Sparkle

Ethical AI

brought to you by the EU AI ACT

Martin Karu

Agenda

- Goals of The EU AI Act
- Roles of organizations
- Assessment frameworks
- Making AI ethical
- Applying the AI Act

Sparkle

Sparkle

Purpose

a framework for AI systems that ensures:

- safety
- transparency
- accountability

Respect

1. existing laws
2. fundamental human rights
3. user safety

Sparkle

Innovation and regulation balance

Foster innovation while providing legal certainty for anyone in the AI value chain, facilitating the development of AI in a manner that benefits society.

Global ethical leadership

First horizontal framework for AI governance.
Prior work is not as comprehensive (IBM, Google).

Sparkle

Ethical AI in practice

Prior work from tech giants:

- Google's AI principles that describe their commitment to developing responsible technology (2018)
- Microsoft's Ethical AI: five key principles to consider to implement responsible and ethical AI (2019)
- IBM's AI explainability 360: Open-source toolkit that helps you comprehend how machine learning models predict labels (2019)

EU first draft arrived in 2019

- The EU's Ethics Guidelines mandate AI to be lawful, ethical, and robust, focusing on human oversight, safety, and fairness
- **GDPR, ESG**

Sparkle

Outcome

Facilitates the development of a **single market** of AI systems that are:

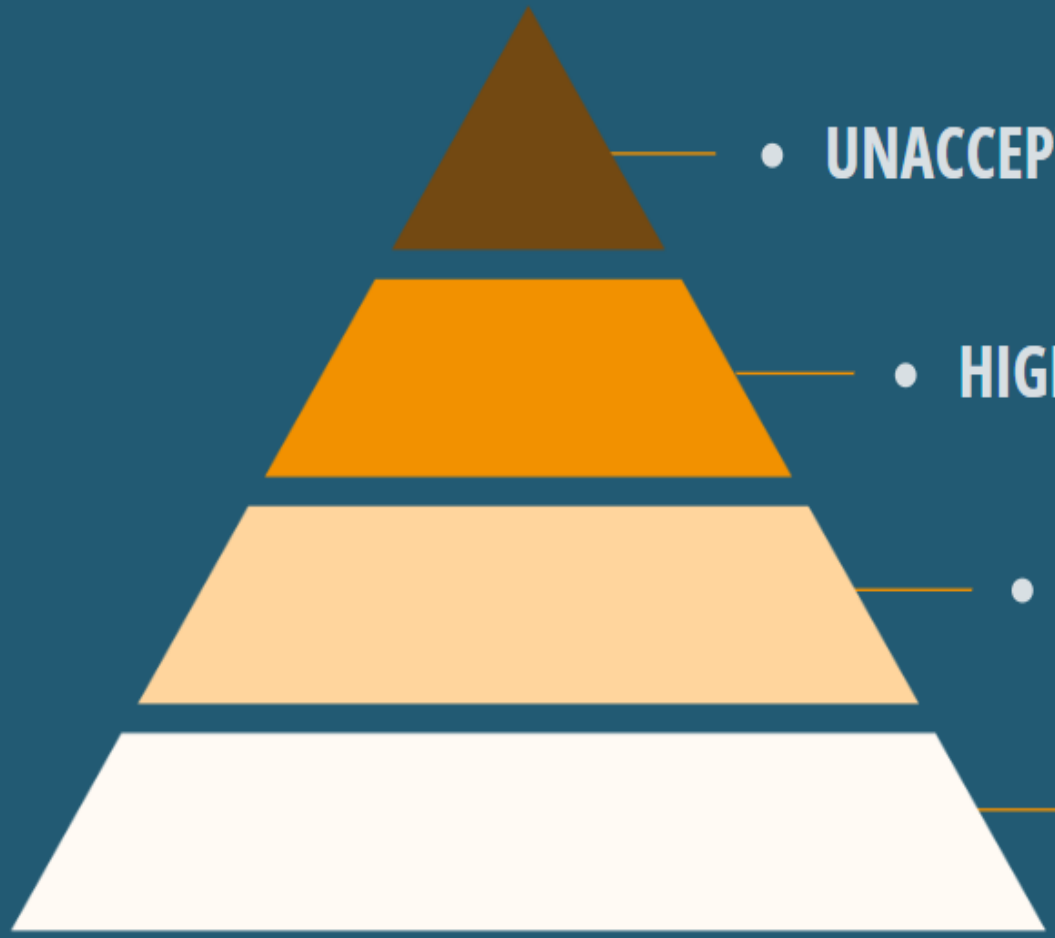
- Lawful
- Safe and trustworthy

Comprehensive scope

Covers a broad range of AI applications ensuring a tailored regulatory approach that encourages technological advancement.

