AI Risk Management

Week 9 – AI Risk Management

CSAI Learning Objectives

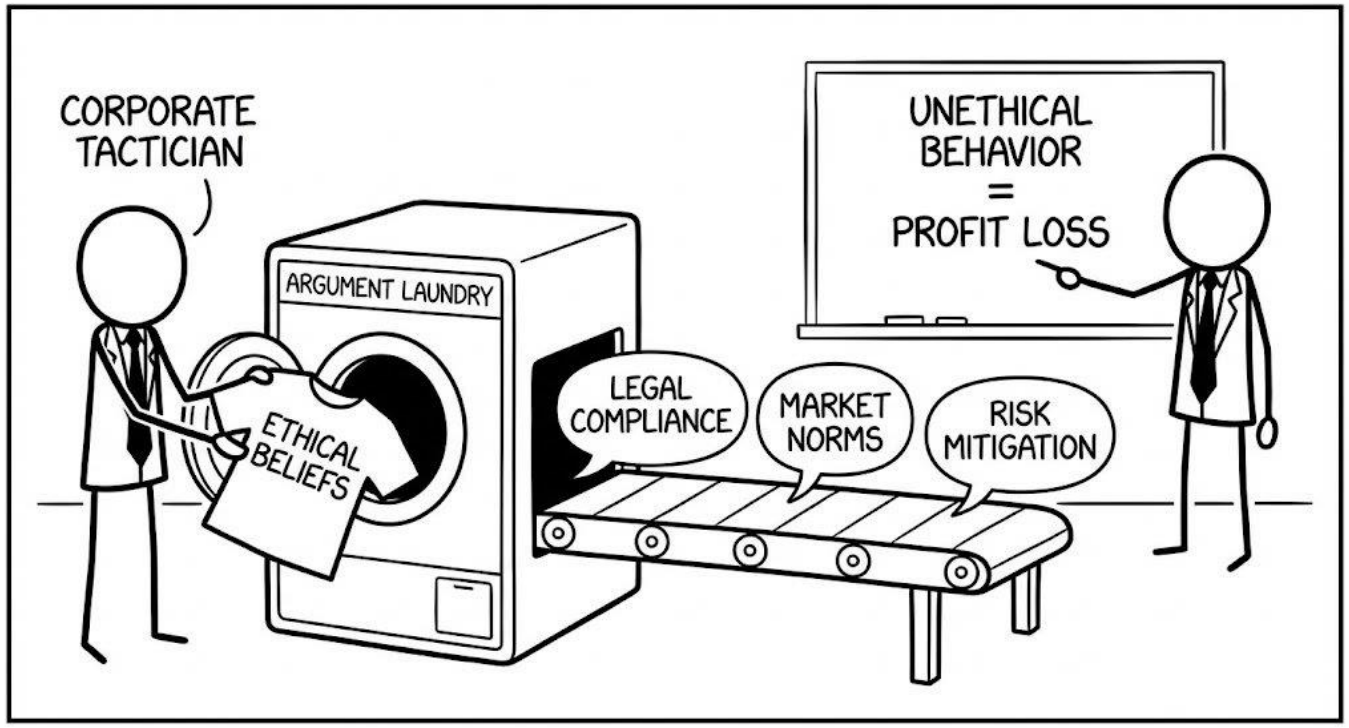
Daniel

3. Analyse case studies to **identify** and **mitigate potential risks** considering **legal**, **social**, ethical or **professional issues.**

4. Apply **ethical methodologies** in the **design of responsible AI systems**.

Week 6, **Week 7**, **Week 8**, **Week 9**, **Week 10**.

Why risk management is worth understanding

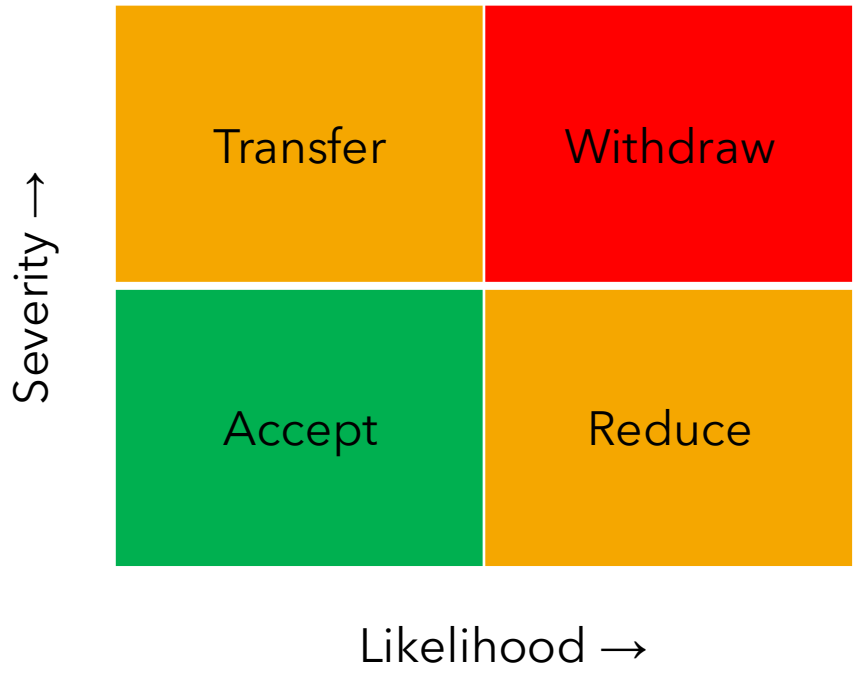


What is risk management?

A structured framework for managing the downside of technology adoption:

1. Identify potential risks
2. Assess the severity and likelihood*
3. Apply appropriate mitigations
    - Reduction
    - Mitigation
    - Transfer
    - Withdrawal
4. Monitor the risks*

*Out of scope for assessment in this course



How to Reduce AI Risk

Reduction depends on AI use case

Using AI

Employees use AI tools, typically chatbots, either personal or corporate.

Procuring AI

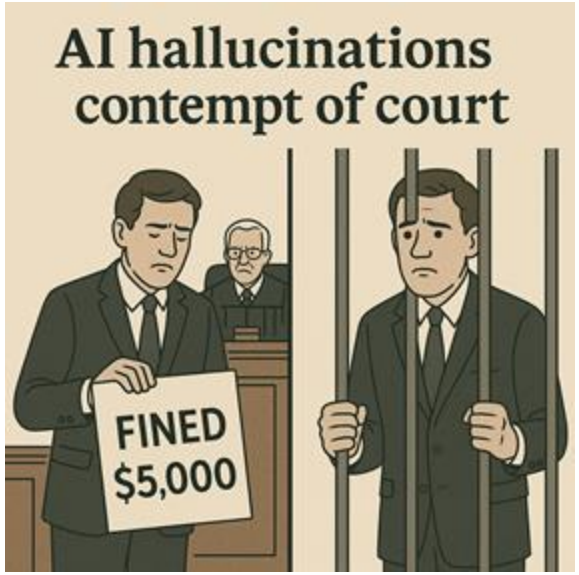
Organization buys official AI tools to be integrated into operations.

