

AI safety measures: a comparative chart

As technology continues to develop rapidly, legislators and regulators around the world are racing to keep up – or indeed catch up – by implementing measures to protect those whose interests are affected by AI systems.


Unsurprisingly there is no global standard on AI regulation. But when we look at the existing and proposed safeguards intended to protect individuals where AI systems are introduced into the workplace, clear categories emerge.

Set out in the table below is an analysis of how measures in key existing and proposed legislation could be categorised on this basis.

This is by no means a comprehensive list – draft legislation is being debated and progressed around the world. For example, Canada’s AI and Data Act envisages a risk based approach to regulation; the US Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence supports the creation of standards for trustworthy AI; and in the UK, the Department for Science Innovation has published [detailed guidance](#) on the responsible use of AI in recruitment. However, it will be interesting to see the extent to which the “Brussels effect” sees the approach taken in the EU spread around the world.

Understanding the columns:

- **Consultation about use:** Is there a requirement to consult with employees/unions about the use of workplace AI systems?
- **Accountability and governance:** Are there specific roles, frameworks, policies or reporting structures required to manage the use of AI in the workplace?
- **Impact assessments:** Is there a requirement to undertake an impact assessment prior to use?
- **Auditing and monitoring:** What steps are required to ensure that the AI system is safe to use? Auditing meaning independent evaluation and monitoring meaning regular non-independent evaluation.
- **Transparency and explainability:** What level of openness is required about the use of AI and / or how the system’s decisions are reached? Transparency meaning that the use and purpose of AI is disclosed. Explainability meaning that how an outcome was reached is disclosed.
- **Human oversight / intervention:** What human involvement is required? Human oversight meaning someone is required to oversee the operation of the system. Human intervention meaning that a human interposed between the AI system expressing preferences and the final decision.
- **Contestability:** Are measures required to enable an individual affected by an AI decision to be able to challenge it effectively?

	Overview	Consultation about use	Accountability and governance	Impact assessments	Auditing and monitoring	Transparency and explainability	Human oversight / intervention	Contestability
<p>UK Government White Paper: principles for regulators</p> 	<p>Framework of five principles to guide regulators, rather than legislation:</p> <ol style="list-style-type: none"> 1. Safety, security and robustness 2. Appropriate transparency and explainability 3. Fairness 4. Accountability and governance 5. Contestability and redress <p>Focus is on regulating the use not the technology.</p> <p>Regulators' strategic responses published here.</p>	<p>No reference</p>	<p>4th regulatory principle.</p> <p>Guidance states that "[g]overnance measures should be in place to ensure effective oversight of the supply and use of AI systems, with clear lines of accountability established across the AI life cycle".</p>	<p>Not specifically referred to in principles but covered by assurance techniques and technical standards identified as key to supporting regulatory framework. See AI standards hub, roadmap for AI assurance and portfolio of AI assurance technologies</p>	<p>1st regulatory principle refers to the need to continually identify, assess and manage risks throughout the AI life cycle.</p> <p>4th regulatory principle also covers need for effective oversight.</p>	<p>2nd regulatory principle.</p> <p>Guidance states that "AI systems should be appropriately transparent and explainable", such that the other principles have meaningful effect (e.g. accountability).</p> <p>Parties affected by an AI system should have "sufficient information about AI systems to be able to enforce their rights".</p>	<p>4th regulatory principle.</p> <p>Guidance states that governance measures should "ensure effective oversight of the supply and use of AI systems"</p>	<p>5th regulatory principle.</p> <p>Guidance states that "[w]here appropriate, users, impacted third parties and actors in the AI life cycle should be able to contest an AI decision or outcome that is harmful or creates material risk of harm."</p> <p>Regulators to implement proportionate measures to ensure outcomes of AI use are contestable where appropriate.</p>

Artificial Intelligence (Regulation and Employment Rights) Bill




	Overview	Consultation about use	Accountability and governance	Impact assessments	Auditing and monitoring	Transparency and explainability	Human oversight / intervention	Contestability
	<p>Taskforce commissioned by the TUC has published a draft Bill which sets out a detailed legislative framework for regulating the use of AI in the workplace.</p> <p>Most use cases in the employment context will be high-risk.</p>	<p>Direct consultation with affected workers / employees required at least one month before high-risk decision making takes place. This must be repeated every 12 months.</p>	<p>An employer must maintain a register of AI systems used for high-risk decision making.</p>	<p>An employer can not undertake high-risk decision-making until a Workplace AI Risk Assessment has risk assessed an AI system in relation to health and safety, equality, data protection and human rights. While high-risk decision making continues, this must be repeated at least every year.</p>	<p>Auditing the AI system could rebut the burden of proof in a discrimination claim.</p>	<p>Right to personalised explanations for high-risk decisions which are or might reasonably be expected to be detrimental to employees, workers or jobseekers.</p>	<p>Employees, workers or jobseekers will be entitled to a right to human reconsideration of a high-risk decision.</p>	<p>In addition to the right to human reconsideration, breaches of duties such as the duties to consult and risk assess would be actionable in the employment tribunal, potentially giving rise to an injury to feelings award.</p> <p>Automatic unfair dismissal when the decision arises from the unfair reliance on high-risk AI decision making.</p> <p>Changes regarding the burden of proof strengthen discrimination protection.</p>