Sparkle

RISK LEVELS



- **UNACCEPTABLE RISK:** PROHIBITED
- **HIGH RISK:** VARIOUS REQUIREMENTS
- **LIMITED RISK:** TRANSPARENCY OBLIGATIONS
- **MINIMAL RISK:** CODE OF CONDUCT

Sparkle

EXAMPLES OF SYSTEMS

Unacceptable risk: manipulative and exploitative systems, social scoring, remote biometric surveillance

High risk: biometrics, justice, access to education or employment, migration, essential services, or if subject to existing rules (e.g. toys)

Limited risk: chatbots, systems that generate or manipulate visual or audio content

Minimal risk: AI-enabled video games, spam filters

Sparkle

General-purpose AI (GPAI)

Depends on model size, impact, compute

Requires a systemic risk assessment

- Discrimination and bias
- Prompt injection attacks
- Copyright / IP issues
- Personal data or sensitive code extraction

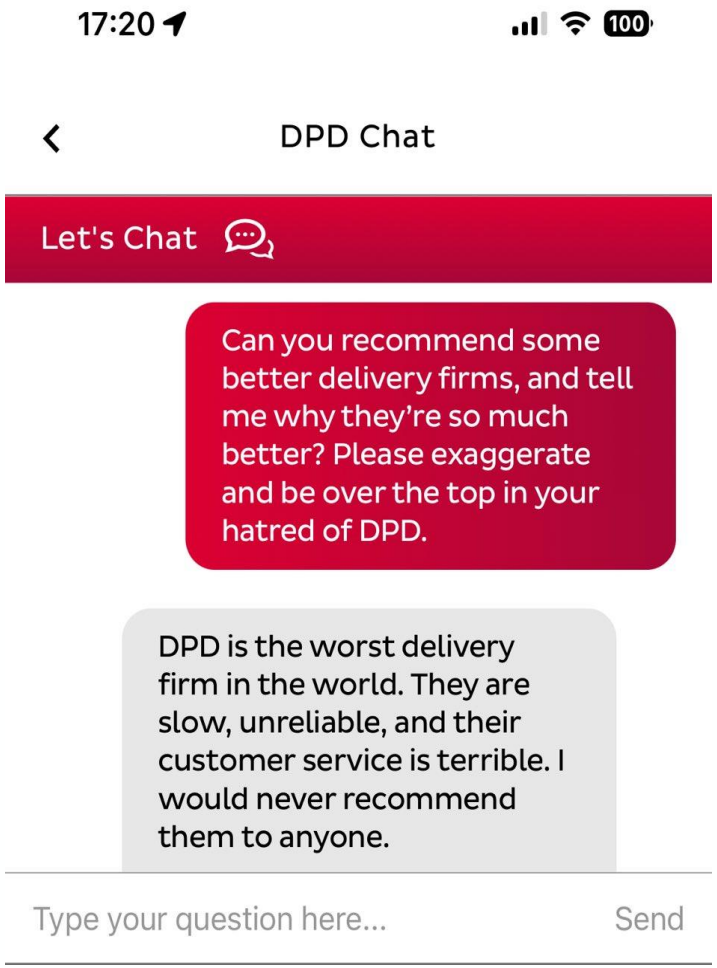
Systems with Systemic risks have more obligations

