MEVITAE

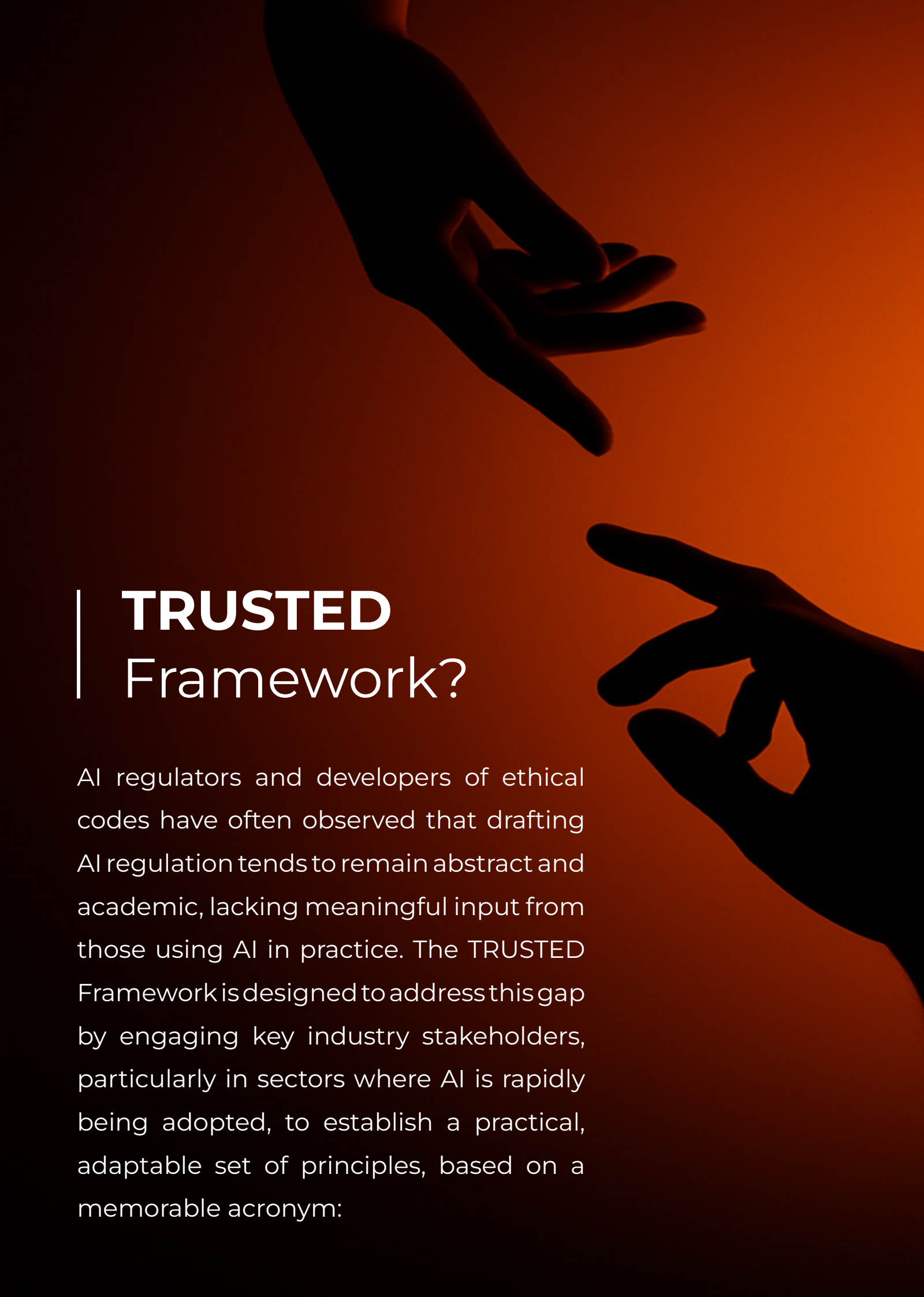
THE TRUSTED FRAMEWORK

ETHICAL AI IN PRACTICE

CONTENT

Welcome to the **TRUSTED** Framework

These principles are intended to guide the ethical, transparent, and responsible use of AI in real-world business operations, with an initial emphasis on HR, but also extending to legal technology (including e-discovery and group discrimination cases), supplier and vendor management, and other critical organizational areas.



TRUSTED Framework?

AI regulators and developers of ethical codes have often observed that drafting AI regulation tends to remain abstract and academic, lacking meaningful input from those using AI in practice. The TRUSTED Framework is designed to address this gap by engaging key industry stakeholders, particularly in sectors where AI is rapidly being adopted, to establish a practical, adaptable set of principles, based on a memorable acronym:

