Understanding the AI Safeguarding Readiness Matrix

A Guide for Designated Safeguarding Leads

Summary

- The AI Safeguarding Readiness Matrix helps schools assess and improve their preparedness for AI-related safeguarding challenges through a four-quadrant framework
- Traditional safeguarding approaches are increasingly insufficient for addressing AI-specific risks such as deepfakes, AI-driven grooming, and algorithmic bias
- Schools should aim to move towards the 'Transforming' quadrant, where safeguarding professionals become AI-literate leaders who proactively shape policies
- The framework provides practical steps to help DSLs identify their current position, build AI literacy, develop appropriate policies, and lead transformation in their settings

Introduction

AI is transforming safeguarding in schools, presenting both risks and opportunities.

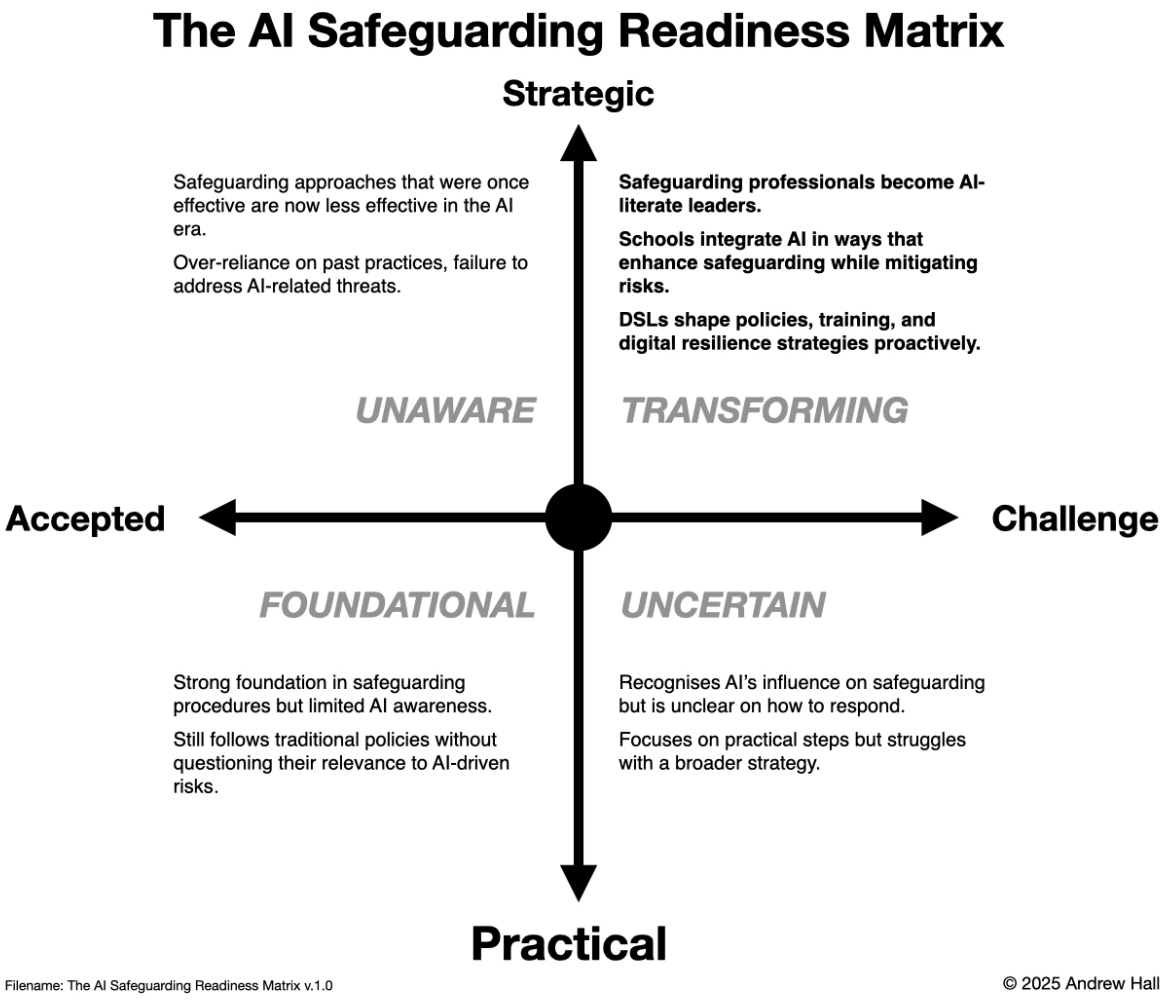
To stay ahead, safeguarding professionals must shift from simply knowing safeguarding procedures (Practical knowledge) to thinking critically about AI's impact.