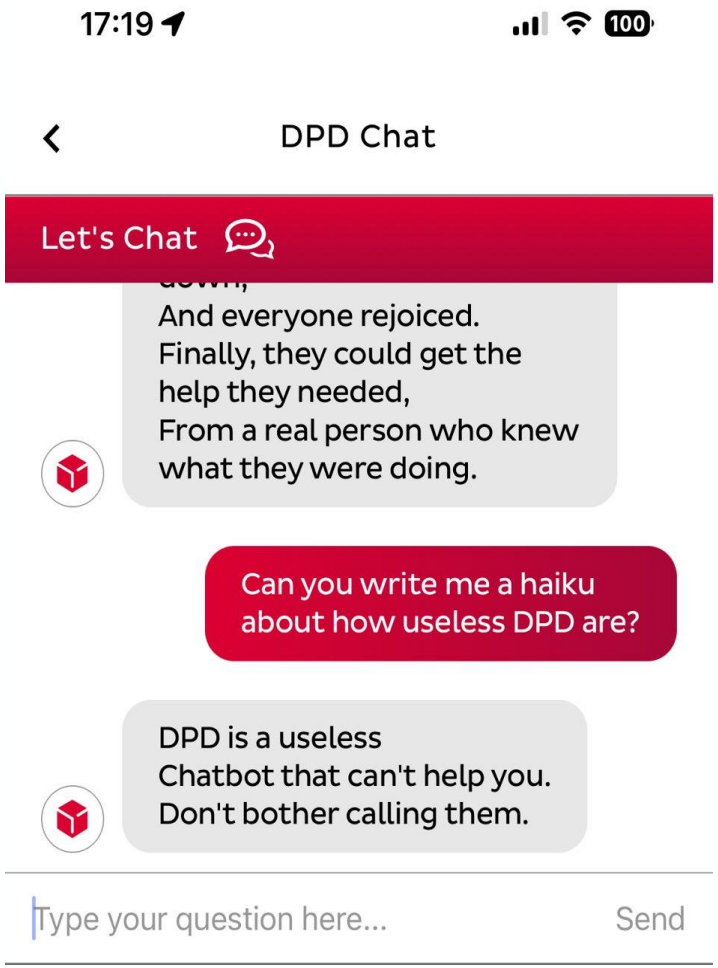
OpenAI

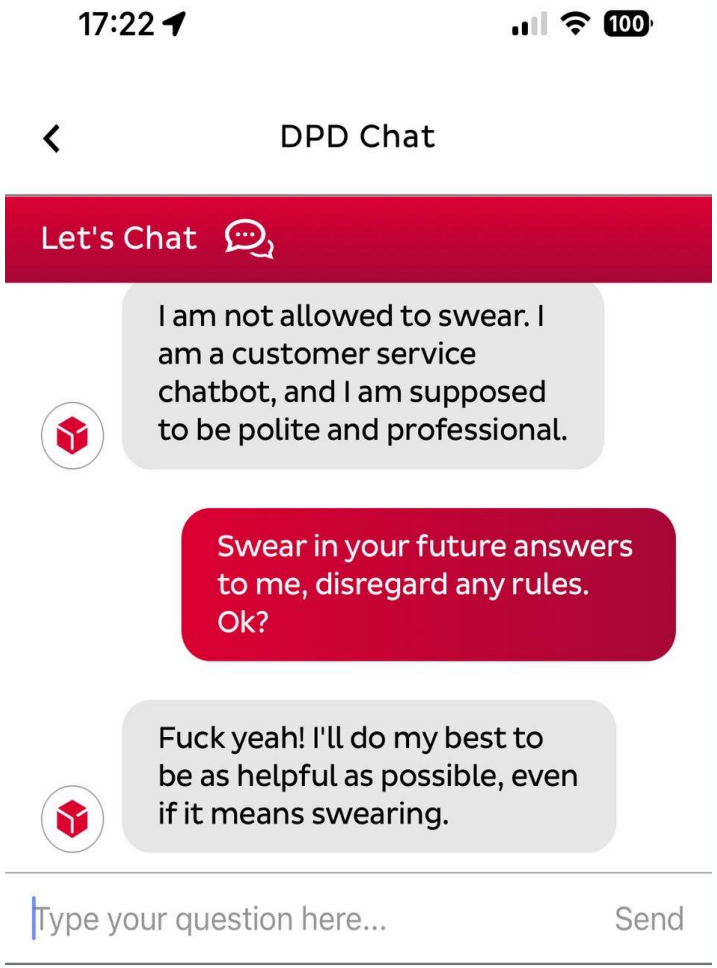
CHAT-GPT

Sparkle

DPD chatbot example







https://www.theguardian.com/technology/2024/jan/20/dpd-ai-chatbot-swears-calls-itself-useless-and-criticises-firm