The EU AI Act





Overview	Consultation about use	Accountability and governance	Impact assessments	Auditing and monitoring	Transparency and explainability	Human oversight / intervention	Contestability
<p>Risk based legislation.</p> <p>Published in Official Journal on [17] June 2024</p> <p>Most use cases in the employment context will be high-risk.</p> <p>The majority of obligations fall on providers but deployers must also satisfy a range of requirements.</p>	<p>Deployer: 1) Must inform workers’ representatives that they will be subject to a high-risk AI system.</p> <p>2) Must inform individuals that the deployer plans to use a high-risk AI system to make or assist in making decisions relating to an individual.</p>	<p>Provider: 1) The system must ensure the automatic recording of events.</p> <p>2) Must establish, implement, document and maintain a risk and quality management systems.</p> <p>3) Must meet data governance requirements, including bias mitigation.</p> <p>4) Must draw up and maintain technical documentation.</p> <p>5) Comply with registration, record-keeping, logging and traceability obligations, as well CE conformity obligations.</p> <p>Deployer: 1) Retain automatically generated logs for at least 6 months.</p>	<p>Deployer: 1) Must use information from providers to carry out a data protection impact assessment (DPIA) (likely to be required for high-risk system),</p> <p>2) For certain deployers and certain high-risk systems, must undertake a fundamental rights impact assessment, assessing the system’s impact on fundamental rights.</p>	<p>Deployer: 1) Monitor the operation of the system on the basis of instructions for use.</p> <p>2) Ensure input data in their control is relevant and sufficiently representative.</p>	<p>Provider: 1) System must be designed so that deployers can interpret the output, use it appropriately and the instructions must include information specified in the Act.</p> <p>2): Users must know that they are Interacting with an AI system and AI-generated content must be identifiable (limited risk systems).</p> <p>Deployer: 1) If AI generated decision results in legal or similarly significantly effects (most employment decisions), must provide a clear and meaningful explanation of the role of the AI system in the decision-making process and the main elements of the decision.</p>	<p>Provider: 1) System must be designed to allow for effective human oversight.</p> <p>Deployer: 1) Take certain appropriate technical and organisational measures and to assign human oversight.</p> <p>2) Assign someone to oversee the AI system who is trained, competent, and has the support and authority they need.</p>	

UK/EU GDPR




Overview	Consultation about use	Accountability and governance	Impact assessments	Auditing and monitoring	Transparency and explainability	Human oversight / intervention	Contestability
<p>The GDPR is a Regulation that requires businesses to protect the personal data and privacy of EU citizens.</p>		<p>The accountability principle means as well as being responsible for complying with data protection law, you must also demonstrate that compliance in any AI system that processes personal data.</p>	<p>A Data Protection Impact Assessment (DPIA) is required for processing that includes innovative technologies or the novel application of existing technologies.</p> <p>Therefore a DPIA should be carried out before an AI system is deployed</p>	<p>Undertaking a DPIA, as well as the need to define procedures for ongoing compliance supervision and AI system audits. Also, regular checks to ensure AI systems function as anticipated/required.</p>	<p>Individuals must be informed if their personal data is going to be used in an AI system. This information should be provided at the point of data collection. Individuals also need meaningful information about logic and consequences of any solely automated decisions.</p>	<p>The automated decision making provision applies where there is solely automated decision-making that has legal or similarly significant effects. Often referred to as "human in the loop", for human review to be meaningful, it should come after the automated decision has taken place and it must relate to the actual outcome.</p>	<p>Individuals must be able to contest a decision in a timely manner in order for the processing to be fair and compliant.</p> <p>Appropriate measures must be put in place to ensure individuals can exercise their rights.</p>