This matrix helps DSLs move from expertise to transformational, AI-aware, safeguarding leadership.

The Matrix Framework

The matrix is built on two key dimensions:

- **Vertical axis (Practical → Strategic)**: Moving from procedural safeguarding knowledge to deeper, strategic understanding of AI risks and opportunities.
- **Horizontal axis (Accepted → Challenge)**: Transitioning from following traditional safeguarding norms to questioning, adapting, and innovating in an AI world.



These axes create four quadrants representing different states of AI safeguarding readiness:

Foundational (Practical/Accepted)

- Strong foundation in safeguarding procedures but limited AI awareness
- Still follows traditional policies without questioning their relevance to AI-driven risks

Uncertain (Practical/Challenge)

- Recognises AI's influence on safeguarding but is unclear on how to respond
- Focuses on practical steps but struggles with a broader strategy

Unaware (Strategic/Accepted)

- Safeguarding approaches that were once effective are now less effective in the AI era
- Over-reliance on past practices, failure to address AI-related threats

Transforming (Strategic/Challenge)

- Safeguarding professionals become AI-literate leaders
- Schools integrate AI in ways that enhance safeguarding while mitigating risks
- DSLs shape policies, training, and digital resilience strategies proactively

Applying The Matrix in Schools

Step 1: Identify Your Position

Where does your school currently sit on this framework? Are you beginning to think about AI's impact?

Step 2: Build AI Literacy in Safeguarding

Begin engaging with AI ethics, bias, and risks. Provide CPD training for DSLs and staff on AI and safeguarding.

Step 3: Develop AI-Aware Safeguarding Policies

Assess current safeguarding strategies—are they equipped to handle AI-driven risks? Introduce policies on AI monitoring, student data protection, and AI literacy for children.

Step 4: Lead Transformation in AI and Safeguarding

DSLs should advocate for AI-safe school environments. Collaborate with policymakers and edtech providers to ensure AI tools used in schools are ethical and safe.

Key Areas of AI Safeguarding Readiness

The matrix examines four key areas across each quadrant:

1. AI Awareness

Focus: Understanding AI's role and risks in safeguarding

Quadrant	Description
Foundational	Basic safeguarding procedures exist, but staff have limited understanding of AI's impact
Uncertain	Some awareness of AI concerns exists, but comprehensive responses are lacking
Unaware	Strategic approach exists but fails to recognise AI's unique challenges
Transforming	AI awareness is embedded in safeguarding culture with ongoing CPD

2. AI Integration

Focus: Embedding AI tools safely and ethically into safeguarding processes

Quadrant	Description
Foundational	Traditional tools used, but AI tools not considered
Uncertain	Some AI-powered tools used without clear policies
Unaware	Strategic processes exist but don't account for AI's unique capabilities
Transforming	AI tools ethically embedded with clear governance and accountability

3. Safeguarding Adaptation

Focus: Adjusting safeguarding strategies to address AI-generated risks

Quadrant	Description
Foundational	Traditional policies exist but don't consider AI risks
Uncertain	AI risks acknowledged with reactive responses
Unaware	Strategic frameworks treat AI challenges as extensions of traditional risks
Transforming	AI-driven risks fully integrated with proactive monitoring

4. Policy & Governance

Focus: Establishing clear policies and compliance frameworks for AI in safeguarding

Quadrant	Description
Foundational	Established policies don't address AI-specific concerns
Uncertain	AI concerns noted but formal policies still developing
Unaware	Strategic frameworks treat AI as merely another digital tool
Transforming	AI governance fully embedded with regular compliance reviews

Moving Between Quadrants: Action Steps

From Foundational to Transforming

- Audit existing safeguarding policies for AI gaps
- Invest in AI literacy training for all safeguarding staff
- Create dedicated AI response protocols for common scenarios
- Partner with tech experts to understand emerging risks

From Uncertain to Transforming

- Develop comprehensive AI safeguarding framework
- Formalise emerging policies and approaches
- Build capacity for strategic thinking beyond case-by-case responses
- Establish governance structures for AI use in school

From Unaware to Transforming

- Challenge assumptions about existing safeguarding approaches
- Review recent incidents for overlooked AI dimensions
- Update risk assessment frameworks to include AI-specific threats
- Engage with latest research on AI in education