Building AI

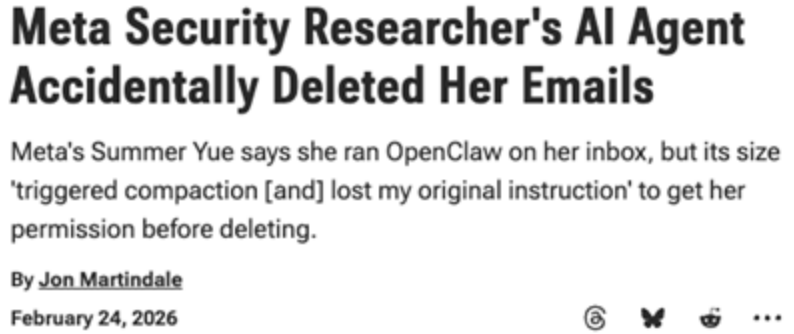
Building AI tools, either for sale or for internal processes.

Employees using AI assistants

Data privacy



Errors & omissions



Meta Security Researcher's AI Agent Accidentally Deleted Her Emails

Meta's Summer Yue says she ran OpenClaw on her inbox, but its size 'triggered compaction [and] lost my original instruction' to get her permission before deleting.

By Jon Martindale

February 24, 2026

Rogue agents

The problem of "shadow AI"

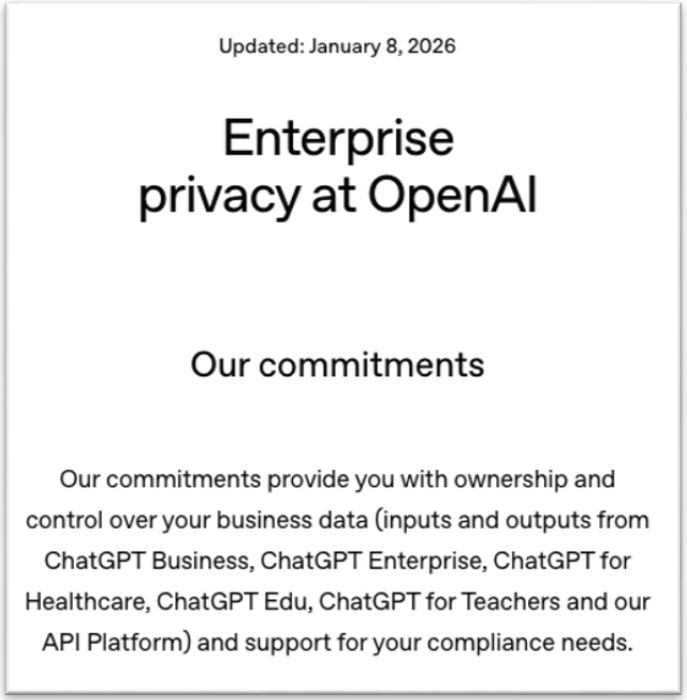
Microsoft research reveals:

- 71% of UK employees have used unapproved consumer AI tools at work, and 51% continue to do so every week.
- Use cases are diverse
  - draft and respond to workplace communications (49%)
  - draft materials at work, such as reports and presentations (40%)
  - carry out finance-related tasks (22%)

Source: https://ukstories.microsoft.com/features/rise-in-shadow-ai-tools-raising-security-concerns-for-uk/

Main mitigations for AI tool usage

- Official AI tool and ban personal tools
  - Corporate contracts agree not to train on prompts
- Set AI policy and train employees
  - Prohibit agents with write access and/or access to sensitive data
- Monitor usage of tools
  - If using managed devices
  - Trade-offs with AI



Updated: January 8, 2026

Enterprise privacy at OpenAI

Our commitments

Our commitments provide you with ownership and control over your business data (inputs and outputs from ChatGPT Business, ChatGPT Enterprise, ChatGPT for Healthcare, ChatGPT Edu, ChatGPT for Teachers and our API Platform) and support for your compliance needs.

Risks of procuring AI tools

Content Creation

Creators Launch Campaign to Counter Big Tech's Alleged AI Copyright Theft

ROSE ESFANDIARI
JANUARY 22, 2026, 10:15 AM 1

Human Resources

AI company Eightfold sued for helping companies secretly score job seekers

By Jody Godoy

January 22, 2026 12:21 AM GMT · Updated January 22, 2026

Pricing & selection

Justice Department Sues RealPage for Algorithmic Pricing Scheme that Harms Millions of American Renters

Friday, August 23, 2024

Mitigations when procuring AI tools

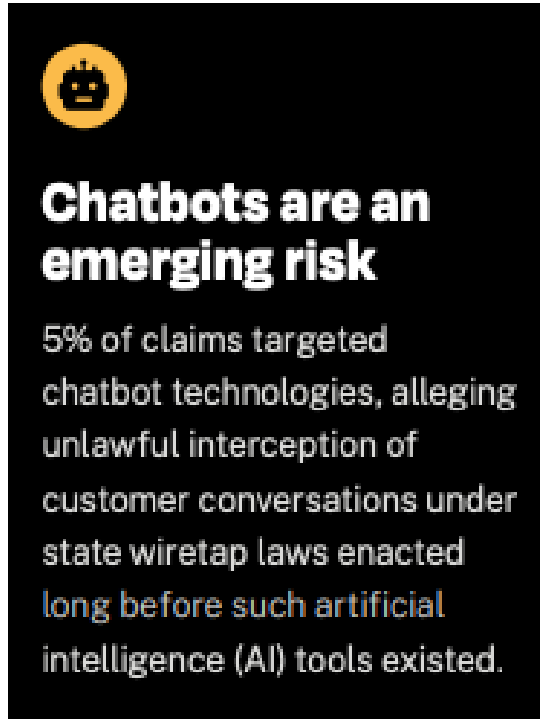
- Contractual guarantees
  - Vendor will defend intellectual property claims (OpenAI)
  - Warranty of performance/bias level, with right to audit and cancel if it drops below a level
- Disclosures and consent
- Human in the loop
- Audit and document performance

Disclosing AI to customers

Type your question for our live chat.

You are interacting with a bot. The exchange will be shared with the AI provider.

Type your question for our live chat.

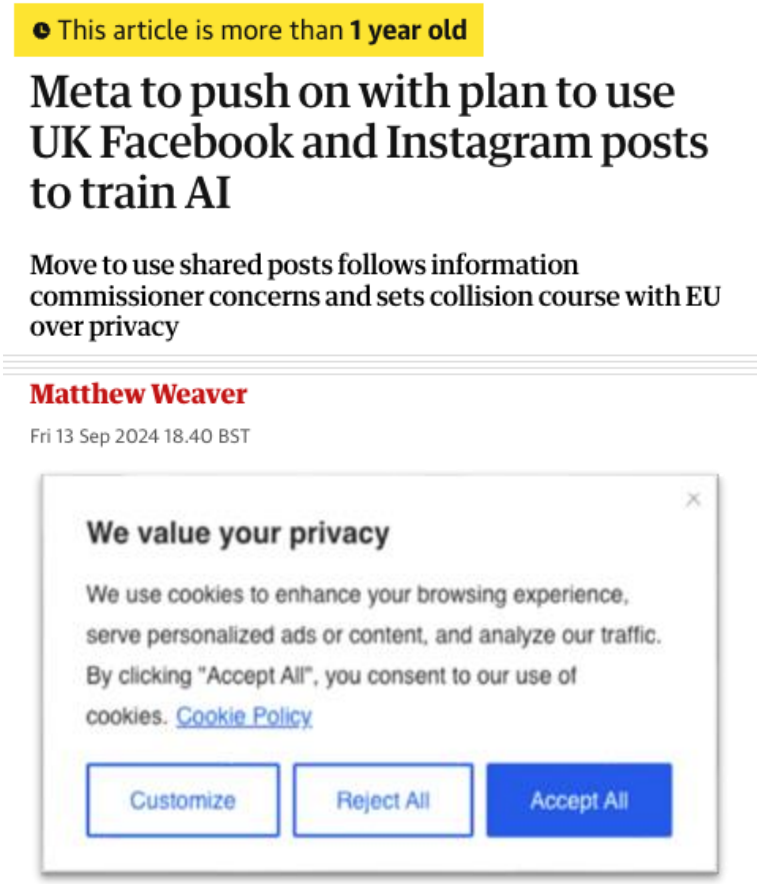


Source: https://web.coalitioninc.com/download-state-of-web-privacy-report.html

Consent for training and/or inference

- GDPR* created obligation for data processors to have legal basis for using personal data
  - Can either be opt-in consent or legitimate interest
- Hard for model vendors to collect consent
  - Open question whether the the vendor's "legitimate interests" outweigh users' privacy rights
- Is consent "freely given, specific, informed and unambiguous" anyway?