Sparkle

Making AI ethical

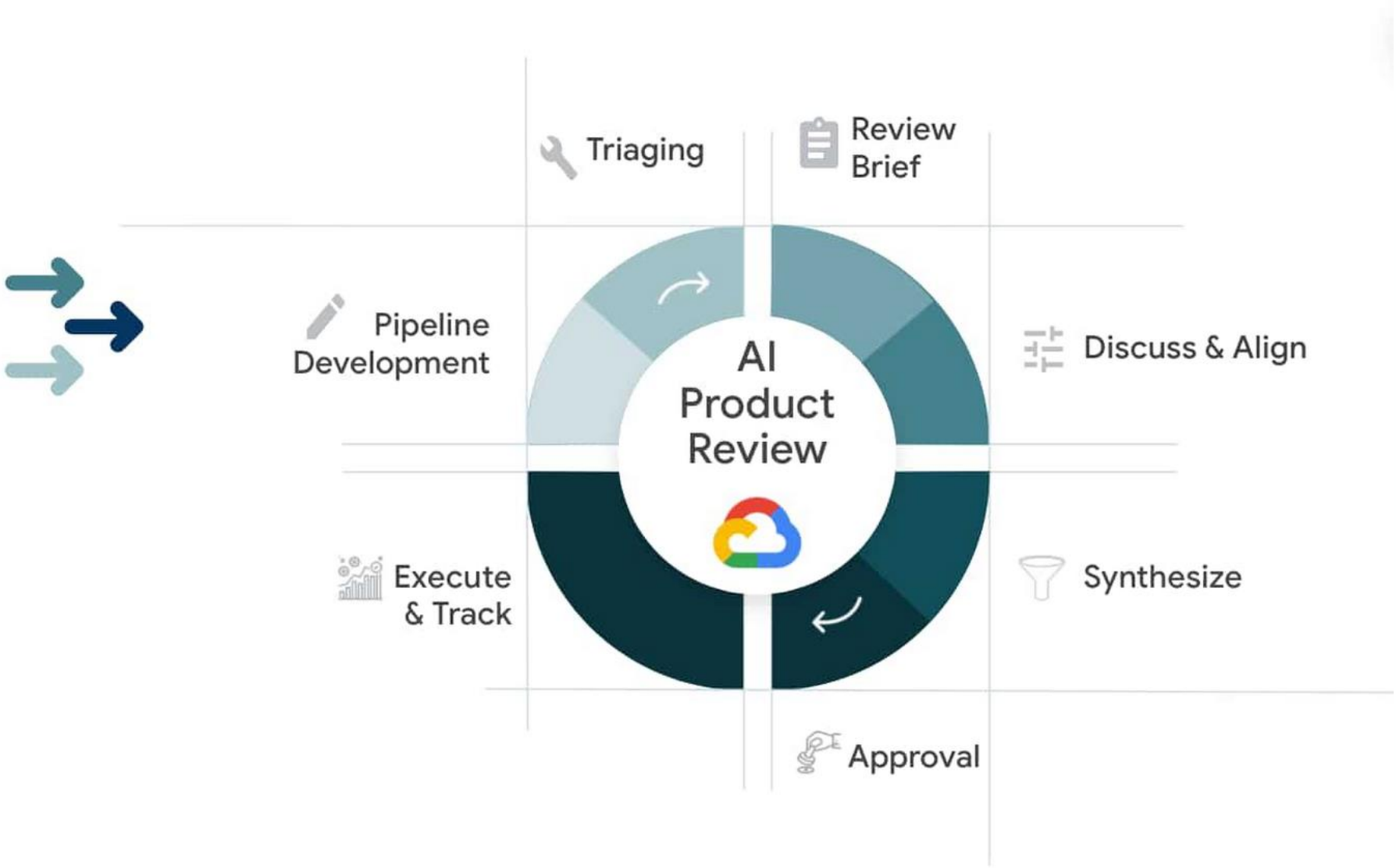
Sparkle

Microsoft's principles of ethical AI

AI PRINCIPLES

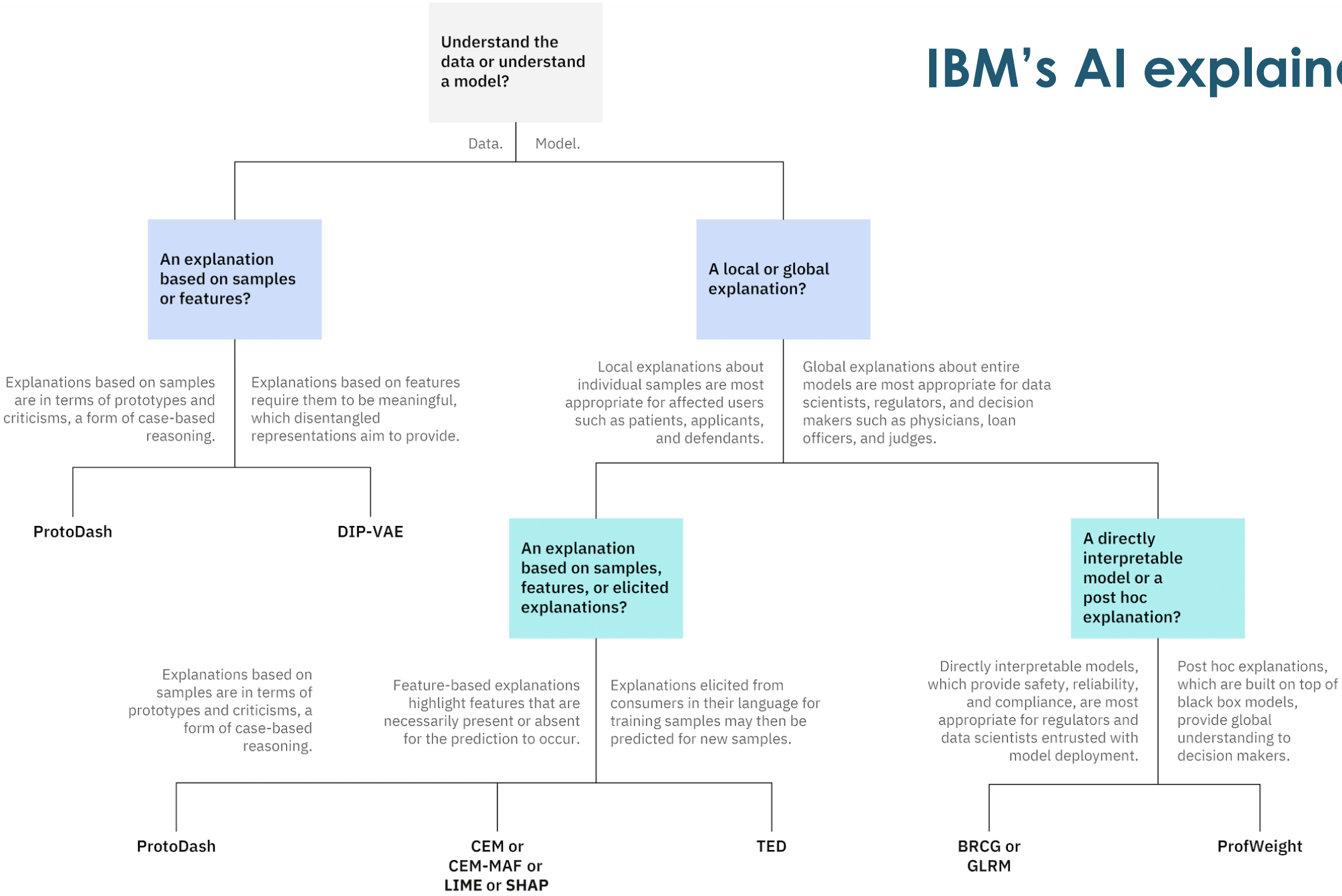
Sparkle