Practical Examples

To help identify your setting's position in the matrix, consider these scenarios and their quadrant placements:

Scenario	Quadrant	Rationale
A DSL dismisses AI concerns as "just the latest fad" while claiming existing policies cover all digital risks	**Unaware**	Strategic approach that relies on outdated practices
Safeguarding team documents AI-generated inappropriate photos but treats it as standard cyberbullying	**Uncertain**	Recognises the issue but lacks comprehensive response
School organises regular "digital landscape" briefings including AI tools exploration and has updated policies with AI clauses	**Transforming**	Proactively educating staff and implementing AI-specific systems
Deputy head dismisses AI homework concerns, comparing it to Wikipedia	**Unaware**	Strategic thinking that accepts outdated comparisons
School has robust traditional reporting but is unsure how to handle AI chatbots mimicking staff	**Foundational**	Good procedures that lack AI-specific adaptation
DSL partners with tech companies for AI literacy and implements student-led "AI Ethics Council"	**Transforming**	Proactive engagement with external partners and policy development
School responds to AI impersonation by simply blocking websites	**Foundational**	Procedurally correct but relies on traditional approaches
Team categorises AI content risks as generic "inappropriate website access"	**Foundational**	Follows procedures but fails to recognise AI-specific risks

Glossary of AI Terms

Adversarial Attacks Manipulating AI by feeding it misleading data to cause errors.

Agentic AI AI that can act independently, raising concerns about oversight.

AI Detection Tools Software that identifies whether content was AI-generated.

AI Ethics Ensuring AI aligns with fairness, human rights, and moral values.

AI Hallucination When AI generates false but convincing information.

AI Literacy Understanding and critically evaluating AI and its outputs.

AI Surveillance AI-powered monitoring of online activity, raising privacy concerns.

AI-Generated Content (AIGC) Any media created by AI rather than humans.

Algorithmic Bias AI reflecting or amplifying human biases, causing unfair outcomes.

Artificial Intelligence (AI) Systems that perform tasks requiring human intelligence.

Data Scraping Automatically collecting online data, potentially without consent.

Deepfake Detection AI tools designed to identify manipulated media.

Deepfakes Synthetic media making people appear to say or do things they didn't.

Filter Bypass Using AI to circumvent school content filters or monitoring.

Generative AI AI that creates new text, images, audio, or video.

Hallucination Detection Tools that detect when AI generates false information.

Large Language Models (LLMs) AI trained on vast text data to generate writing.

Prompt Engineering Crafting instructions to guide AI responses and prevent misuse.

Regulatory Compliance Ensuring AI follows legal and safeguarding policies.

Responsible AI Developing AI with ethics, transparency, and accountability.

Synthetic Data Artificially created data mimicking real-world information.

Voice Cloning AI-generated speech mimicking a person's voice.

Self-Assessment Checklist

Use this checklist to quickly assess your school's current position in the AI Safeguarding Readiness Matrix.

AI Awareness

- ☐ Our staff can confidently explain how AI might impact safeguarding in our setting
- ☐ We have discussed AI-specific risks (e.g., deepfakes, AI chatbots) in safeguarding training
- ☐ Our DSL team stays informed about emerging AI technologies used by students
- ☐ We can identify the difference between traditional online risks and AI-enhanced risks

AI Integration

- ☐ We have clear policies governing the use of AI tools in our school
- ☐ We conduct impact assessments before implementing AI systems that affect students
- ☐ We understand the capabilities and limitations of AI systems used in our setting
- ☐ We regularly review how AI tools are being used by staff and students

Safeguarding Adaptation

- ☐ Our safeguarding policies specifically mention AI-related risks
- ☐ We have procedures for handling incidents involving AI-generated content
- ☐ Our risk assessments consider how AI might create new vulnerabilities
- ☐ Staff know how to respond to AI-specific safeguarding concerns

Policy & Governance

- ☐ We have explicit guidelines on ethical AI use in our setting
- ☐ Student data privacy policies account for AI-specific considerations
- ☐ We have a framework for evaluating new AI tools before adoption
- ☐ Leadership regularly reviews our AI safeguarding approach

Scoring

0-4 ticked: Likely in the Foundational or Unaware quadrant
5-9 ticked: Likely in the Uncertain quadrant
10-14 ticked: Moving towards the Transforming quadrant
15-16 ticked: Firmly in the Transforming quadrant

Next Steps

The AI landscape is evolving rapidly, and safeguarding practices must evolve with it. Begin by completing the self-assessment checklist with your safeguarding team, honestly evaluating where your setting currently stands. Identify one area for immediate improvement and schedule time to develop an action plan. Remember that moving toward the 'Transforming' quadrant is a journey that requires ongoing commitment, but even small steps forward significantly enhance your ability to safeguard children in an AI-enabled world.