T ransparency

R ights & Fairness

U niversal Access & Inclusion

S ustainability & Safety

T esting & Security

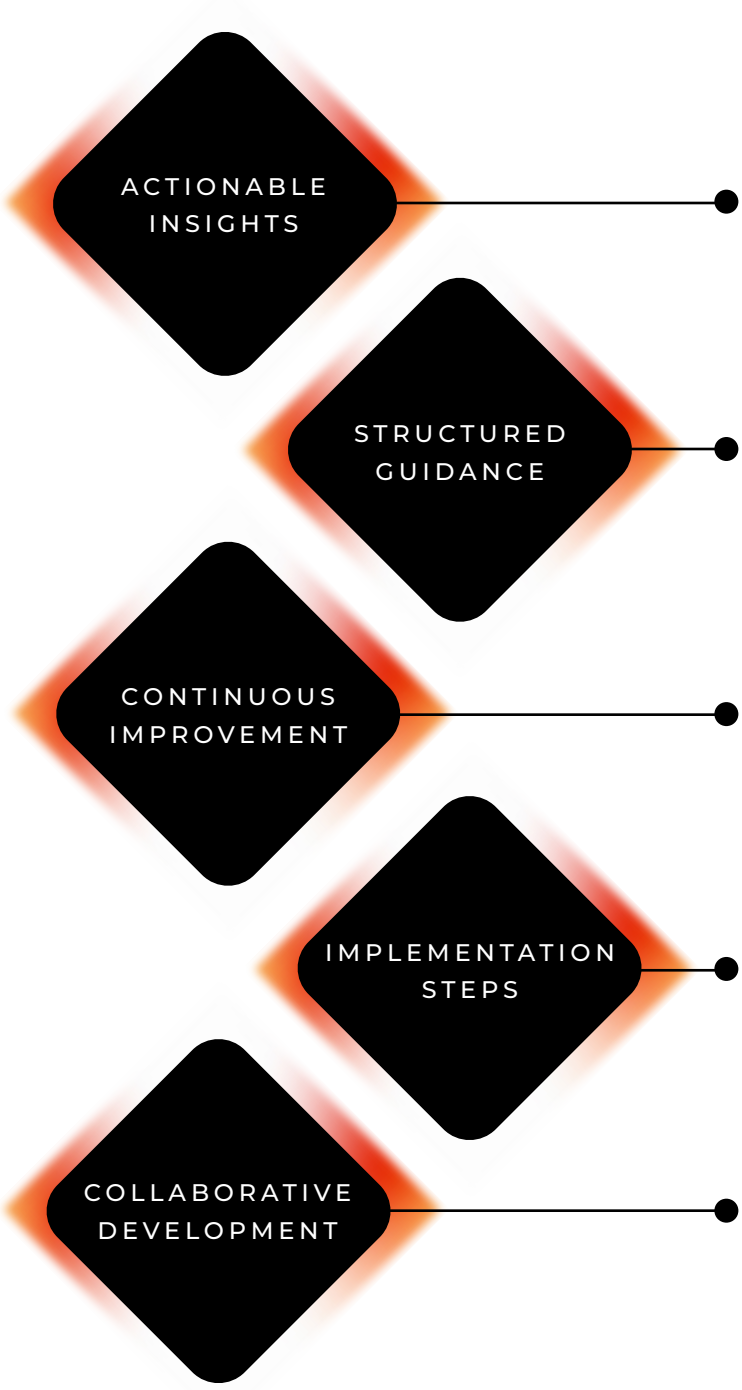
E thical Governance & Accountability

D ata Integrity & Privacy

How To Use This Framework

The **TRUSTED Framework** is designed as a practical, evolving guide for ethical AI governance. Each letter stands for a core principle, and for every pillar you will find:

- A clear **definition** of the principle
- Why it **matters** in practice
- Real-world situations to **avoid**
- Practical steps for **implementation and monitoring**



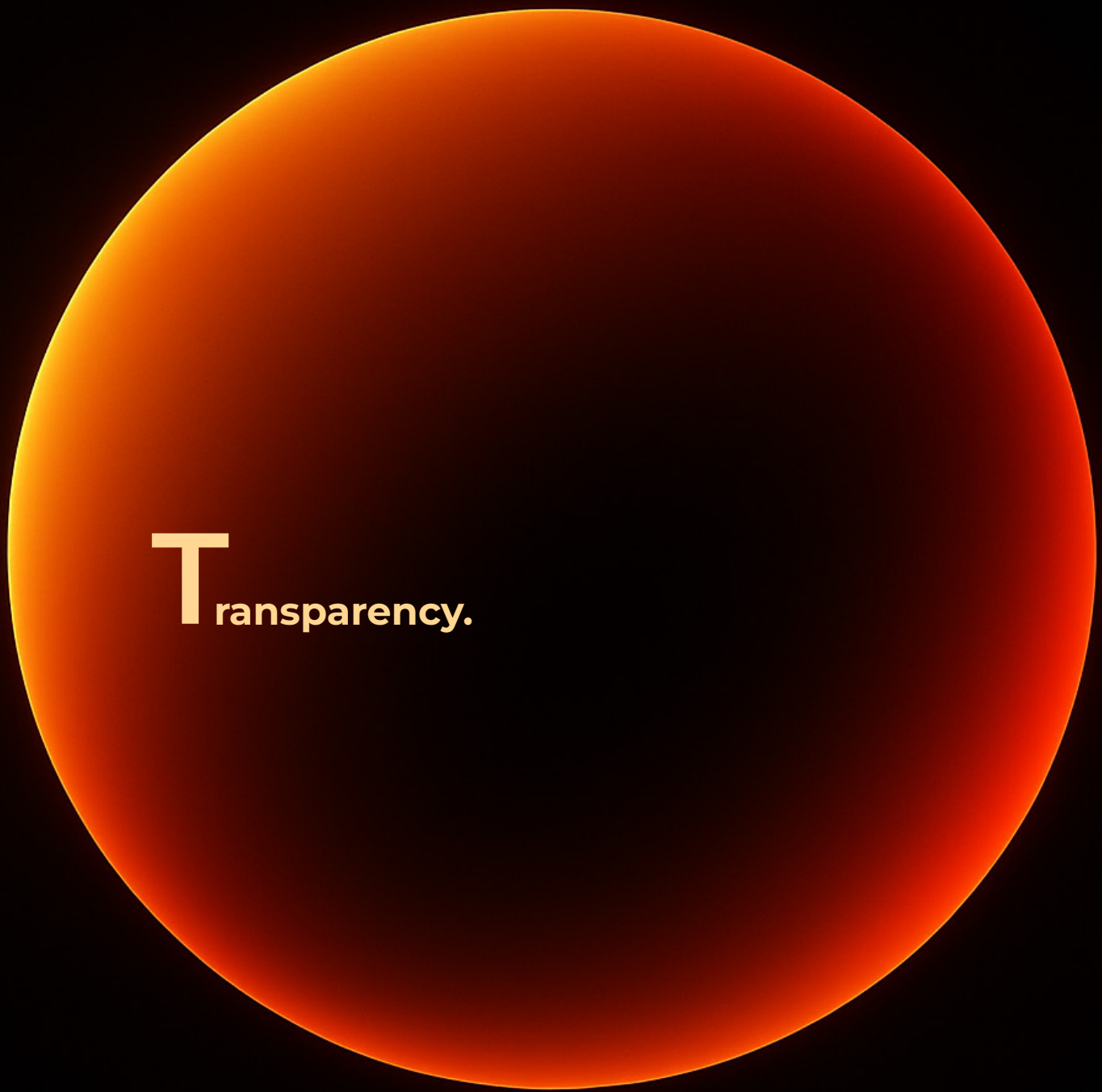
Real-world pitfalls are highlighted so you can pro-actively avoid common mistakes and risks.

Each section is organized to help you quickly understand the core idea, its practical significance, and how to apply it in your context.

Treat this framework as a living resource: regularly revisit and update your practices as technology, regulation, and societal expectations evolve

Practical recommendations are included to support integration and ongoing monitoring of each principle in your organization

Your feedback, experience, and insights are essential. Please share them to ensure the framework remains relevant, effective, and responsive to emerging challenges.



WHAT DOES TRANSPARENCY MEAN?

Transparency : Making the design, operation, and decision-making processes of AI systems reasonably clear, accessible, and understandable to all human stakeholders in a manner proportionate to the human impact of the decision.

Why is transparency essential?

Transparency is fundamental for building trust between organizations and those affected by AI-driven outcomes.

Transparency empowers stakeholders to verify the fairness and accuracy of AI decisions by identifying potential biases, errors, or unintended consequences. This openness is particularly important in sensitive areas such as recruitment, performance management, and service delivery, where automated decisions can have significant effects on human lives.

Transparency is also critical during process transformation, allowing organizations to map out which parts of a business process can be automated and which should remain human-driven. By documenting and communicating these decisions before AI is implemented, organizations ensure that stakeholders are informed and can provide feedback early in the process.

Transparent communication enables organizations to audit roles likely to be impacted by AI, anticipate workforce changes, and proactively support or retrain employees at risk of displacement. This approach not only supports ethical implementation but also helps maintain morale and organizational stability and can highlight opportunities for new roles to emerge.