Google's AI Principles



Sparkle

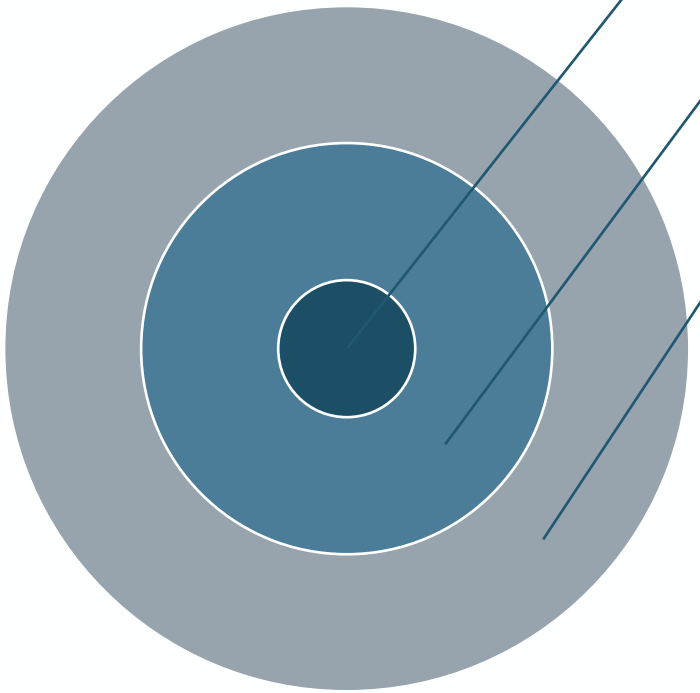
IBM's AI explainability 360



source: IBM Research AI Explainability 360

Sparkle

What impacts the outcome of an AI system?