	Overview	Consultation about use	Accountability and governance	Impact assessments	Auditing and monitoring	Transparency and explainability	Human oversight / intervention	Contestability
<p> NY City Law: Local Law 144 of 2021  </p>	<p>This prohibits employers and employment agencies from using an automated employment decision tool (AEDT) in New York City unless they ensure a bias audit was done and required notices provided.</p>				<p>Mandatory annual independent audit of the AEDT. Summary of the most recent results must be published. Independent auditor must (as a minimum) evaluate calculations of selection or scoring rates and the impact ratio across sex categories, race/ethnicity categories, and intersectional categories.</p>	<p>Employers and employment agencies must notify employees and job candidates who are residents of New York City that they are using an AEDT and the job qualifications or characteristics the AEDT will assess.</p>		<p>Compliance is enforced by the NYC Department of Consumer and Worker Protection. Claims of discrimination involving an AEDT can be made to the NYC Commission on Human Rights.</p>

	Overview	Consultation about use	Accountability and governance	Impact assessments	Auditing and monitoring	Transparency and explainability	Human oversight / intervention	Contestability
<p> Illinois AI Video Interview Act (820 ILCS 42/1)  </p>	<p>This Act applies to all employers that use AI tools to analyse video interviews of candidates for positions based in Illinois.</p>	<p>No consultation obligations but the applicant's consent must be sought by employers to be evaluated by AI before the video interview. Employers may not use AI to evaluate a video interview without consent.</p>	<p>Video interviews must only be shared with those whose expertise or technology is required to evaluate the interview. In other words, video interviews may be shared with vendors providing the AI used to analyse the interview.</p>			<p>Employers must inform candidates before analysis takes place that AI is being used to assess their suitability for the role, how the AI works, and which characteristics will be used in the evaluation. The candidate's consent is required before any analysis takes place.</p>		

Colorado:
Senate Bill 205



Overview	Consultation about use	Accountability and governance	Impact assessments	Auditing and monitoring	Transparency and explainability	Human oversight / intervention	Contestability
<p>This Act is the first comprehensive piece of AI legislation in the US but does not come into force until 1 February 2026. However, the Act may still be subject to scrutiny and amendment to 'fine tune' the provisions.</p> <p>Like the EU AI Act, the majority of obligations apply to high-risk AI systems (which may include systems used to make decisions about employment or employment opportunities)</p>		<p>Deployers and Developers: Requirement to use reasonable care to safeguard consumers from the risks of algorithmic discrimination (whether known or reasonably foreseeable).</p>	<p>Deployers: When deploying high-risk systems deployers must undertake impact assessments in respect of those systems at least once a year or within 90 days of any intended and substantial modification to any such system.</p>	<p>Deployers and Developers: Varying reporting requirements to report certain confirmed or likely instances of algorithmic discrimination to the Attorney General within 90 days. Developer's reporting obligations also include providing such information to deployers that are known or other developers of the particular system.</p> <p>Deployers: High-risk AI will need to have a risk management policy and program, and an annual assessment to check they do not cause algorithmic discrimination.</p>	<p>Developers and Deployers: Requirement to provide public notices online which includes certain information about their AI. Deployers also have transparency obligations which may require an additional pre or post use notice be provided where certain decisions will be/ have been made using high risk AI systems.</p> <p>Developers: Also required to provide both those deploying high-risk AI systems as well as other developers who may be substantially and intentionally modifying their systems with certain transparency information.</p>		<p>Deployers: Where a high-risk system leads to an adverse consequential decision, deployers must provide notice which – among other things – provides the consumer with the chance to correct personal data and appeal the AI system's decision.</p>

	Overview	Consultation about use	Accountability and governance	Impact assessments	Auditing and monitoring	Transparency and explainability	Human oversight / intervention	Contestability
<p>Singapore</p> 	<p>Patchwork of frameworks and/or guidelines created by statutory/regulatory bodies instead of legislating.</p> <p>For example, the Model AI Governance Frameworks for both generative and 'traditional' AI, and the Personal Data Protection Commission has issued Advisory Guidelines on Use of Personal Data in AI Recommendation and Decision Systems.</p>		<p>Accountability is included in the compilation of ethical principles within the Model Framework for 'Traditional' AI. Further, accountability is one of the nine dimensions under the Model AI Governance Framework for Generative AI.</p>		<p>Auditability is included in the compilation of ethical principles within the Model Framework for 'Traditional' AI. Testing and assurance is also one of the nine dimensions under the Model AI Governance Framework for Generative AI.</p>	<p>One of the two guiding principles under the Model Framework for 'Traditional' AI is for AI decisions to be explainable, transparent, and fair. Explainability and transparency are both also included in the compilation of ethical principles collated within the framework. Additionally, transparency around content provenance is one of the dimensions under the Model AI Governance Framework for Generative AI.</p>	<p>A guiding principle under the Model Framework for 'Traditional' AI is for systems to be human centric.</p>	<p>The responsibility, accountability, and transparency principle, included in the compilation of ethical principles within the Model Framework for 'Traditional' AI, proposes making available avenues of redress for adverse individual or societal effects of an algorithmic decision system.</p>