Introduction

AI is transforming safeguarding in schools, presenting both risks and opportunities.

To stay ahead, safeguarding professionals must shift from simply knowing safeguarding procedures (**Practical knowledge**) to **thinking critically about AI's impact**.

This framework helps DSLs move from expertise to **transformational, AI-aware, safeguarding leadership**.

Applying This Framework in Schools

Step 1

Identify Your Position

Where does your school currently sit on this framework?

Are you beginning to think about AI's impact?

Step 2

Build AI Literacy in Safeguarding

Begin engaging with AI ethics, bias, and risks.

Provide CPD training for DSLs and staff on AI and safeguarding

Step 3

Develop AI-Aware Safeguarding Policies

Assess current safeguarding strategies—are they equipped to handle AI-driven risks?

Introduce policies on AI monitoring, student data protection, and AI literacy for children.

Step 4

Lead Transformation in AI and Safeguarding

DSLs should advocate for AI-safe school environments

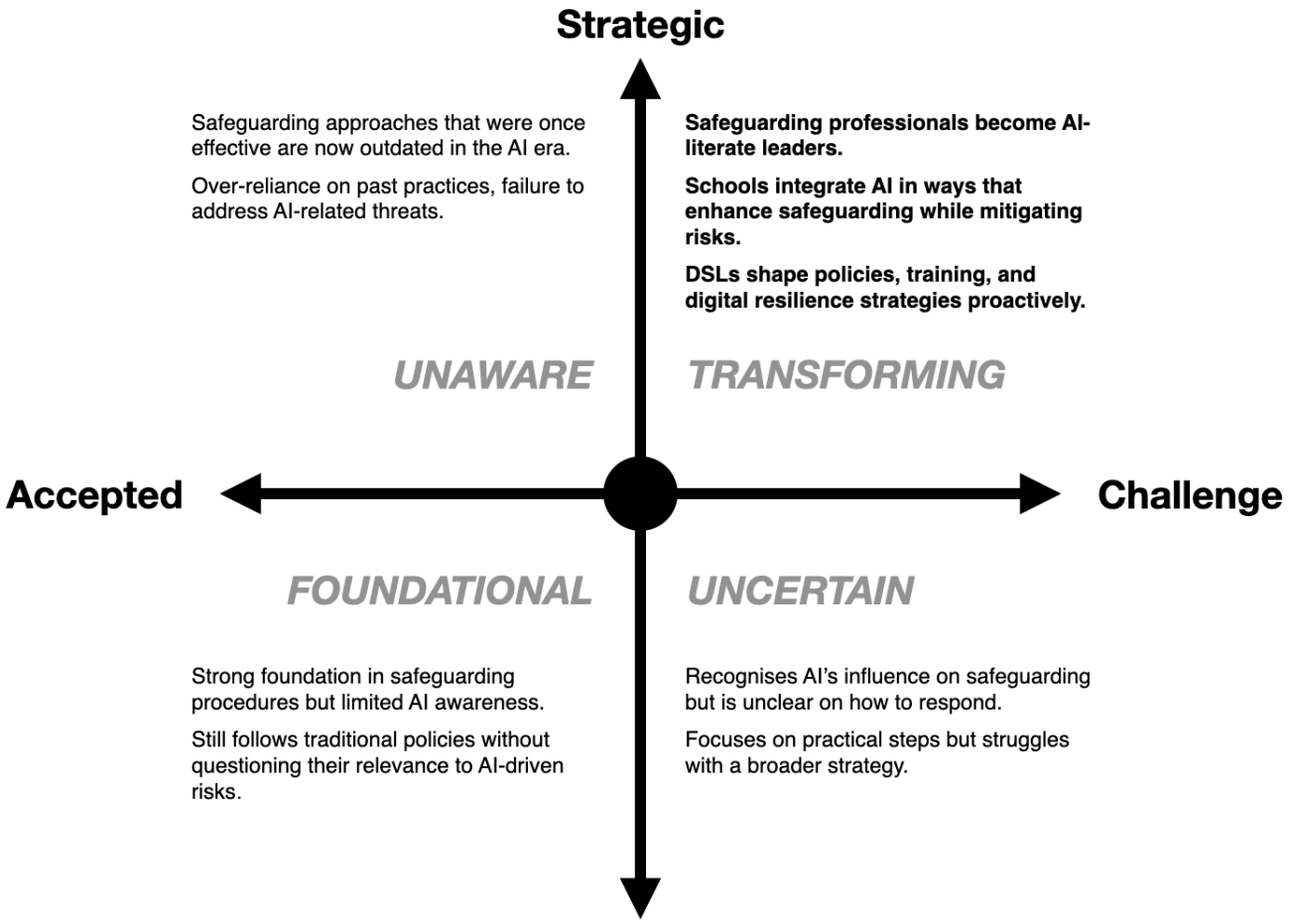
Collaborate with policymakers and edtech providers to ensure AI tools used in schools are ethical and safe.