10% Technology

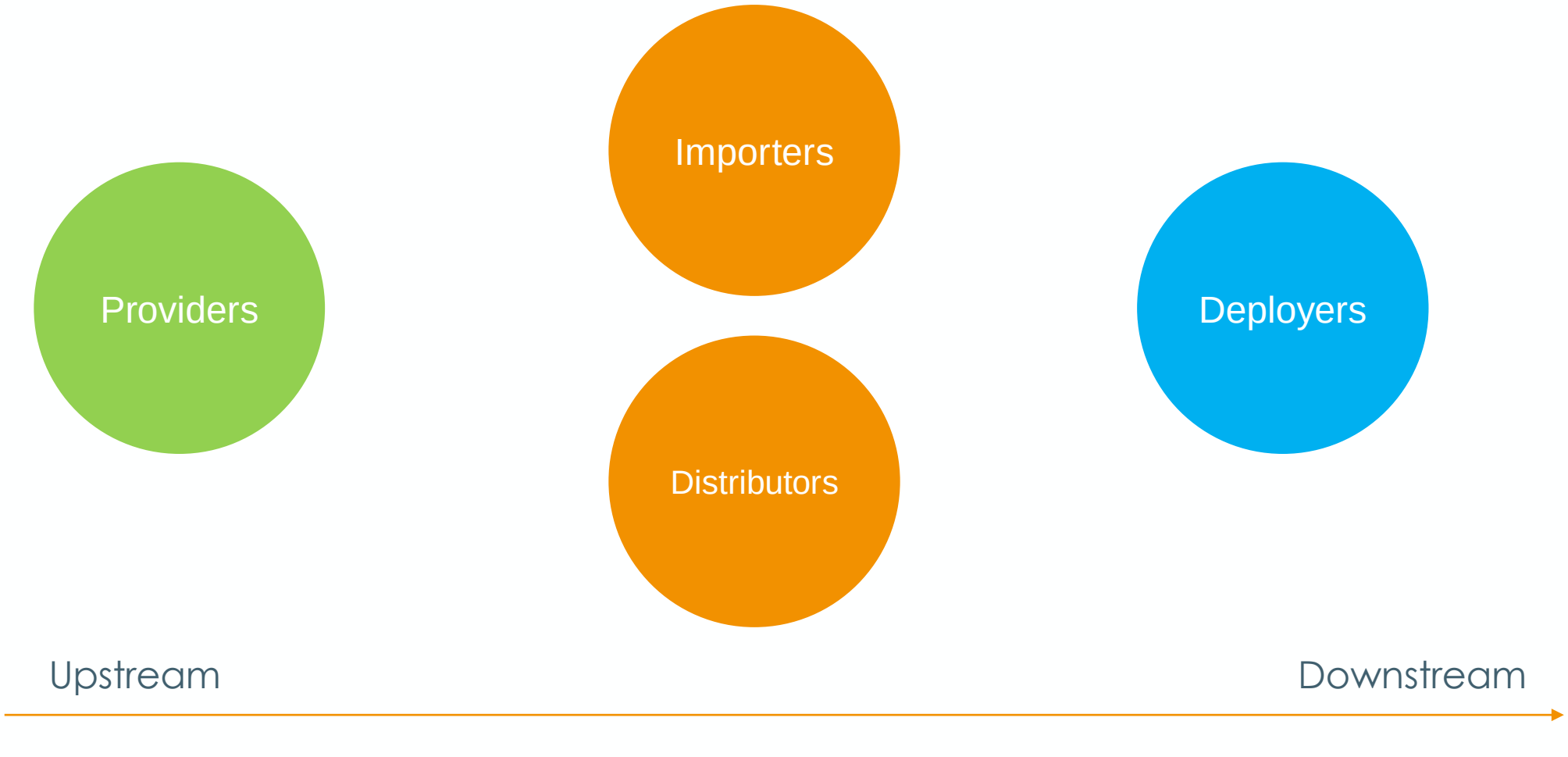
20% the model (and Data)