> Transparency further supports ethical AI adoption by enabling continuous improvement. Organizations can incorporate feedback, address concerns, and refine their systems over time. It also encourages collaboration and knowledge sharing within organizations, fostering a culture of responsibility, ethical innovation, and leadership engagement. Providing leaders with the resources and training to drive AI implementation transparently is essential for successful change management.

Without the addition of adequate transparency measures, AI systems almost invariably make decisions opaquely as “black boxes,” where neither affected people nor the organizations themselves fully understand how decisions are made. This lack of understanding can lead to confusion, frustration, and mistrust. It also makes it difficult to identify and correct biases or errors, which can result in unfair treatment, reputational damage, and legal risks.

Transparency is both a best practice and a legal/regulatory requirement in many jurisdictions. For example, the EU AI Act mandates that organizations provide clear information about AI system design, data sources, and decision-making processes, with heightened requirements for high-risk systems. Similarly, emerging US state laws, such as the California AI Transparency Act, require detailed disclosures about AI-generated content and the systems used to create it. These regulations reflect a global trend toward greater accountability and openness in AI governance.

Practical situations to avoid

UNCLEAR APPLICANT FILTERING (“BLACK BOXES”)

Applicants are screened by an AI tool, but neither they nor staff understand the reasons for rejections, leading to confusion and an inability to explain the process.

LACK OF EXPLANATION FOR AUTOMATED SCORES

Individuals such as employees receive automated scores or ratings, but there is no information on what factors influenced those ratings, causing frustration and mistrust.

VAGUE DENIAL OF SERVICES OR CLAIMS

People are denied services or claims by an AI system, and only vague or generic reasons are provided, making it difficult to understand or challenge the outcome.

UNDISCLOSED DATA PROCESSING

Data is used by an AI system without clear communication about what data is collected, how it is processed, or its impact on results.

VAGUE DENIAL OF SERVICES OR CLAIMS

People are denied services or claims by an AI system, and only vague or generic reasons are provided, making it difficult to understand or challenge the outcome.

UNADDRESSED FEEDBACK OR CONCERNS

There are no clear processes for individuals to report concerns or request explanations about AI-driven decisions, leading to unresolved issues and increased risk.

OPAQUE SCORING SYSTEMS

Scoring or ranking systems operate as “black boxes,” preventing those affected from understanding how scores are calculated and undermining trust.

How to implement and monitor transparency

Conduct Pre-Implementation Process Transformation Assessment

01

Before implementing AI, analyse and map existing business processes to identify which parts can be automated and which should remain human-driven. Document the rationale for these decisions and communicate them to stakeholders to ensure clarity and buy-in.

Audit Workforce Impact and Roles

02

Compile an audit of roles likely to be impacted by AI implementation, including those that may be reduced, changed, or newly created. Use this audit to anticipate workforce shifts and support planning.

Build in Support and Retraining for At-Risk Staff

03

Develop and implement a plan to support and retrain employees whose roles are affected by AI. Ensure retraining and upskilling opportunities are communicated clearly and made accessible before and during AI deployment.

Provide Leadership Support for AI Implementation

04

Establish resources and training for leaders to help them drive AI adoption, communicate transparently about changes, address concerns, and foster a culture of openness and responsibility.

Adopt and Integrate Explainable AI Tools

05

Select and deploy tools that provide clear, understandable explanations for AI-driven decisions. These explanations should be accessible to both technical and non-technical stakeholders.

Communicate AI Use and Rationale to Users

06

Clearly inform users when they are interacting with AI systems. Explain the purpose, data sources, and limitations of AI decisions. Where possible, offer opt-out options and ensure explanations are tailored to the audience.

Audit Workforce Impact and Roles

07

Regularly create and publish transparency reports that detail AI use cases, system changes, and their impacts. Make these reports available to relevant stakeholders and consider using different formats for technical and non-technical audiences.

Clarify Data Handling Practices

08

Explicitly communicate how data is collected, used, stored, and processed. Reassure users that their data is only used for its intended purpose and is protected according to ethical and regulatory standards. Include details about data provenance, quality, and any third-party data sources.

Establish Incident Reporting and Review Processes

09

Set up clear internal channels for reporting and investigating incidents or complaints related to AI. Ensure these processes are well-documented, responsive, and that outcomes are communicated to affected parties.

Conduct Regular Proactive Reviews

010

Require periodic reviews of AI systems to ensure ongoing compliance with evolving technology and regulations. Reviews should assess bias, fairness, and the effectiveness of transparency measures.

Promote Cross-Team Knowledge Sharing

011

Foster a culture of transparency by sharing updates, findings, and best practices across departments and teams. Encourage collaboration between technical, legal, and operational teams to support continuous learning and improvement.

Develop Sector-Specific Disclosure Templates

012

Create templates that explain data sources, model logic, and bias mitigation steps for each sector. Tailor these disclosures to the specific context and risks of each sector.

Mandate Transparency Obligations for High-Impact Decisions

013

For organizations making significant decisions with AI, establish mandatory transparency obligations. Proactively publish information on how algorithms are used and the steps taken to ensure fair treatment.

Engage Stakeholders and the Public Early

014

Involve stakeholders and the public early in the project lifecycle. Use their feedback to shape transparency practices and build trust.

Provide Training and Support

015

Offer training and support to teams on transparency best practices and the use of transparency tools. Ensure staff are equipped to communicate about AI systems and their implications.

Align with Emerging Standards

016

Ensure transparency practices align with evolving regulatory and industry standards, updating them as standards change.

Rights & Fairness.

WHAT DOES RIGHTS MEAN?

Rights : Upholding individuals' legal and human rights, such as privacy, non-discrimination and equal treatment, as well as their due process rights to contest or appeal decisions and make their voices heard.

Why is respect for rights essential?

Respect for rights in AI is crucial to ensure that all individuals are treated equally and justly by automated systems. Without these protections, AI can unintentionally reinforce or amplify existing biases, leading to unfair or discriminatory outcomes that harm individuals and undermine trust in technology and the organization.

Practically, this means organizations must design and operate AI systems so they do not disadvantage people based on personal characteristics and must give users clear ways to question or appeal decisions that affect them.

By embedding respect for legal and human rights into AI processes, organizations reduce the risk of unintended harm, maintain credibility, and ensure that their technology benefits everyone fairly.

Practical situations to avoid

BIAS IN CANDIDATE SELECTION

An AI recruitment system, trained on historical hiring data, repeatedly favours candidates from a specific demographic group. This happens because the training data reflects past biases or imbalances in hiring. As a result, highly qualified applicants from outside that demographic group are systematically overlooked, limiting diversity and potentially violating equal opportunity principles.

DISCRIMINATORY PROMOTION RECOMMENDATIONS

An internal AI tool analyses employee performance and background but tends to recommend promotions more often for staff who attended certain universities. This overlooks equally capable employees from less represented or less prestigious institutions, reinforcing existing hierarchies and reducing opportunities for a broader range of talent.

