solution to this problem. Avoiding the use of identifiable personal data for model training, where possible, and focusing on minimising its use in any system that operates on a continuous learning basis will reduce the risk. Some vendors, including IBM[40], claim to have toolkits that allow deletion and correction of personal data in trained models. This paper is not a in a position to endorse any particular toolset or approach.

[39] https://www.sciencedirect.com/science/article/abs/pii/S0267364917302091

[40] https://developer.ibm.com/blogs/data-minimization-for-machine-learning/

Practical Considerations *(contd)*

Abuse

Surveillance

The European regulations in particular highlight the risk of excessive use of machine-recognition of biometrics but the underlying concern is more broad. The scalability of automated processing permits extensive and ubiquitous surveillance across a wide range of technologies and circumstances – whether, e.g., monitoring working patterns by analysing keystrokes, software use and network traffic, or combining CCTV, facial recognition and device detection to track individuals' movements. Regulators in most countries have expressed profound concerns about the growth of surveillance and in particular its application in asymmetric relationships such as that between employer and employee. Again this is at its root an issue of fairness, but there are also potential clashes with fundamental human rights enshrined in e.g. the European Charter of Fundamental Rights.

> *White Castle, a US fast food chain, is presently being sued by former employees protesting over-collection and processing of biometric data (specifically fingerprints). The case is still to settle but the potential damages alone run to some US$17bn.[41] The recent parliamentary committee discussion of movement tracking by Royal Mail is also instructive.[42]*

Recommendation

Take care to consider the balance of individual rights and productivity (or security) when considering the implementation or extension of surveillance. From a litigation risk perspective especial care is needed regarding remote biometrics (face recognition, centralised fingerprint recognition and so forth) but modern working patterns also see temptations to use location or activity data (such as keystrokes) in automated surveillance models.

[41] https://www.natlawreview.com/article/17-billion-slider-illinois-supreme-court-decides-white-castle-bipa-case

[42] https://www.telegraph.co.uk/business/2023/02/02/royal-mail-dystopian-tracking-system-tells-bosses-when-postmen/

Practical Considerations *(contd)*

Misuse of capability including recombination risk

The pattern-matching capabilities of machine learning models are powerful; as noted from some of the examples above they are particularly good at recognising patterns in data that are not apparent to human observers – allowing, for example, the deduction of gender, health conditions or identity from data that has theoretically been stripped of the relevant indicators. Given wide access to datasets the possibilities for abuse of this capability are considerable; such abuse may be intentional and malicious or equally possible a well-intentioned effort to derive useful information without recognition of the privacy, regulatory or broader ethical risks involved.

One extension of this risk, but one not confined to learning models, is that data from multiple sources when combined may reveal information, including sensitive details, about data subjects that were either thought to have been suppressed or which were never intentionally collected in the first place. However, since many machine learning projects specifically attempt to derive useful insight from the combination of multiple datasets this risk is particularly pronounced.

> *Machine-based pattern recognition has been used to de-anonymise individuals based on supposedly anonymous location data bought in from apps. Notably this featured in two cases where interest groups used machine learning to reveal sexuality – so also breaching rules around processing of sensitive data. In both cases the targets were Catholic priests who were "outed" as being gay and in consequence dismissed or risked dismissal from the priesthood[43] [44].*

Recommendation

Uses of machine learning systems to derive targeting or categoric information from datasets collected for other purposes must always be subject to ethical review. Ethics committees should not only consider whether the intended use is acceptable but also both whether any unintended negative consequences are possible and whether the dataset, technology or outputs could be abused.

[43] https://arstechnica.com/tech-policy/2021/07/catholic-priest-quits-after-anonymized-data-revealed-alleged-use-of-grindr/

[44] https://www.theguardian.com/world/2023/mar/10/colorado-catholic-group-identify-priests-gay-apps

Practical Considerations *(contd)*

Real world examples

Recommendation algorithms on social media platforms like Instagram, X and YouTube have been shown to prioritise extreme and emotionally compelling content, regardless of accuracy. In 2015-8, Facebook's tendency to promote fake news articles designed to maximise clicks based on sentiment overwhelmed legitimate news and information sources in Myanmar and helped foment racial and political tensions in the country. In 2018, a United Nations investigation determined that the violence against the Rohingya constituted a genocide and that Facebook had played a "determining role" in the atrocities.[45]

AI models tend to be trained on biased data sets, which can lead to biased outputs. Such bias can reinforce existing social inequities, as was the case with a machine learning algorithm deployed by the city of Rotterdam in 2017, that was intended to help detect welfare fraud by assigning people risk scores. While the data used to train the model did not explicitly include race, ethnicity, or place of birth among the 315 variables, it used to calculate the risk score, some second order characteristics (for example, migrant status) were used by the model as inadvertent proxies for ethnicity and gender, and it began discriminating unfairly against people on these bases. Such discrimination is illegal under Dutch law, and the tool had to be withdrawn from service in 2021.[46]