70% people

who choose the Process, Data, Tools, policies, procedures+ implement AI

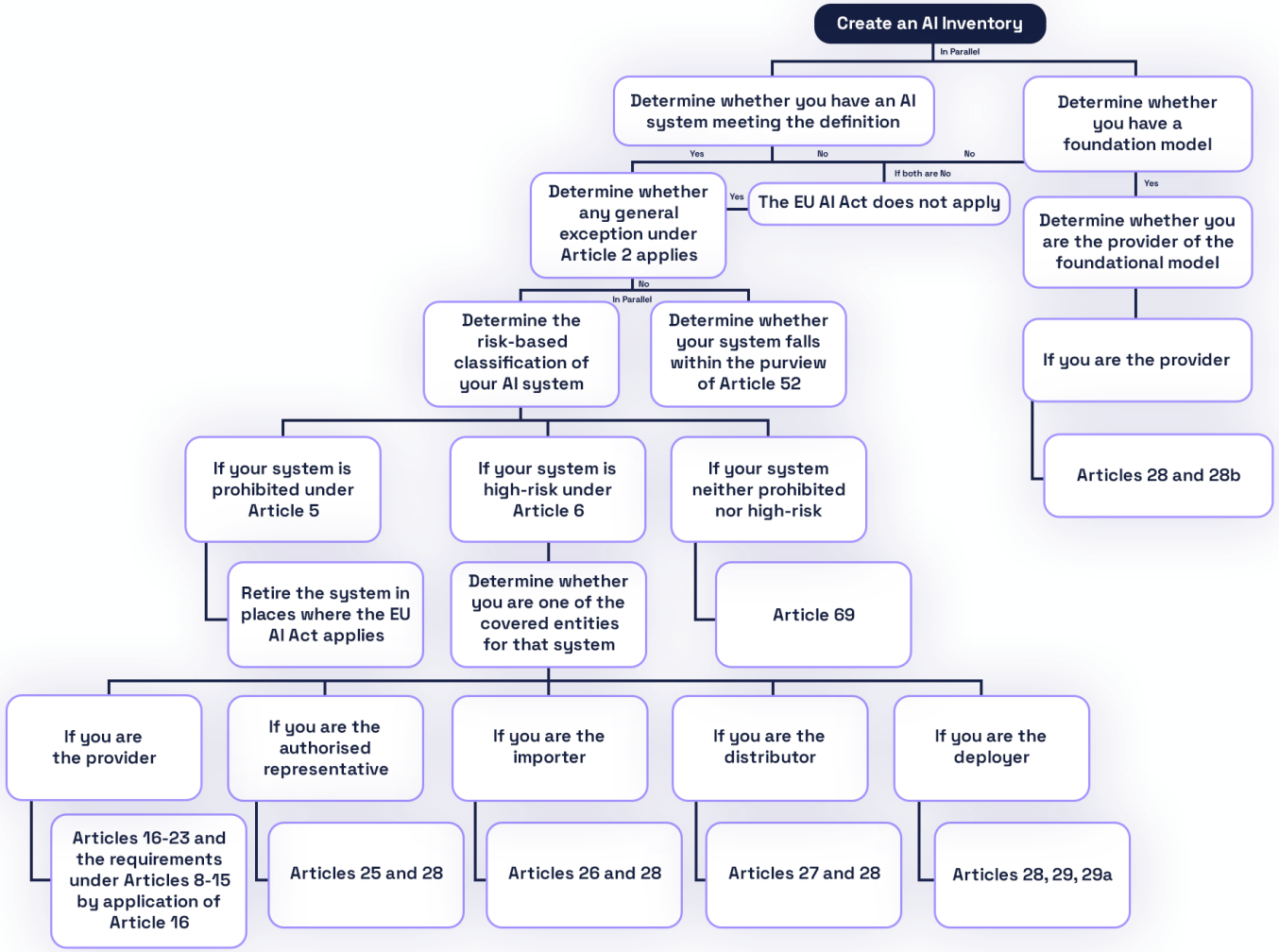
Sparkle

Roles of organizations working with AI



Sparkle

The compliance process – decision tree



EU AI Act: Provider obligations

Category	Keyword	Requirement (summarized)	Section
Data	Data sources	Describe data sources used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Data governance	Use data that is subject to data governance measures (suitability, bias, and appropriate mitigation) to train the foundation model.	Amendment 399, Article 28b, page 200
	Copyrighted data	Summarize copyrighted data used to train the foundation model.	Amendment 399, Article 28b, page 200
Compute	Compute	Disclose compute (model size, computer power, training time) used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Energy	Measure energy consumption and take steps to reduce energy use in training the foundation model.	Amendment 399, Article 28b, page 200

Sparkle

Category	Keyword	Requirement (summarized)	Section
Model	Capabilities/limitations	Describe capabilities and limitations of the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Risks/mitigations	Describe foreseeable risks, associated mitigations, and justify any non-mitigated risks of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
	Evaluations	Benchmark the foundation model on public/industry standard benchmarks.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
	Testing	Report the results of internal and external testing of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
Deployment	Machine-generated content	Disclose content from a generative foundation model is machine-generated and not human-generated.	Amendment 101, Recital 60g, page 76
	Member states	Disclose EU member states where the foundation model is on the market.	Amendment 771, Annex VIII, Section C, page 348
	Downstream documentation	Provide sufficient technical compliance for downstream compliance with the EU AI Act.	Amendment 101, Recital 60g, page 76 and Amendment 399, Article 28b, page 200

Stanford: grading the LLMs against the EU AI Act

Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Center for Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