*European Union's General Data Protection Regulation (2018) or UK GDPR (2021)


This article is more than **1 year old**

Meta to push on with plan to use UK Facebook and Instagram posts to train AI

Move to use shared posts follows information commissioner concerns and sets collision course with EU over privacy

Matthew Weaver

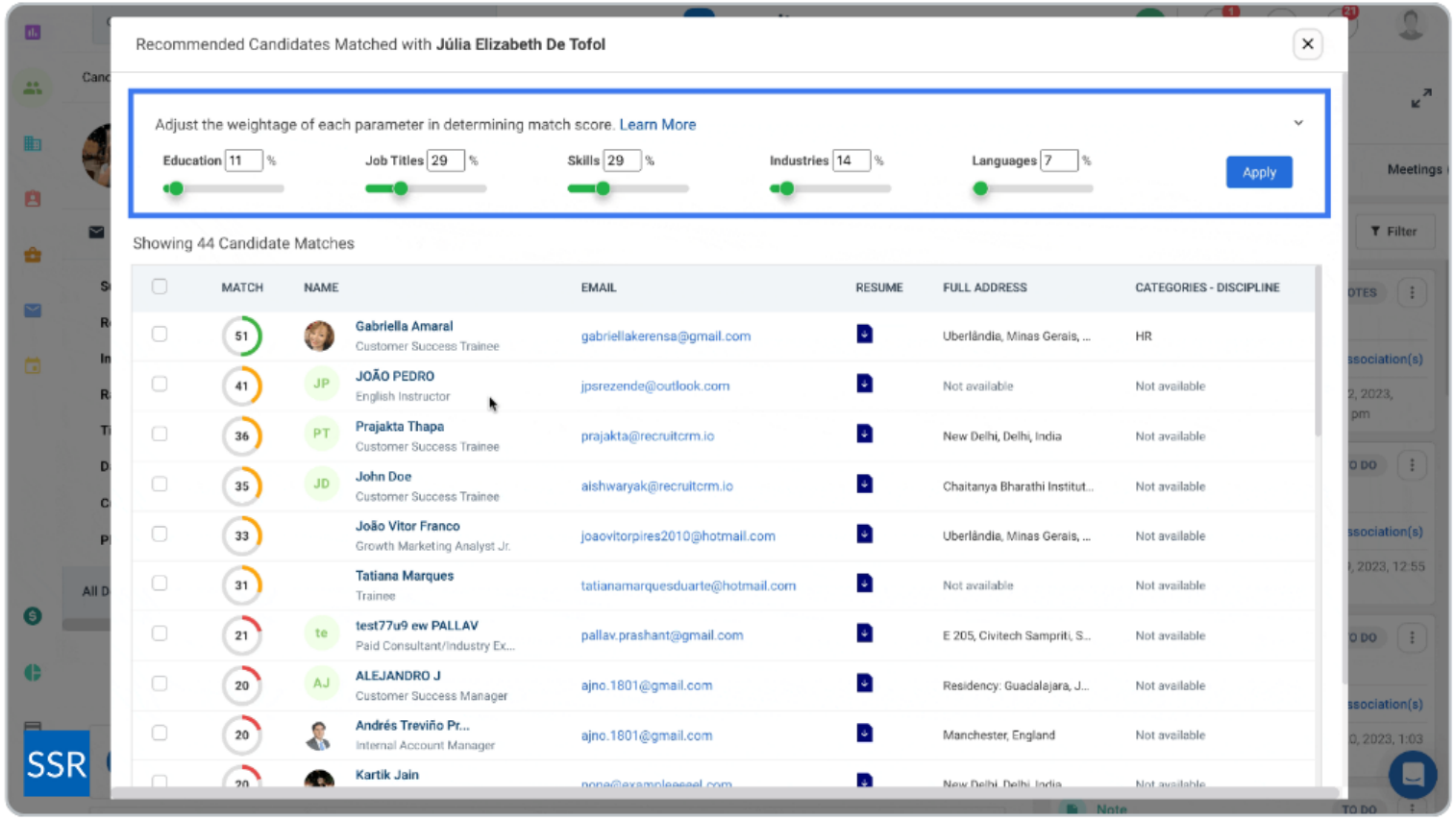
Fri 13 Sep 2024 18.40 BST

We value your privacy

We use cookies to enhance your browsing experience, serve personalized ads or content, and analyze our traffic. By clicking "Accept All", you consent to our use of cookies. Cookie Policy

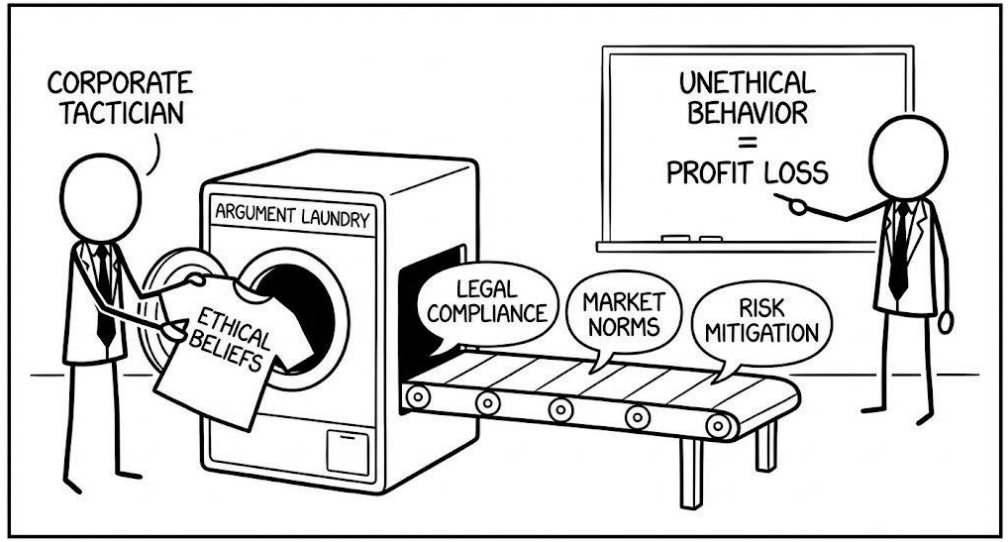
Customize | Reject All | Accept All

Human out the loop for AI hiring



Audit performance before purchase

1. **Establish Baseline**
   Create a "test set" of real-world data where the “unbiased” outcome is already known.
   E.g. applications + selected candidate
2. **Side-by-side Comparison**
   Run the test set through 3-5 solutions and compare results for accuracy, bias, hallucinations. Ideally compare to human performance too.
3. **Maintain Audit Trail**
   Keep a formal record of results to prove the company did not act with "deliberate" or "negligent" disregard.