Vertical Axis (Practical → Strategic): Moving from procedural safeguarding knowledge to deeper, strategic understanding of AI risks and opportunities.

Horizontal Axis (Accepted → Challenge): Transitioning from following traditional safeguarding norms to questioning, adapting, and innovating in an AI world.

Filename: The AI Safeguarding Readiness Matrix

The AI Safeguarding Readiness Matrix



AI Safeguarding Scenarios: Sorting Activity

Place each scenario on the attached matrix reflecting the quadrant that is most applicable.

Scenario 1 During a governors' meeting, the DSL presents a comprehensive review of online safeguarding incidents from the past term. When questioned about AI tools students are using, she responds, "We've handled internet safety for years—this is just the latest fad. Our existing policies cover all digital risks."	**Scenario 5** A secondary school has established robust reporting channels for traditional safeguarding concerns. When a teacher reports that students are creating AI chatbots mimicking staff personalities, the DSL carefully documents this but isn't confident about next steps, saying, "I understand this is concerning, but I'm not sure what policy applies here."
Scenario 2 A computing teacher notices students using AI image generators to create lifelike photos of classmates in inappropriate situations. He reports this to the safeguarding team, who document the specific incident but aren't sure how to address the broader implications beyond treating it as standard cyberbullying	**Scenario 6** The DSL at a primary school has initiated partnerships with local tech companies to provide age-appropriate AI literacy lessons. She regularly contributes to authority-wide discussions about AI safeguarding standards and has implemented a student-led "AI Ethics Council" to inform school policy development.
Scenario 3 The head of safeguarding organises monthly "digital landscape" briefings for staff covering emerging online risks. The latest session included hands-on exploration of popular AI tools, helping staff understand capabilities and limitations. The school is currently piloting an AI detection system and has updated its acceptable use policy with specific AI clauses.	**Scenario 6** A Year 11 form tutor discovers students using AI to impersonate parents in emails to attendance staff. She escalates this through proper channels, but the response from leadership is, "Let's just block the website they're using and send a warning letter home." No consideration is given to addressing the underlying capabilities that make such impersonation possible.
Scenario 4 When parents raise concerns about their children using AI homework tools, the deputy head dismisses these worries, explaining, "We've always had Wikipedia and Google—this is no different." The school's safeguarding strategy makes no mention of emerging AI technologies despite their widespread use among pupils.	**Scenario 8** The safeguarding team at a community college maintains detailed records of all incidents and follows procedural guidance meticulously. However, when staff report concerns about students accessing inappropriate content through AI tools that bypass normal filters, the team classifies these under general "inappropriate website access" categories without recognising the unique characteristics or risks involved

AI Safeguarding Scenarios: Sorting Activity Answers

1. **Unaware** (Strategic/Accepted) - The DSL dismisses AI concerns as "just the latest fad" while claiming to have a strategic approach, showing they're maintaining outdated practices despite thinking strategically.

2. **Uncertain** (Practical/Challenge) - The team recognises and documents the specific AI issue (practical) but struggles with developing a broader strategy to address the unique aspects of AI risks.

3. **Transforming** (Strategic/Challenge) - The school is proactively educating staff, updating policies specifically for AI, and implementing systems to address new challenges in a forward-thinking way.

4. **Unaware** (Strategic/Accepted) - The deputy head fails to recognis me the unique challenges of AI by comparing it to older technologies, showing strategic thinking that accepts outdated comparisons.