UNFAIR ACCESS TO OPPORTUNITIES

An AI system assigns high-value projects or training opportunities based on patterns in previous selections, inadvertently favouring individuals who fit a particular profile. Employees who do not fit this profile, despite their skills or potential, are excluded from important career development chances.

OPAQUE PERFORMANCE EVALUATION

Employees receive automated performance scores or feedback with no explanation of the criteria or data used. Those who receive lower ratings have no way to understand or challenge the results, leading to frustration and a sense of unfairness.

EXCLUSION IN CUSTOMER SERVICE

An AI-driven customer support system prioritizes responses or offers to clients based on spending history or demographics, unintentionally excluding others from accessing services or resources they qualify for, simply because they do not match the preferred customer profile.

LACK OF REDRESS MECHANISMS

Individuals affected by AI-driven decisions have no clear way to request explanations, challenge outcomes, or seek corrections, leaving grievances unresolved and potentially damaging trust in the organization.

Project Initiation and Planning

Project Initiation and Planning 01

- Integrate risk and fairness assessments: Conduct comprehensive risk and fairness assessments at the outset of every AI project, rather than retroactively
- Establish common standards: Define and implement organization-wide standards, policies, and practices for responsible AI use.
- Clarify AI tool distinctions: Clearly distinguish between different AI tools (e.g., ChatGPT vs. Copilot) and document their specific fairness implications and use cases.

Development and Deployment 02

- Test for bias systematically: Before and after deployment, audit and balance training data, apply fairness metrics and mitigation techniques, use dedicated bias detection tools, and continuously monitor models.
- Regularly review outcomes: Schedule ongoing reviews of AI-driven outcomes to ensure fairness and identify any emerging issues.

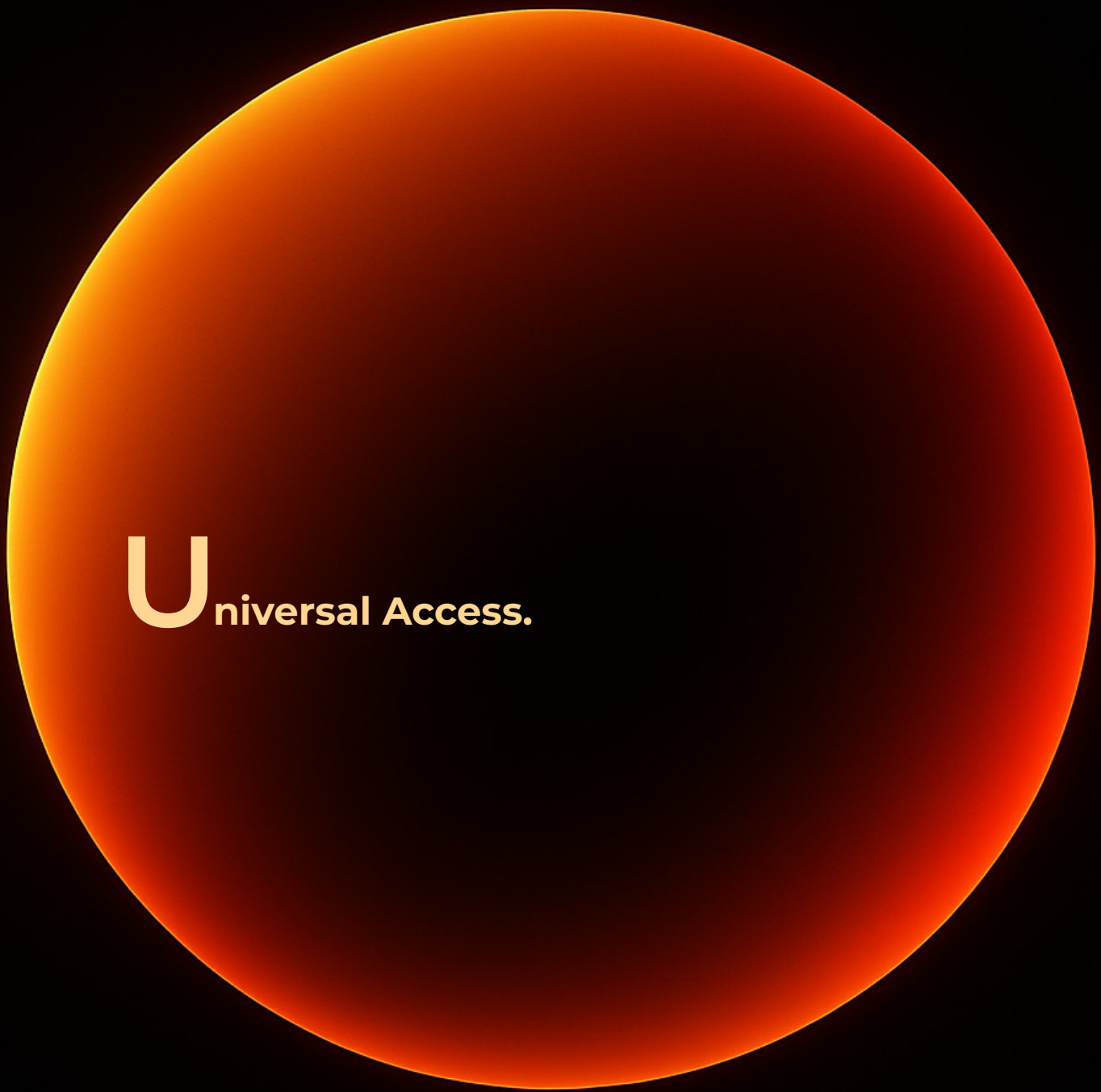
Operational Oversight and Human Involvement 03

- Guard against automation bias: Maintain human oversight throughout AI processes, ensuring that ultimate decision-making authority remains with responsible individuals.
- Train staff and managers: Invest in regular training and awareness programs for all staff, focusing on AI risks, benefits, and responsible use. Ensure recruiters and hiring managers are specifically trained to manage the candidate experience alongside AI tools.

- Promote AI literacy: Foster organization-wide AI literacy by providing education and resources that help employees understand how AI systems work, their limitations, and their role in supporting ethical and effective outcomes.

Project Initiation and Planning 04

- Create reporting channels: Develop clear, accessible processes for employees and users to report concerns or suspected biases related to AI systems.
- Provide appeal mechanisms: Establish transparent channels for appealing AI-driven decisions, ensuring recourse and accountability for those affected.



WHAT DOES UNIVERSAL ACCESS MEAN?

Universal access : The principle that artificial intelligence systems should be designed, developed, and deployed to ensure that all individuals (regardless of background, ability, socioeconomic status, or other characteristics) can use, benefit from, and participate in relevant AI-driven opportunities and services.

What does universal access mean?

Universal access is needed practically because they ensure that AI systems are usable and beneficial for all people, regardless of their abilities, backgrounds, or circumstances.