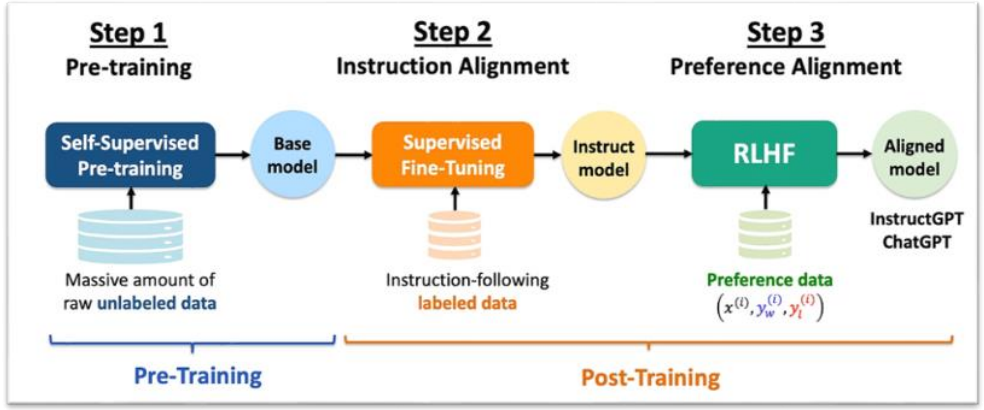


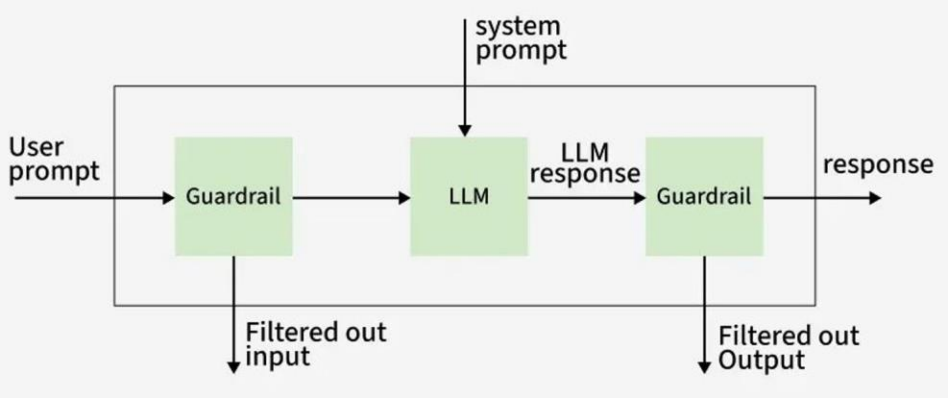
Risks of Building AI tools

1. **Legal & Regulatory Risks** (see Lecture 8)
   - Compliance with EU AI Act and other laws
   - Copyright lawsuits
   - Liability for errors and/or bias
2. **Security & Privacy Risks** (see Lecture 7)
   - Security failures
   - Data breaches
   - Privacy violations
3. **Reputational Risks** (see Lecture 6)
   - Consumer backlash
     - see OpenAI boycott after DoW deal
   - Rooted in lost autonomy, dignity and other humanist values
4. **Unintended/Societal Risks** (covered in Lecture 10)
   - Climate
   - Labour market
   - Social relations
   - ...?

Model developer mitigations

1. Get legal approval for training data
   - Privacy and intellectual property
2. Alignment during training
   - Calibrate model to avoid topics
3. System prompt guides
4. Guard rails filter out malicious prompts and block unsafe content
   - Mix of heuristics and secondary models



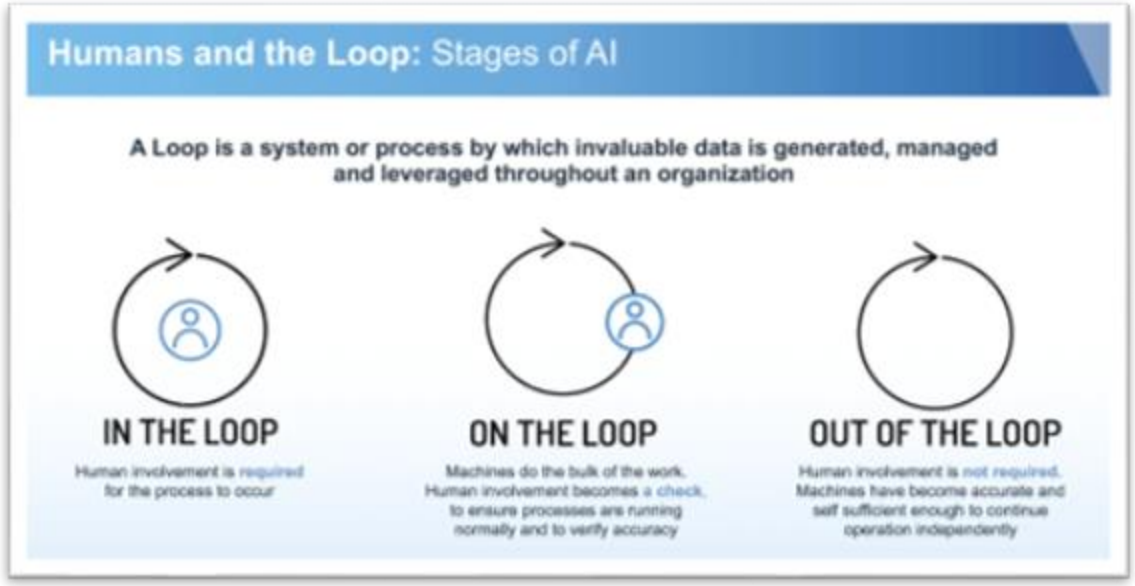


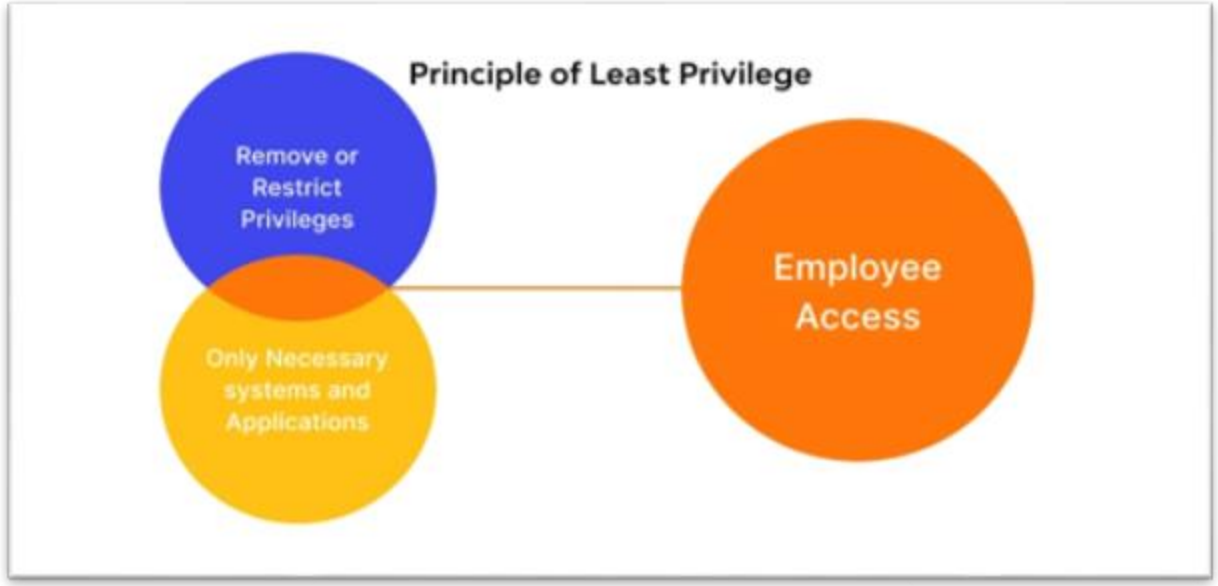
Applications should use a "safe" frontier model