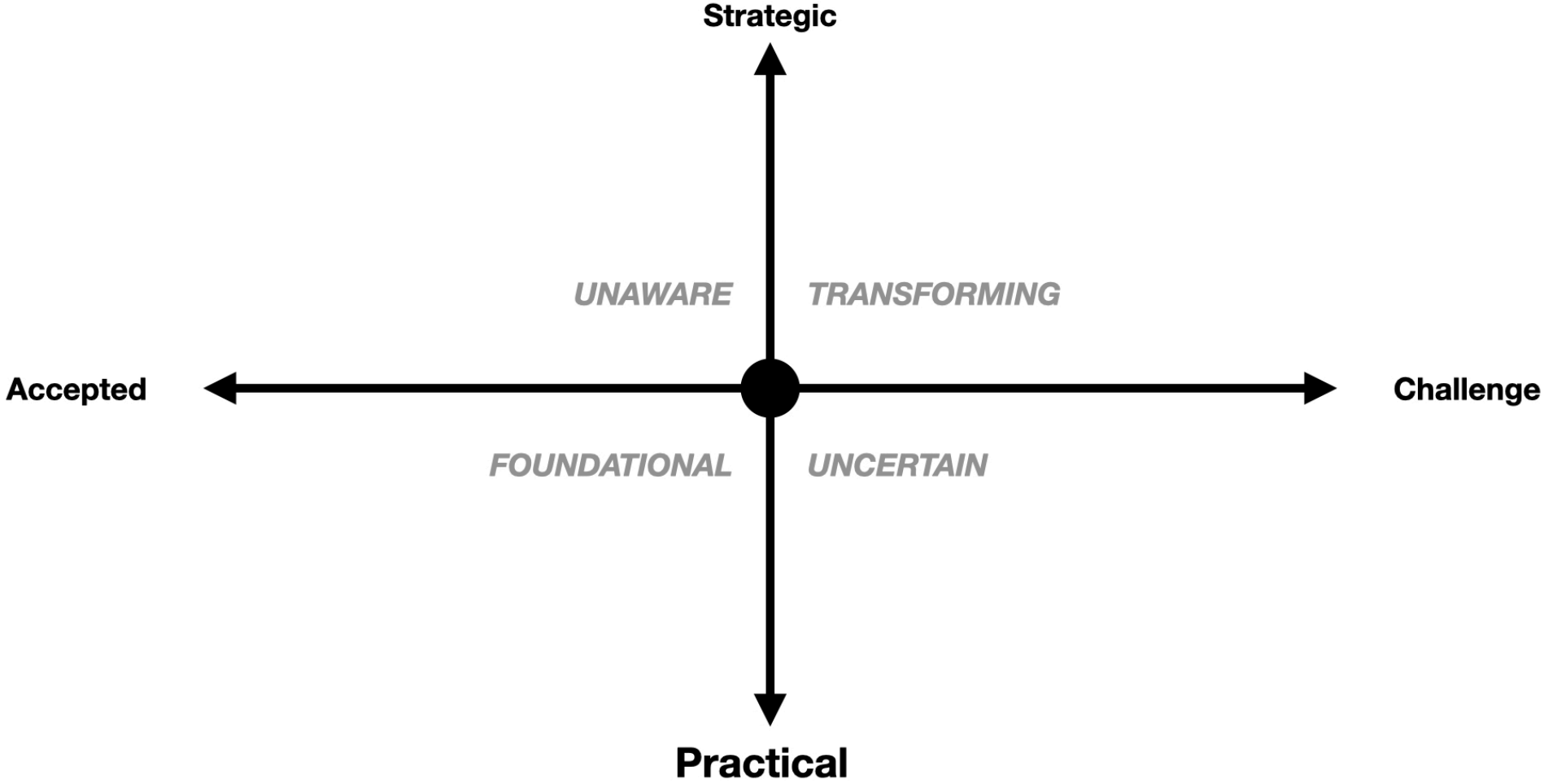
5. **Foundational** (Practical/Accepted) - The school has good traditional safeguarding procedures but lacks AI-specific understanding, following established protocols without adaptation.

6. **Transforming** (Strategic/Challenge) - The DSL works proactively with external partners, contributes to broader policy discussions, and involves students in shaping responses to AI challenges.

7. **Foundational** (Practical/Accepted) - The response is procedurally correct but relies on traditional approaches (blocking websites) without addressing the strategic implications of AI impersonation capabilities.

8. **Foundational** (Practical/Accepted) - The team meticulously follows established procedures but lacks awareness of AI-specific risks, categorizing new threats under traditional classifications.

The AI Safeguarding Readiness Matrix



Filename: The AI Safeguarding Readiness Matrix v.1.0