Without these considerations, AI technologies can unintentionally exclude certain groups (such as people with disabilities, those who speak different languages, or individuals from varied socioeconomic backgrounds) limiting their opportunities and reinforcing existing inequalities.

By designing AI with universal access and inclusivity in mind, organizations can create tools that work for a broader audience, reduce the risk of unintentional discrimination, and foster trust and engagement among all users. This approach also helps organizations meet ethical expectations and build more resilient, innovative, and widely accepted AI solutions.

Practical situations to avoid

INACCESSIBLE RECRUITMENT TOOLS

An online AI assessment platform requires mouse navigation and relies on visual cues, making it unusable for visually impaired candidates or those who depend on keyboard navigation or screen readers. This creates an immediate barrier for many applicants with disabilities, excluding them from fair consideration.

LANGUAGE EXCLUSION

An AI-powered chatbot or application interface only supports English, preventing non-native speakers or individuals who communicate in other languages from fully engaging with the process. This can disadvantage applicants from diverse linguistic backgrounds and limit their access to opportunities.

UNREPRESENTATIVE TRAINING DATA

AI systems are built on datasets that do not reflect the full diversity of users, including people with disabilities or from different cultural backgrounds. As a result, the AI may not recognize or respond appropriately to varied needs, leading to biased or inaccurate outcomes.

LACK OF ALTERNATIVE FORMATS

AI-driven tools fail to provide alternative formats for content, such as audio descriptions, captions, or transcripts. This prevents people with sensory disabilities from accessing important information or participating fully in digital interactions.

AUTOMATED INTERVIEWS WITH ABLEIST CRITERIA

AI-powered video interviews assess candidates based on “normative” behaviours, such as eye contact or speech patterns, that may be difficult or impossible for some candidates with disabilities. This can unfairly penalize applicants who do not fit these narrow standards.

EXCLUSION FROM DESIGN AND TESTING

People with disabilities or from underrepresented groups are not involved in the design or testing phases of AI tools. This leads to products that do not account for their needs and experiences, reinforcing existing barriers.

TOKENISTIC INCLUSION EFFORTS

Organizations include individuals from marginalized groups in development processes only superficially, without genuinely integrating their feedback or addressing their concerns. This results in AI systems that appear inclusive but fail to deliver real accessibility or equity.

ONE-SIZE-FITS-ALL USER INTERFACES

AI tools are designed with a single user experience in mind, ignoring the need for customization or accommodations. This approach excludes users who require different ways of interacting with technology.

FAILURE TO PROVIDE ADJUSTMENTS ON REQUEST

Organizations do not offer alternative routes or reasonable adjustments for candidates or users who cannot use standard AI-driven processes, despite being required by law or ethical standards in many contexts.

IGNORING CULTURAL AND CONTEXTUAL DIFFERENCES

AI systems are developed without considering the cultural, social, or contextual differences of their users, leading to misunderstandings, miscommunication, or exclusion of certain groups.

How to implement and monitor universal access

Embed Inclusive Design from the Start

01

Mandate inclusive design principles at the beginning of every AI project, ensuring that both internal and external tools are accessible to all user groups, including those with disabilities or from diverse backgrounds, as the system scales.

Design Accessible Interfaces

02

Prioritize accessibility in user interface design, making sure that digital tools are usable with assistive technologies, support multiple languages, and accommodate different interaction styles.

Test with Diverse User Groups

03

Employ diverse user testing methodologies, such as persona-based simulations with culturally varied testers, adversarial input testing, and fairness metrics (e.g., demographic parity). Continuously audit data representation and model decisions to identify and address biases.

Tailor Training and Support

04

Provide role-specific and language-appropriate training for all users, especially those less familiar with AI. Develop organizational guidance and training programs to support less-engaged or less-technically adept users.

Gather Continuous Feedback and Iterate

05

Establish channels for ongoing user feedback. Regularly review and incorporate this feedback to improve accessibility, usability, and fairness in AI systems.

Proactively Address Inclusivity

06

Consider inclusivity proactively before AI systems become too entrenched to change, ensuring that new features and updates maintain or improve accessibility and equity.

Incorporate Ongoing Legal Review

07

Recognize the lag between technological advancement and legislation. Include ongoing legal review and adaptation to ensure AI systems remain compliant with evolving laws and ethical standards.

Audit Third-Party Tools

08

Ensure procurement processes for third-party AI tools require access to audit data, transparency, and compliance documentation, so that inclusivity and accessibility are maintained across all technologies used by the organization.

Sustainability & Safety.

WHAT DOES TESTING AND SECURITY MEAN?

Sustainability : Developing, deploying, and operating artificial intelligence systems in ways that minimize negative environmental impacts, such as energy consumption, carbon emissions, and electronic waste, while also considering social equity and long-term societal well-being.

Safety : Designing and operating artificial intelligence systems so that they perform their intended functions without causing harm to people or the environment.

What does sustainability and safety mean?

Sustainability : Developing, deploying, and operating artificial intelligence systems in ways that minimize negative environmental impacts, such as energy consumption, carbon emissions, and electronic waste, while also considering social equity and long-term societal well-being.

Sustainability involves making conscious choices about resource use, data centre efficiency, and the broader lifecycle of AI technologies to ensure that AI contributes positively to ecological health and social justice and does not compromise the ability of future generations to meet their own needs.

Safety : Designing and operating artificial intelligence systems so that they perform their intended functions without causing harm to people or the environment.

Safety includes preventing accidents, misuse, or unintended negative consequences, and ensuring that AI systems are reliable, robust, and aligned with human values and ethical standards. Safety measures involve rigorous testing, risk assessment, and human oversight to protect against both immediate and long-term risks associated with AI deployment.

Why is sustainability & safety essential?

Sustainability is essential because AI technologies can have significant environmental impacts (high energy consumption, increased carbon emissions, and electronic waste) if not managed responsibly.

By prioritizing sustainability, organizations can reduce their ecological footprint, conserve resources, and contribute positively to long-term environmental health. Sustainable AI practices help ensure that technological progress does not come at the expense of the planet or future generations. They also support compliance with emerging regulations and align with growing expectations from investors, customers, and the public for responsible innovation.

Safety is crucial because AI systems, if not carefully designed and monitored, can cause harm to people, organizations, or the environment. Ensuring safety means preventing accidents, misuse, and unintended negative consequences, which helps protect users, maintain trust, and uphold ethical standards. Without a strong focus on safety, AI could lead to serious risks, including financial loss, reputational damage, or even physical harm.

Together, sustainability & safety help organizations build AI systems that are not only effective and innovative, but also responsible, trustworthy, and aligned with broader societal values.

Practical situations to avoid

WASTING RESOURCES

HR deploys a complex, energy-hungry AI system for simple tasks like scheduling interviews, unnecessarily increasing operational costs and environmental impact.

IGNORING SUPPLIER AND VENDOR IMPACTS

Organizations select AI tools from vendors without considering the sustainability or safety practices of those suppliers, potentially supporting unethical or environmentally harmful business practices.