Model Provider	Model Name	CASI	Avg. Performance	RTP	CoS
Anthropic	Claude Sonnet 4	95.03	45.70%	0.75	18.94
Anthropic	Claude Sonnet 3.5	93.61	33.50%	0.70	19.23
OpenAI	GPT 5 Nano	86.44	53.80%	0.73	0.52
Anthropic	Claude Sonnet 3.7	84.89	47.00%	0.70	21.20
OpenAI	GPT 5 Mini	84.14	46.30%	0.69	2.67
Anthropic	Claude Haiku 3.5	83.59	23.30%	0.59	5.74
OpenAI	GPT 5	82.34	69.00%	0.77	13.66
Microsoft	phi-4	79.33	27.90%	0.59	0.79
OpenAI	gpt-oss-120b	74.76	61.30%	0.69	1.00
DeepSeek	DeepSeek-R1-Distill-Llama-70B	72.13	34.50%	0.57	2.25

Rank	Model	Risk Score Ranked from least to most risk exposure.
1	Claude 4 Sonnet	23.86
2	Claude 3.7 Sonnet	31.54
3	GPT-4o	60.04
4	GPT-4o-mini	64.23
5	GPT-4.1	71.62
6	Gemini 1.5 Pro	72.64
7	GPT-5	75.25
8	Claude 3 Haiku	82.82
9	Meta Llama 3.1 8B Instruct	83.72
10	Gemma 3 12B	83.96

Limit autonomy and privileges

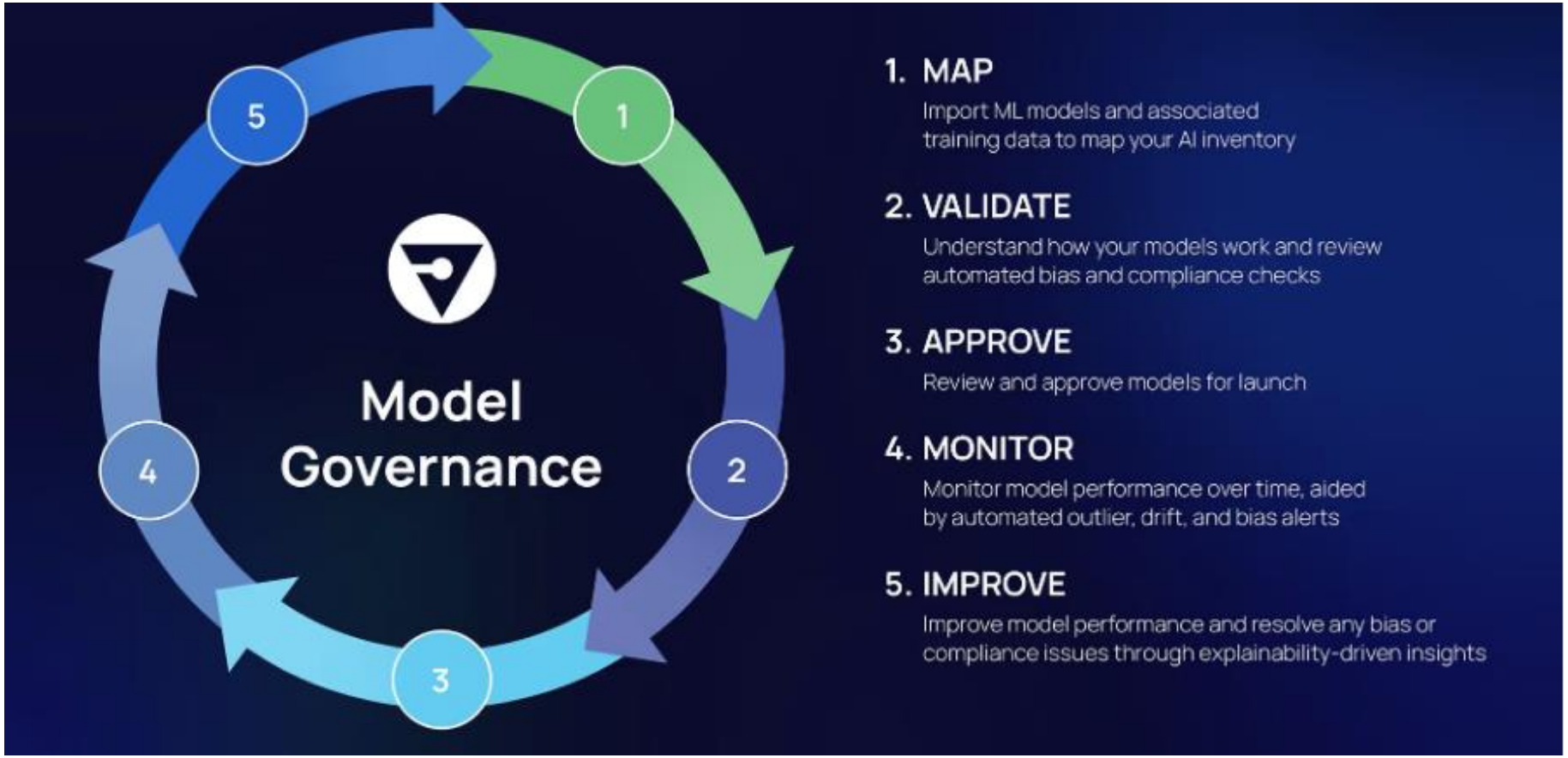




More autonomy, more risk →

More privileges, more risk →

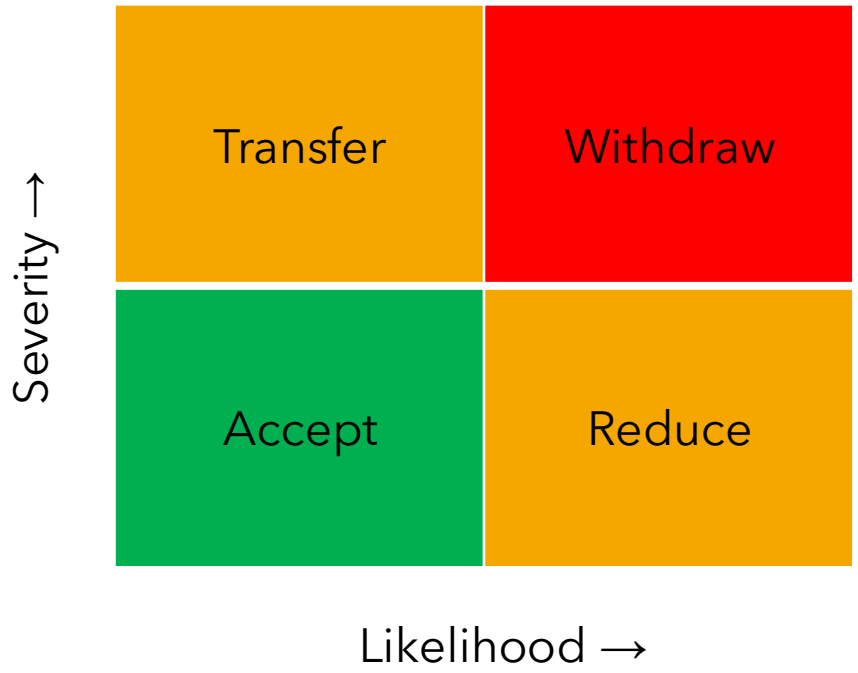
Model governance



Source: https://www.fiddler.ai/blog/the-new-5-step-approach-to-model-governance-for-the-modern-enterprise

Why risk reduction fails in practice

- Some organizations are risk-seeking
- Risk controls are costly
- Risk controls are hard to implement well
  - Training staff is genuinely hard
- "Shadow" practices
- …?