In 2019 the US National Institute of Standards and Technology (NIST) tested 200 commercially available facial recognition algorithms on their accuracy for "one-to-one" matching (matching a photo of someone to another photo of the same person in a database, used to unlock smartphones or check passports), and "one-to-many" searching (determining whether a photo of someone has any match in a database, used by police departments to identify suspects in an investigation). All the algorithms developed in the US proved less accurate at matching Asian, African American and Native American faces than they were at matching white faces.[47]

A recruitment engine developed by Amazon in 2014 had to be shut down after three years because it was not rating potential applicants in a gender-neutral manner. The company had trained the algorithm on resumés from its previous 10 years of hiring, but

[45] https://www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait/

[46] https://www.wired.com/story/welfare-state-algorithms/

[47] https://www.technologyreview.com/2019/12/20/79/ai-face-recognition-racist-us-government-nist-study

Practical Considerations *(contd)*

as it had predominantly hired men in that period, the model had learned that women were not suitable for the job. Despite multiple attempts to fix the bias the project eventually had to be abandoned.[48]

Minimum requirements – and next steps

With so much regulatory activity planned or proposed, the situation with regard to AI compliance is confusing to say the least. And in confusing situations, the temptation is often to do nothing, to see how things pan out. However, now that the EU AI Act has been enacted, we can be sure that AI regulation is coming, and the shape of it – in the EU especially – is becoming very clear. It therefore makes sense to be prepared for when the legislation does arrive, rather than wait for it to happen and then to try to retrofit it to tools and models your business already has in place.

One certainty is that AI regulations will follow on from and build upon existing data protection legislation and practices, as these provide the best existing exemplar for audit and assessment. As with most compliance, internal transparency and good record keeping are crucial. Assessing your AI infrastructure as you develop and deploy it in line with the main risk assessment principles being used by international legislators for their own guidance will not only save you unnecessary time, expense and effort later but will equally guide your own exploration of this exciting new technological space and make sure that your efforts produce outcomes that are genuinely beneficial to your employees and customers and a credit to your business.

[48] https://www.reuters.com/article/idUSKCN1MK0AG/

aiEthix from Securys

Your AI compliance partner

In a world of fast evolving AI regulations, it is difficult for companies to understand how they should be using artificial intelligence within their businesses. aiEthix from Securys, the global privacy and data protection consultancy, provides practical AI implementation and governance solutions that take the guesswork out of managing AI responsibly and allow firms to explore machine learning opportunities that will withstand future developments in technology and legislation.

What we offer

Our team has many years of experience helping organisations navigate many kinds of data protection and privacy challenges. Now, we are applying our advanced knowledge and unique insights into managing risk in these areas to the burgeoning and exciting field of machine learning.

Our approach is refreshingly practical and our methods leave no stone unturned. Our start point is our clients' aspirations to ensure their AI infrastructure aligns with wider organisational goals.

AI discovery

Mapping your existing AI footprint

An AI discovery draws upon our experience of data privacy discovery and audit for organisations of all types to conduct low impact high insight analysis of your workflows and data flows involving internally built and externally sourced AI/ ML tools as well as unsanctioned "shadow AI", which we investigate in a non-confrontational, non-judgemental way.

AI Governance

Making light work of your regulatory requirements

Data knows no border. Even a modest AI/ML deployment exposes you to regulatory and legislative enforcement across multiple jurisdictions. Let our experts work alongside your risk and compliance teams to identify, monitor and mitigate all the risk that groundbreaking technologies run, without sacrificing your ability to profit from the opportunities that AI/ML bring.

AI by Design

Put data protection to work

Our consultants will work with you at every stage of the AI product cycle, from inception to delivery, advising on data preparation and system architecture to ensure that your models source and handle data responsibly, conform to relevant regulation, and are adequately risk-assessed and proactively risk-mitigated.

About Securys

Global knowledge, locally delivered

Our teams of international experts, operating across dozens of jurisdictions, deliver relevant advice and see the world through your lens.

A deep understanding of data

We bring decades of experience delivering privacy-by-design programmes to clients of all scopes and scales, across multiple industries and jurisdictions.

Practical approach

We will tackle any bespoke AI project in a constructive fashion, breaking down barriers within teams to enable better compliance and a united approach to implementation.

About Securys

Securys is a specialist data privacy consultancy with a difference. We're not a law firm, but we employ lawyers. We're not a cybersecurity business but our staff qualifications include CISSP and CISA. We're not selling a one-size-fits-all tech product, but we've built proprietary tools and techniques that work with the class-leading GRC products to simplify and streamline the hardest tasks in assuring privacy.

We're corporate members of the IAPP, and all our staff are required to obtain one or more IAPP certifications. We are ISO 27001 and ISO 27701 certified with a comprehensive set of policies and frameworks to help our clients achieve and maintain certification.

Above all, our relentless focus is on practical operational delivery of effective data privacy for all your stakeholders.

We're not just a consultancy. We're your privacy engine room. We can stand in your boardroom and do strategy with the top team, and work with your compliance teams to solve knotty problems. We can audit your compliance and deliver drillable risk dashboards across the organisation. But above all we can get involved at ground level and help your frontline teams get the job done. That's Privacy Made Practical®.

Securys®
Global Data Privacy Experts

T 0800 193 8700
E info@securys.co.uk
W www.securys.co.uk

in securys-limited