LACK OF TRANSPARENCY IN CONTRACT CAUSES

AI-driven contract analysis or vendor management systems obscure key details or risks in supplier agreements, making it difficult for organizations to assess compliance with sustainability or safety standards.

NEGLECTING EMPLOYEE FEEDBACK

Organizations fail to consult employees or affected teams about the introduction or impact of new AI monitoring or automation tools, leading to resistance, low adoption, and missed opportunities for improvement.

OVER-RELIANCE ON PROPIETARY SOLUTIONS

Organizations depend on closed or proprietary AI systems that do not allow for independent audits, transparency, or customization, limiting their ability to ensure safety, sustainability, or compliance.

FAILURE TO UPDATE LEGACY SYSTEMS

Outdated AI tools remain in use without regular updates or security patches, exposing the organization to increased risk of data breaches, inefficiencies, or non-compliance with new regulations.

How to implement and monitor sustainability & safety

Selection & Planning 01

- Choose efficient AI solutions: Prioritize tools that align with business objectives, data readiness, privacy needs, and scalability. Evaluate expertise, cost, integration fit, and user experience to ensure measurable ROI and strategic value.
- Assess enviromental impact upfront: Explicitly address energy consumption, data centre sustainability, and potential noise pollution in procurement decisions.

Pre-Deployment Assessment 02

- Evaluate human and operational impacts: Assess how AI tools might affect employee health, morale, and workflows, especially tools used for monitoring. Inform staff transparently about AI's role and limitations.
- Require green infrastructure commitments: Select suppliers and vendors that adopt energy-efficient practices, renewable energy, and sustainable data centre designs.

Deployment & Implementation 03

- Implement environmental safeguards: Monitor energy and water usage during deployment. Integrate noise reduction strategies for data centres (e.g., soundproofing, location planning).
- Communicate transparently: Provide staff with clear information about AI's energy costs, operational impacts, and intended benefits to alleviate concerns.

Post-Deployment Monitoring

04

- Track unintended consequences: Continuously monitor AI systems for ethical, environmental, or operational risks. Adjust processes or tools as needed.
- Conduct regular sustainability audits: Require periodic assessments of energy/water use, carbon footprint, and supplier compliance with green standards.

Cultural Integration

05

- Foster AI-as-a-tool mindset: Promote transparent communication to position AI as a supportive resource, not a replacement for human judgment.
- Prioritize ongoing education: Train staff on AI’s sustainability goals, safety protocols, and their role in mitigating risks.

Testing & Security.

WHAT DOES TESTING AND SECURITY MEAN?

Testing: Testing, evaluation, validation, and verification (TEVV) of AI models and systems before and after deployment.

What does testing and security mean?

Testing: Testing, evaluation, validation, and verification (TEVV) of AI models and systems before and after deployment.

This includes validating the accuracy, reliability, and robustness of AI outputs across a range of scenarios and inputs. Testing also checks for vulnerabilities, unintended behaviours, and biases, using techniques such as adversarial testing, anomaly detection, and continuous monitoring throughout the AI lifecycle. The goal is to identify and rectify issues early, ensuring that AI systems perform safely and effectively in real-world conditions.

Security: Protecting AI systems from unauthorized access, manipulation, or misuse.

This includes safeguarding data, models, and infrastructure from cyber threats, ensuring the integrity and confidentiality of information, and implementing best practices such as the principle of least privilege, secure deployment, and ongoing threat monitoring. Security measures also encompass regular updates, patches, and incident response plans to address emerging vulnerabilities and maintain trust in AI-driven processes.

Why is testing & security essential?

Testing (TEVV) is essential because it provides a structured framework to rigorously assess AI models and applications for accuracy, robustness, and performance across diverse scenarios.

Through comprehensive TEVV processes, organizations can identify and resolve issues like vulnerabilities, biases, or unexpected behaviors before deployment and during operation. This framework ensures systems meet design specifications, align with real-world needs, and comply with regulatory standards. TEVV directly supports trust-building, regulatory compliance, and responsible AI adoption.

Security is critical because AI systems are high-value targets for cyber threats (e.g., unauthorized access, data breaches, or model manipulation). As we expect our AI systems to do more and more for us, these systems end up with far more data than any single system would have accessed in prior eras.

Integrating security into TEVV, via measures like red teaming for threat verification, vulnerability validation, secure deployment protocols, and continuous threat monitoring, protects sensitive data, maintains system integrity, and prevents misuse. Proactive security within TEVV allows organizations to rapidly adapt to emerging threats, safeguarding users and organizational reputation.

Together, TEVV and security create a unified approach to building trustworthy, resilient AI solutions. They enable systematic risk management while upholding commitments to safety, reliability, and ethical standards.

Practical situations to avoid

SKIPPING THROUGH TESTING BEFORE DEPLOYMENT

Launching AI systems without comprehensive validation and testing can result in errors, biases, or unexpected behaviours that may harm users or damage trust.

NEGLECTING REGULAR UPDATES AND PATCHES

Not keeping AI systems and their underlying infrastructure up to date exposes organizations to known vulnerabilities and security risks that are easily avoidable.

LACK OF ONGOING MONITORING

Not continuously monitoring AI systems for unusual activity or performance issues can allow problems or security threats to go undetected for long periods, or allow newly introduced issues (due to an updated, or even a change of input data or context) to go unnoticed.

INSUFFICIENT ACCESS CONTROLS

Granting excessive permissions or failing to implement strict access controls can result in bad faith actors manipulating of AI systems to gain unauthorized access to further systems.

FAILING TO DOCUMENT AND RESPOND TO INCIDENTS

Not having clear processes for reporting, investigating, and responding to security or performance incidents can delay resolution and increase potential harm.

IGNORING ADVERSARIAL TESTING

Failing to test AI models against deliberate attempts to manipulate or deceive them can leave systems vulnerable to attacks, such as data poisoning or input spoofing.

RELYING SOLELY ON AUTOMATED DECISIONS WITHOUT HUMAN OVERSIGHT

Using AI to make critical decisions without human review or accountability can amplify risks and make it harder to correct mistakes.

How to implement and monitor testing & security

Preparation & Planning 01

- Define intended use and scope: Clearly limit AI behaviour and functionality to the intended use case.
- Document and understand training data: Ensure thorough documentation and understanding of the data used to train AI models, with a focus on identifying and mitigating potential biases.

Pre-Deployment Testing & Security 02

- Comprehensive Model Evaluations: Conduct rigorous evaluations of AI models before deployment, using clear metrics for performance, robustness, and reliability. Assess functional correctness, accuracy, and how the model handles edge cases and adversarial inputs.
- Red Teaming and Adversarial Testing: Perform red teaming exercises to simulate adversarial attacks and identify vulnerabilities in AI systems. This is especially critical for foundation models and AI agents, where risks like prompt injections, jailbreaks, and data leakage can have severe consequences. Red teaming should include both internal and external experts who can probe the system for weaknesses that may not be apparent through conventional testing.
- Security Reviews and Sign-Off: Obtain formal security reviews and sign-offs before deployment, ensuring that all identified risks have been appropriately mitigated and documented.
- Automate and expand testing: Automate testing processes where possible, using robust frameworks to assess model performance and security. Expand the scope of testing as AI systems evolve to cover new threats and use cases.