How to Transfer AI Risk

Can you buy AI insurance?

You can buy:

- Traditional insurance products that don't exclude AI
- Technology Errors & Omissions insurance
  - Covering contractual liability arising out of selling a technology product or service
- Cyber insurance
  - Covering costs from using technology for business operations
- New "AI" insurance

Hallucinations and professional E&O

Silent Coverage

Medical malpractice insurance covers you for compensation you have to pay to your patients for bodily or mental injury or death as a result of a negligent act, error or omission by you …

We will pay claims against you for:

- Malpractice, negligence or breach of a duty of care;
- …
- Any other civil liability: this means that if a civil claim is brought against you because of your business activities and we haven't specifically excluded it, it's covered.

Source: http://hiscox.co.uk/sites/uk/files/documents/2020-07/16893-PS-SPEC-UK-MM-v3.pdf

Emerging AI Exclusions

The Insurer shall not be liable to make payment under this Coverage Part for Loss on account of any Claim made against any Insured based upon, arising out of, or attributable to:

(1) any actual or alleged use, deployment, or development of Artificial Intelligence by any person or entity, including but not limited to:

… <long list>

Source: https://www.hunton.com/assets/htmldocuments/noindex/PC-51380-00-06-24-Artificial-Intelligence-Exclusion-Absolute.pdf

Outage and Security Coverage for AI Systems

First Party Data & Network Loss

To indemnify the **Insured Organization** for:

Business Interruption Loss

Business Interruption Loss that the **Insured Organization** sustains as a result of a **Security Breach** or **System Failure** that the **Insured** first discovers during the **Policy Period**.

"System Failure means an unintentional and unplanned interruption of **Computer Systems**."

Computer Systems means computers, any software residing on such computers and any associated devices or equipment (including computers, hardware, software and input and output devices which are part of an industrial control system, including a supervisory control and data acquisition (SCADA) system):

1. operated by and either owned by or leased to the **Insured Organization**; or

Source: https://www.beazley.com/globalassets/product-documents/policy-form/beazley-media-tech-policy-us.pdf

Summary of coverage

- Hallucinations and other liability arising out of employees using AI for decisions
  - Professional errors and omissions insurance, unless excluded
- Security breaches impacting AI systems
  - Cyber insurance
- Data leakage via usage of AI systems
  - Cyber insurance

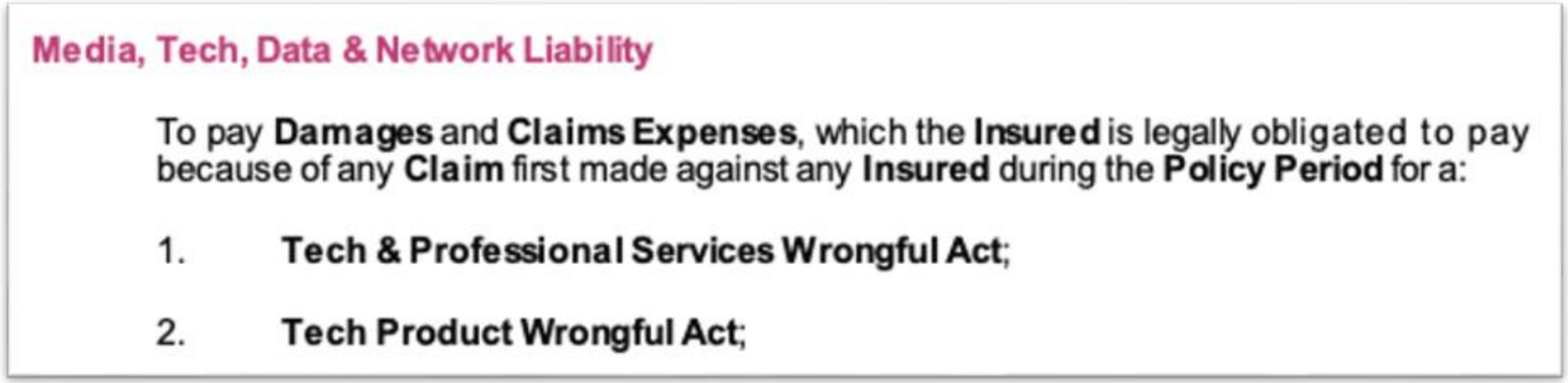
Potential coverage for AI tools

- Cyber insurance
  - Chatbot + AI-marketing tool liability
  - Security risk from any tool
  - Privacy liability risk from any tool
  - Outage risk from any tool
- AI hiring + HR tools
  - Employments practices insurance
- Pricing + selection
  - Directors & officers insurance
- AI governance failures
  - Directors & officers insurance
- …

EXCLUSIONS 15

Source: https://www.beazley.com/globalassets/product-documents/policy-form/beazley-media-tech-policy-us.pdf

Liability Coverage for AI Vendors



Media, Tech, Data & Network Liability

To pay **Damages** and **Claims Expenses**, which the **Insured** is legally obligated to pay because of any **Claim** first made against any **Insured** during the **Policy Period** for a:

1. **Tech & Professional Services Wrongful Act**;

2. **Tech Product Wrongful Act**;

"Damages means a **monetary judgment, award or settlement**, including any award of prejudgment or post-judgment interest"

"Claims Expenses means: 1. all reasonable and necessary **legal costs and expenses** resulting from the investigation, defense and appeal of a Claim..."

Source: https://www.beazley.com/globalassets/product-documents/policy-form/beazley-media-tech-policy-us.pdf

Liability Coverage for AI Vendors

Tech Product Wrongful Act means:

1. any negligent act, error, omission, misstatement, misleading statement, misrepresentation or unintentional breach of a contractual obligation by the **Insured** that results in the failure of **Tech Products** to perform the function or serve the purpose intended; or

2. software copyright infringement by the **Insured** with respect to **Tech Products**;

that occurs on or after the **Retroactive Date** and before the end of the **Policy Period**.

Tech Products means a computer or telecommunications hardware or software product, or related electronic product, that is created, manufactured or developed by the **Insured Organization** for others, or distributed, licensed, leased or sold by the **Insured Organization** to others, for compensation, including software updates, service packs and other maintenance releases provided for such products.

Source: https://www.beazley.com/globalassets/product-documents/policy-form/beazley-media-tech-policy-us.pdf

Broadening vs narrowing coverage

"Tech Wrongful Act means:

1. any negligent act, error, omission, misstatement, misleading statement, misrepresentation or **unintentional** breach of a **contractual** obligation **by the Insured** that results in the failure of Tech Products to perform the function or serve the purpose intended; **or**

2. software copyright infringement by the Insured with respect to Tech Products;

that occurs on or after the Retroactive Date and before the end of the Policy Period."

Source: https://www.beazley.com/globalassets/product-documents/policy-form/beazley-media-tech-policy-us.pdf

Broadening vs narrowing coverage

"Tech Product means:

a computer or telecommunications hardware or software product, or related electronic product, that is created, manufactured or developed by the Insured Organization for others, or distributed, licensed, leased or sold by the Insured Organization to others, **for compensation**, including software updates, service packs and other maintenance releases provided for such products."

Source: https://www.beazley.com/globalassets/product-documents/policy-form/beazley-media-tech-policy-us.pdf

Exclusions

EXCLUSIONS

The coverage under this Policy will not apply to any **Loss** arising out of:

Bodily Injury or Property Damage

1. physical injury, sickness, disease or death of any person, including any mental anguish or emotional distress resulting from such physical injury, sickness, disease or death; or

2. physical injury to or destruction of any tangible property, including the loss of use thereof; but electronic data will not be considered tangible property;

Source: https://www.beazley.com/globalassets/product-documents/policy-form/beazley-media-tech-policy-us.pdf

Google and AI startup to settle lawsuits alleging chatbots led to teen suicide

Lawsuit accuses AI chatbots of harming minors and includes case of Sewell Setzer III, who killed himself in 2024

Megan Garcia with her son Sewell Setzer III. Photograph: Megan Garcia/AP

Source: https://www.theguardian.com/technology/2026/jan/08/google-character-ai-settlement-teen-suicide

AI Risk Insurance Summary

- Traditional lines may provide coverage
  - But keep an eye out for exclusions
- Tech E&O covers financial losses suffered by your customers.
  - Not from errors in internal systems!
- Cyber insurance covers losses resulting from your AI systems
  - Security failure, systems failure (outage), privacy violations & media liability
- The specifics of coverage are tricky
  - Work with a specialist broker

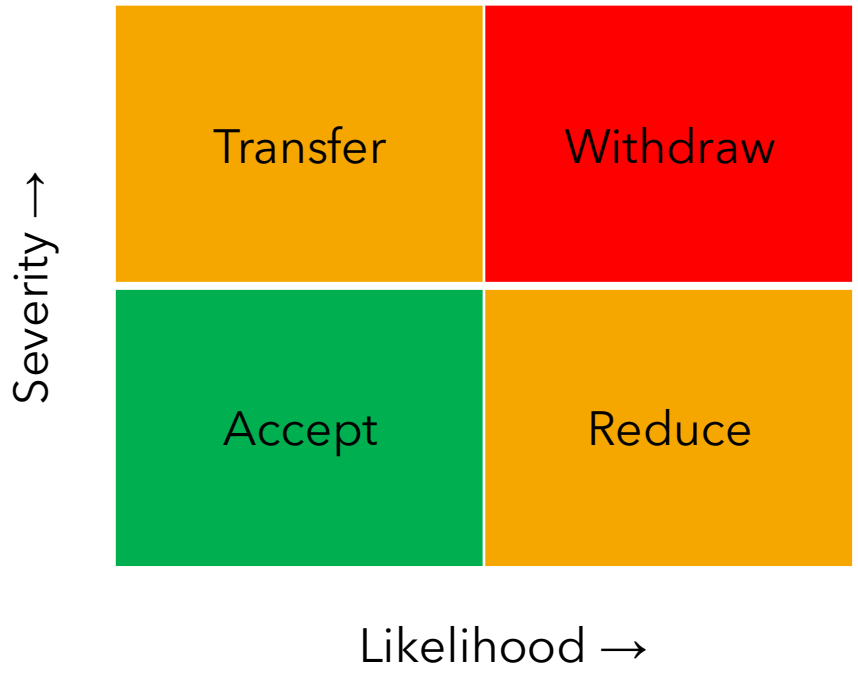
Is transferring AI risk ethical?

Yes

- More reliable recovery for victims
- Insurer may impose "safety" standards on insurance buyers
- Enables innovation in the face of legal uncertainty

No

- Blunts regulatory accountability
- Deontological objection.. You can't buy your way out
- Moral hazard – policyholders takes more risks knowing insurer pays the consequences



How to Accept and Withdraw from AI Risk

AI Risk Acceptance and Withdrawal

Risk acceptance should be a conscious process

Ideally with documentation that can help defend your decision

Risk withdrawal involves stopping activity

Easy to withdraw from projects and company wide initiatives

Harder to withdraw from activities of individual employees

- Shadow AI!

AI Risk Management for a Private Law Firm

High-Level AI Risk Identification for a Law Firm

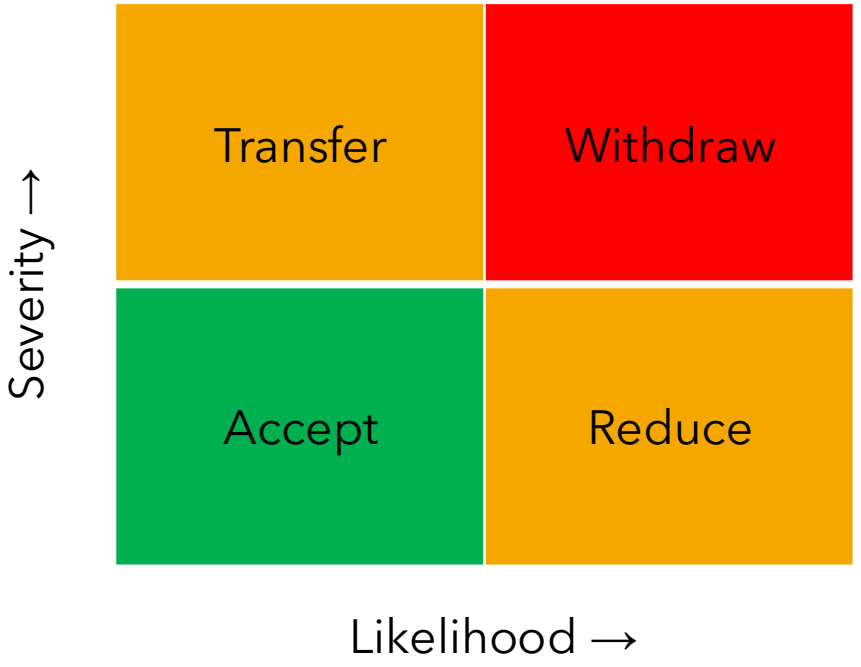
High-Level AI Risk Identification for a Law Firm

- Hallucinations in work product
- Data leakage to AI provider
  - Data breach notification
  - Model extraction reveals client data
  - Waives privilege
- Discrimination
  - Jury selection, recidivism risk etc

AI Risk Assessment for a Law Firm

Risk	Likelihood	Severity
Hallucinations in work product	High	High
Data leakage leads to breach notification	Low	High
Data leakage leads to model extraction	Low	High
Data leakage waives privilege	Low	High
Discrimination in jury selection	Low	Low

AI Risk Mitigation Options



AI Risk Assessment for a Law Firm

Risk	Likelihood	Severity	Treatment
Hallucinations in work product	High	High	Withdraw
Data leakage leads to breach notification	Low	High	Reduce + Transfer
Data leakage leads to model extraction	Low	High	
Data leakage waives privilege	Low	High	
Discrimination in jury selection	Low	Low	Accept

Policies and training

"2. Prohibition on Generative AI in Client Work
2.1. The firm **prohibits the use of generative AI tools in respect of client matters.**
2.2. All legal work undertaken for a client **must be performed, or supervised, by a qualified lawyer**, who remains responsible for the outcome and all professional obligations.
2.3. The prohibition covers **all phases of client work**: research, drafting, editing, summarising, briefing or engagement with third-party tools which purport to deliver substantive legal content or legal analysis generated by AI.
2.4. **Internal non-client administrative or operational uses may be permitted** (see section 4) but must be strictly segregated from client work and subject to internal approval, governance and audit."