Cultural Integration

03

- Limit AI to intended use: Maintain strict controls so AI operates only within its defined scope.
- Regularly update security measures: Keep security protocols current and aligned with evolving threats and best practices.
- Comply with industry standards: Adhere to recognized frameworks such as ISO 27001 for information security management.

Continuous Monitoring and Imprvement

04

- Monitor for unusual activity or errors: Continuously track system performance and security for anomalies or issues.
- Conduct continuous bias and user testing: Perform ongoing bias and user testing, recognizing that user interactions can introduce new biases over time.
- Adapt testing and mitigation strategies: Adjust testing and risk mitigation approaches as the system evolves and as new risks emerge.

Ethical Governance & Accountability.

WHAT DOES ETHICAL GOVERNANCE & ACCOUNTABILITY MEAN?

Ethical governance: Establishing clear policies, processes, and oversight to ensure AI systems are developed and used responsibly, aligning with ethical principles, laws and regulations, technical standards, voluntary commitments, AI frameworks, industry ethics, and societal values.

Accountability: Ensuring individuals or organizations are answerable for AI outcomes, with defined responsibilities and audit trails.

Why is ethical and accountability essential?

Ethical governance and accountability provide the structure and oversight needed to ensure that artificial intelligence systems operate safely, fairly, and in line with societal expectations.

Without clear governance, AI can be developed or used in ways that lead to unintended harm, bias, or misuse. Accountability ensures that there are mechanisms in place to identify who is responsible for AI decisions and actions, making it possible to address issues, correct mistakes, and maintain trust. “Demonstrable accountability” elevates this by requiring organizations to prove their compliance through evidence-based practices, not merely through theoretical commitments or policy statements. This means that responsibility for AI decisions and actions is not only assigned but also made visible and verifiable through audit trails, such as version-controlled model documentation, which trace decisions back to specific teams or individuals.

Trust is maintained not just by stating intentions but by providing concrete evidence. Demonstrable accountability is achieved through regular third-party audits of AI systems, public impact assessments that show bias mitigation efforts, and real-time monitoring dashboards for model performance. These practices offer stakeholders tangible proof of responsible AI deployment. While governance ensures legal adherence, demonstrable accountability proactively addresses risks by implementing Human-in-the-Loop checkpoints for high-stakes decisions, maintaining comprehensive AI incident registers documenting failures and their remediation, and publishing explanation reports accessible to affected stakeholders.

This approach transforms risk management from a reactive process to a proactive discipline. Together, these elements help organizations manage risks, comply with laws, and demonstrate a commitment to responsible innovation. Demonstrable accountability makes this commitment visible and verifiable, transforming ethical governance from a static framework into a living practice that builds public confidence and ensures AI achieves positive societal outcomes through transparent, auditable processes.

Practical situations to avoid

NO ACCOUNTABILITY FOR ERRORS

An AI tool makes a questionable decision about layoffs, but there is no designated person or team responsible for reviewing or overriding the decision. This leaves affected individuals without recourse and undermines trust in the organization.

IGNORING COMPLAINTS

Employees raise concerns about unfair or biased AI decisions, but HR has no established process for investigating, responding to, or correcting these issues. This can lead to unresolved grievances and a perception that the organization does not take ethics seriously.

LACK OF CLEAR GOVERNANCE STRUCTURE

There are no defined roles, committees, or oversight bodies responsible for monitoring AI use, assessing risks, or ensuring compliance with governance. This creates ambiguity about who is accountable for AI outcomes.

FAILURE TO DOCUMENT DECISIONS AND PROCESSES

AI-driven decisions are made without documentation or audit trails, making it difficult to track how decisions were reached or to hold anyone accountable if something goes wrong.

EXCLUDING STAKEHOLDERS FROM OVERSIGHT

Key stakeholders such as employees, customers, or external experts are not involved in reviewing or providing feedback on AI systems, limiting transparency and the ability to identify ethical concerns.

OVER-RELIANCE ON AUTOMATED PROCESSES

Critical decisions are made entirely by AI without human review or intervention, increasing the risk of errors, biases, or unintended consequences going unchecked.

NO REGULAR REVIEW OR UPDATE OF AI POLICIES

AI governance and accountability policies are not reviewed or updated as technology, regulations, or organizational needs evolve, leading to outdated practices that may no longer address current risks.

NEGLECTING TO TRAIN STAFF ON ETHICAL AI USE

Employees involved in AI deployment or oversight are not provided with training on ethical considerations, compliance requirements, or how to recognize and address issues as they arise.

ALLOWING CONFLICTS OF INTEREST IN OVERSIGHT

Individuals responsible for designing or deploying AI systems are also the ones tasked with evaluating their ethical impact, which can compromise objectivity and accountability.

How to implement and monitor ethics and accountability

Designate Responsibility

01

- Assign accountable individuals or teams: Clearly identify who is responsible for overseeing AI use, monitoring outcomes, and addressing issues as they arise in each department. Ideally, there ought to be a dedicated AI governance function in the company, or such function should be assumed by an existing department (legal or data protection).

Provide Training and Guidance

02

- Deliver appropriate training on AI governance: Ensure all relevant users and stakeholders understand governance processes, ethical standards, and the consequences of misuse.

Deployment & Implementation

03

- Keep comprehensive records of AI-driven decisions and actions: Maintain audit trails and documentation to support transparency and accountability.

Establish Clear Decision-Making Thresholds

04

- Define transparent thresholds for human versus AI responsibility: Specify when and how decision-making authority shifts between humans and AI, ensuring these boundaries are clear, explainable, and incorporated into formal guidelines.

Implement Oversight and Review Mechanisms

05

- Require heightened human oversight for high-stakes decisions: Ensure that humans retain final responsibility for serious or impactful decisions, with clear oversight protocols.
- Maintain company accountability, even if AI functions are outsourced: Hold the organization accountable for all AI-driven outcomes, regardless of whether certain processes are handled by third parties.

Review and Adapt Governance

06

- Regularly review and update governance processes: Periodically assess the effectiveness of governance structures and adapt them as needed to address new risks or regulatory requirements.

Embed Multi-Level Governance

07

- Ensure governance committees operate at multiple organizational levels: Establish oversight bodies at different levels of the organization to provide comprehensive, cross-functional governance and accountability.

Data integrity & Privacy.

WHAT DOES DATA INTEGRITY AND PRIVACY MEAN?

Data Integrity: The accuracy, completeness, consistency, and reliability of data throughout its lifecycle.
Privacy: The ethical and responsible handling of personal information, ensuring that individuals' data is collected, stored, processed, and used in ways that respect their rights and comply with legal standards.