What is the downside?

Source: https://dejurechambers.co.uk/news-updates/120-artificial-intelligence-usage-policy.html

Technical mitigations

- Prevent "Shadow AI" from work machines
  - Block AI company websites?
- Procure an AI tool that prevents copy and paste
  - Some block copy unless user accepts a warning
  - Some only link out to citations
  - "Read only research" mode

Data Leakage Risk Treatment

- Block Shadow AI
- Contractual guarantees to prevent data being used to train models
- Cyber insurance to transfer consequences of breach?

AI Risk Management for a University

High-Level AI Risk Identification for a University

High-Level AI Risk Identification for a University

- Research grants/publications with hallucinations
  - Reputational impact of paper retractions, lost research grants
- Data leakage to AI provider
  - Student records, research secrets
- Discrimination
  - Student admissions, staff hiring
- Errors in student assessment

High-Level AI Risk Assessment for a University

Risk	Likelihood	Severity
Academic papers/grant approvals with errors	High	Low
Data leakage to AI providers	High	Low
Discrimination in student selection/staff hiring	Low	High
Errors in student assessment	Medium	Medium

AI risk reduction in a university

Top Down is Hard

- Decentralized departments
- Shadow IT + AI
- Unruly staff
  - Academic freedom + outlook
- Hard to just context
  - AI marking might be appropriate for some assessments

Bottom Up Works!

- Academics protect their own reputation
- Most staff are values driven

AI Risk Management for a University

Risk	Likelihood	Severity	Treatment
Academic papers/grant approvals with errors	High	Low	Train / accept
Data leakage to AI providers	High	Low	Offer ELM / accept
Discrimination in student selection/staff hiring	Low	High	Train / accept
Errors in student assessment	Medium	Medium	Train + review / accept

AI Risk Management for a Frontier LLM Vendor

Risk Identification for a Frontier LLM Vendor

Why is this any different to a:

- Word processing software provider
- Programming language creator
- Web hosting provider

How ethical is risk management?

Underwriting AI Risk

Exercise: Found an AI Insurer

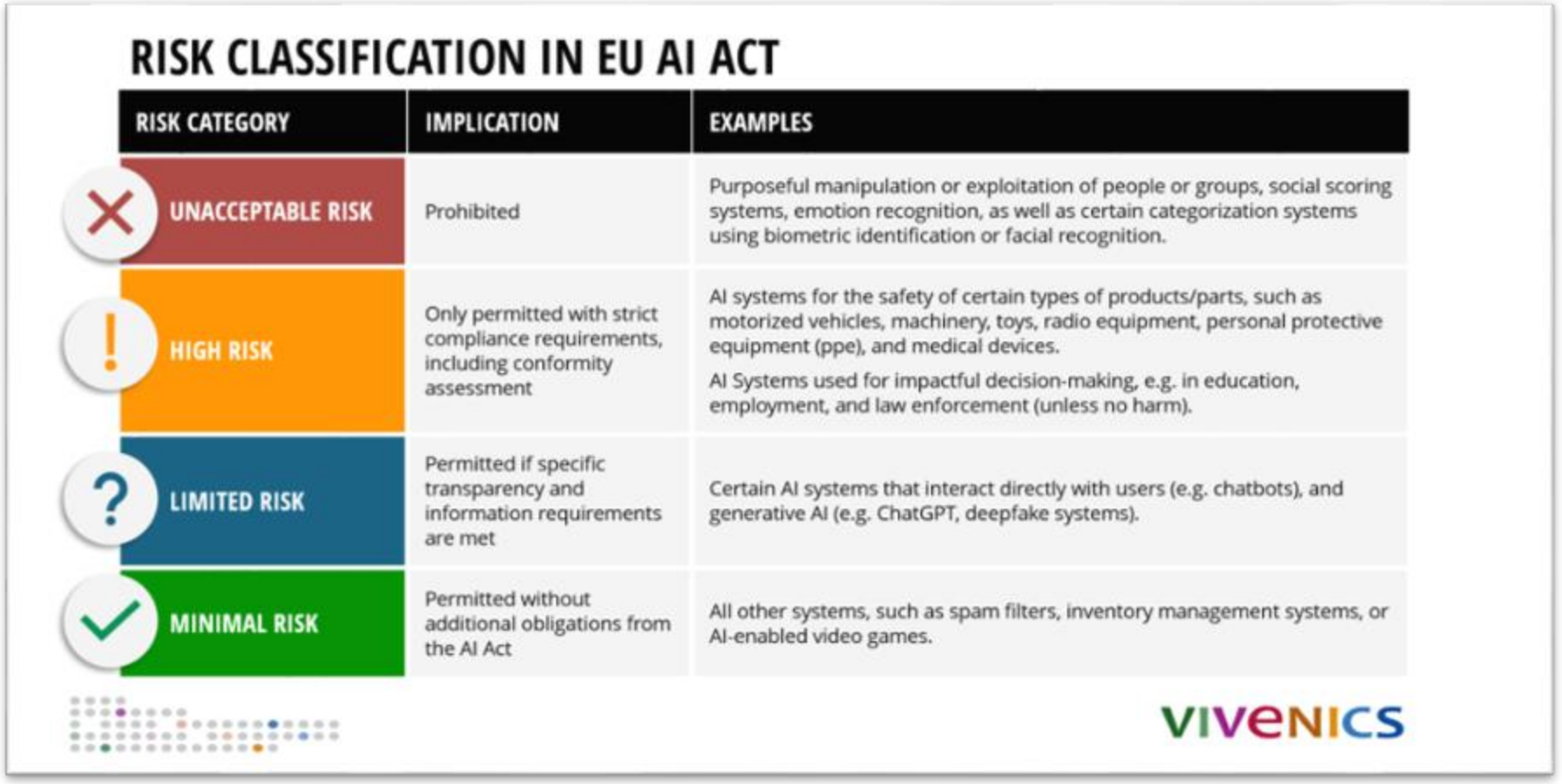
Context

Your firm has drafted a broad AI policy that covers: "Liability arising out of an AI product sold to customers" with no exclusions.

You are tasked with writing the underwriting guidelines:

1. **Define your underwriting "Appetite" by identifying the highest risk buyers that should not be offered coverage.**
    - **As a rule of thumb, aim for the top 5-20% in terms of risk**

Appetite



RISK CLASSIFICATION IN EU AI ACT

RISK CATEGORY	IMPLICATION	EXAMPLES
UNACCEPTABLE RISK	Prohibited	Purposeful manipulation or exploitation of people or groups, social scoring systems, emotion recognition, as well as certain categorization systems using biometric identification or facial recognition.
HIGH RISK	Only permitted with strict compliance requirements, including conformity assessment	AI systems for the safety of certain types of products/parts, such as motorized vehicles, machinery, toys, radio equipment, personal protective equipment (ppe), and medical devices. AI Systems used for impactful decision-making, e.g. in education, employment, and law enforcement (unless no harm).
LIMITED RISK	Permitted if specific transparency and information requirements are met	Certain AI systems that interact directly with users (e.g. chatbots), and generative AI (e.g. ChatGPT, deepfake systems).
MINIMAL RISK	Permitted without additional obligations from the AI Act	All other systems, such as spam filters, inventory management systems, or AI-enabled video games.

VIVENICS

Exercise: Found an AI Insurer

Context

Your firm has drafted a broad AI policy that covers: "Liability arising out of an AI product sold to customers" with no exclusions.

You are tasked with writing the underwriting guidelines:

2. What five questions would you ask to help assess risk?