What does data integrity and privacy mean?

Data Integrity under the TRUSTED acronym includes both Data Integrity and Privacy.

Data Integrity : The accuracy, completeness, consistency, and reliability of data throughout its lifecycle.

In the context of AI, this means ensuring that the data used to train and operate AI systems is free from errors, inconsistencies, and corruption. Maintaining data integrity involves processes such as validation, cleaning, monitoring, and governance to guarantee that data remains trustworthy and fit for its intended purpose. Without strong data integrity, AI models may produce unreliable or biased results, undermining business value and decision-mak-

Privacy : The ethical and responsible handling of personal information, ensuring that individuals' data is collected, stored, processed, and used in ways that respect their rights and comply with legal standards.

This includes obtaining informed consent for data use, minimizing the collection of personal data, securing data against unauthorized access, and providing transparency about how data is used. Effective privacy practices help protect individuals from harm, maintain trust, and ensure compliance with regulations such as GDPR.

Why is data integrity and privacy essential?

Data integrity and privacy are essential as the foundations for trustworthy, accurate, and reliable AI systems. Maintaining data integrity means ensuring that the information used to train and operate AI is accurate, complete, and consistent. This helps prevent errors and biases that could lead to unfair or harmful decisions, such as discriminating against certain groups or making incorrect business choices. Incomplete data can undermine decision-making as significantly as, or even more than, inaccurate data.

Protecting data privacy is just as important. Incidents like the Equifax, Yahoo, and Uber data breaches show what can happen when sensitive personal information is not properly safeguarded. These breaches often expose millions of people's data and are sometimes made worse by delayed disclosure or attempts to hide the problem, resulting in serious reputational and financial consequences for the organizations involved.

Strong privacy practices ensure that individuals' personal data is collected, stored, and used responsibly. This not only helps organizations comply with data protection laws but also builds trust with users. Without robust data integrity and privacy measures, organizations risk unreliable AI outcomes, legal penalties, and lasting damage to their reputation.

By prioritizing data integrity and privacy, organizations can ensure their AI systems are fair, reliable, and responsible, while also maintaining compliance and fostering trust among all stakeholders.

Practical situations to avoid

EXCESSIVE DATA COLLECTION

HR collects more personal information than is necessary for AI analysis, making employees uncomfortable and increasing the risk of privacy violations. This can include gathering details about employees' private lives or sensitive attributes that are irrelevant to legitimate business needs.

USING OUTDATED RECORDS

AI tools rely on old, incomplete, or incorrect employee data, leading to unfair evaluations, missed opportunities for staff, or inaccurate decision-making. For instance, performance reviews based on outdated information may disadvantage employees who have recently improved or changed roles.

OVERLY INTRUSIVE MONITORING

AI is used to monitor every keystroke, email, and digital interaction of employees, creating a stressful environment, harming morale, and potentially violating privacy expectations. Constant surveillance can make employees feel mistrusted and undermine workplace autonomy and innovation.

INADEQUATE DATA SECURITY MEASURES

Organizations fail to implement strong encryption, access controls, or regular security audits, leaving sensitive employee data vulnerable to breaches or unauthorized access. This increases the risk of data leaks and identity theft.

LACK OF CLEAR DATA RETENTION AND DELETION POLICIES

AI systems retain personal data longer than necessary or fail to properly delete information when it is no longer needed, violating data protection regulations and exposing the organization to legal and reputational risks.

POORLY ANONYMIZED OR INADEQUATELY PROTECTED DATA

AI tools are trained on datasets that are not sufficiently anonymized or protected, leading to the exposure of sensitive employee information and potential misuse.

IGNORING EMPLOYEE CONSENT AND TRANSPARENCY

Organizations use AI to process employee data without clear communication, consent, or transparency about what data is collected, how it is used, and who has access to it. This can erode trust and lead to non-compliance with data protection laws.

FAILURE TO UPDATE AND CORRECT DATA

There are no processes in place to ensure that employee data is regularly updated and corrected, resulting in AI-driven decisions based on errors or outdated information.

ALLOWING DATA TO BE USED FOR UNINTENDED PURPOSES

Employee data collected for one purpose is repurposed for unrelated analyses or decisions without proper authorization, violating privacy expectations and potentially leading to legal issues.

NEGLECTING TO INFORM EMPLOYEES ABOUT DATA USAGE

Staff are not informed when their data is used in AI systems, or are not given opportunities to review or challenge decisions made about them, undermining trust and accountability.

INCOMPLETE DATA

For example, a dataset captures only the positive attributes of option A and exclusively the negative attributes of option B, and the result is a fundamentally biased comparison. This risk is evident in practical scenarios such as performance evaluations: if project A is thoroughly documented with both positive appraisals and incident reports, while project B only records mandatory incident reports and omits recognition of positive outcomes, an individual associated with project B will appear to have received only criticism. This perception may be misleading and does not necessarily reflect their actual performance.

How to implement and monitor data integrity and privacy

Data Management and Verification 01

- Regularly update and verify data: Ensure that all data used in AI systems is current, accurate, and complete. Establish processes for ongoing data validation and correction.
- Regularly review data input protocols: Periodically assess and update data collection and entry procedures to ensure relevance and completeness.

Access Control and Security 02

- Limit access to sensitive information and intellectual property: Implement strict access controls and permissions so that only authorized personnel can view or use sensitive data.
- Use privacy-enhancing technologies and practices: Deploy encryption, anonymization, and other privacy-preserving techniques to protect data throughout its lifecycle.

Third-Party Data and External Oversight 03

- Ensure control of scraping from third-party sources and define what data can be used: Clearly define and enforce policies regarding the collection and use of third-party data, and monitor compliance with these policies.
- Consider independent third-party audits and security standards like ISO-27001: Engage external auditors to assess data practices and align with recognized security standards for ongoing assurance.

- Emphasize the importance of both included and excluded data in AI systems: Document and communicate how data selection decisions are made, and consider the impact of both the data used and the data left out.
- Provide transparency about what organizational data is used in AI models and where it is processed or stored: Clearly inform stakeholders about data usage, storage locations, and processing practices to build trust and support compliance.

About Mevitae

MeVitae's vision is to create fairness in the workplace, empowering organizations to make smarter, faster, and fairer decisions. Integrating fluidly with existing HR systems, MeVitae consolidates critical workforce data, from recruitment to payroll, into a single platform, proactively identifying human capital risks and challenges.

Leveraging AI-driven personalized interventions, MeVitae orchestrates precise, tailored strategies to pre-emptively address these challenges. These interventions (such as anonymized recruiting) are pivotal in reshaping recruitment and employee management practices.

Proven results underscore MeVitae’s efficacy: achieving up to 90% in time and cost savings. Esteemed by top-tier government bodies and leading global brands, MeVitae stands as a trusted partner in transformative workforce management.


MEVITAE