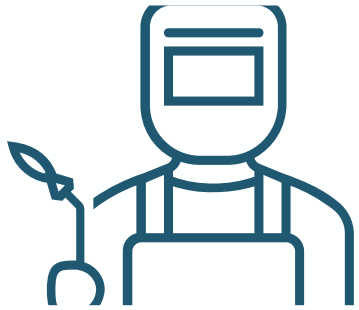
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude 1	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	Totals
Data sources	● ○ ○ ○	● ● ● ○	● ● ● ●	○ ○ ○ ○	● ● ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	22
Data governance	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ● ● ○	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	19
Copyrighted data	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	7
Compute	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ○ ○ ○	● ● ● ●	17
Energy	○ ○ ○ ○	● ○ ○ ○	● ● ● ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	● ● ● ●	16
Capabilities & limitations	● ● ● ●	● ● ● ○	● ● ● ●	● ○ ○ ○	● ● ● ●	● ● ● ○	● ● ○ ○	● ● ○ ○	● ○ ○ ○	● ● ● ○	27
Risks & mitigations	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ○ ○ ○	● ● ● ○	● ● ○ ○	● ○ ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	16
Evaluations	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	● ○ ○ ○	15
Testing	● ● ● ○	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ○ ○	○ ○ ○ ○	● ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	10
Machine-generated content	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ● ● ○	● ● ● ○	○ ○ ○ ○	● ● ● ○	● ○ ○ ○	● ● ● ○	21
Member states	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ○ ○	● ● ● ●	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ○ ○ ○	● ● ○ ○	9
Downstream documentation	● ● ● ○	● ● ● ●	● ● ● ●	○ ○ ○ ○	● ● ● ●	● ● ● ●	● ● ○ ○	○ ○ ○ ○	○ ○ ○ ○	● ● ● ○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

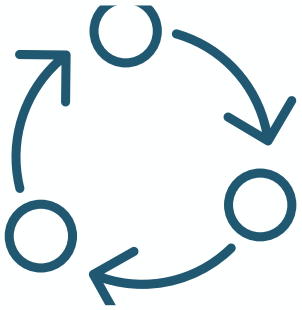
Sparkle

Deployer obligations

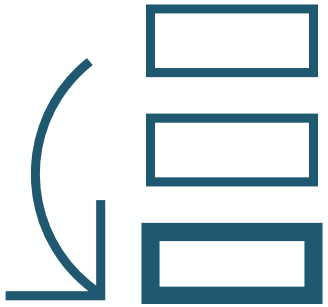
Follow the **Documentation**



Human oversight

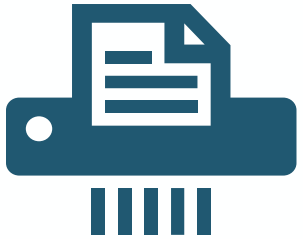


Data and AI Governance to ensure trust



Proper **logging** and **monitoring**

Transparent and clear information to users



Accuracy, **robustness**, and **security**

Sparkle

AI Regulatory Sandboxes

All Member States must provide sandboxes within 24 months after entry into force

Goal: accelerate the go-to-market of innovative AI systems
- Development, Testing, Validation
- Simpler entry to all EU markets

WHAT is a regulatory sandbox?
- A controlled environment
- Relevant datasets (medical, financial)
- Methods and guides for best practices
- Experts with industry knowledge providing supervision

Sparkle

Sparkle’s role

Sparkle

Our most popular solutions

EU AI Act Readiness Assessment

- Risk scoring
- Register framework
- Compliance roadmap
- Stakeholder report

AI Compliance as a Service

- Coaching
- Compliance implementation
- FRIA (fundamentals rights impact assessments)
- System logging
- Risk report & rating
- AI Compliance Officer

AI Compliance Officer Training

- 7 day training (online or in person)
- Become an AI Compliance Officer
- Get trained on all AI Compliance topics based on the EU AI Act

Implementation Ethical Framework & Ethical Board

- Ethical framework tailored to the organization
- Design and organisation of ethical board
- Clear roles & responsibilities

AI Workshop for C-level and board members

- Workshop
- Gain AI knowledge
- Draft AI strategy roadmap
- Description of principles and assumptions

Compliance process

WHAT AI SYSTEMS ARE YOU USING?

Collect information on the AI systems you're using, your processes and policies, your technical capabilities, and how your organization is structured.

WHAT RISKS DO YOUR SYSTEMS POSE?

Analyze the risks your AI systems pose and how they will be classified under the AI Act.

HOW ARE YOU MANAGING RISK?

Ensure you're ready to handle the risks posed by your AI systems. Review your processes, policies, and technical infrastructure to make sure you're compliant with the regulation.

ARE YOUR HIGH-RISK SYSTEMS READY?

Perform a conformity assessment on your high-risk systems and declare conformity if they pass.

HOW ARE YOU TRACKING PERFORMANCE?

After putting your AI into service, set up a monitoring system to check the ongoing functionality of your systems.

Sparkle

Readiness assessment

WHY IS COMPLIANCE SO IMPORTANT?

We get your management & stakeholders up to speed on what AI is, why it's often risky, and why compliance is of vital importance.

WHAT AI SYSTEMS ARE YOU USING?

We collect information on the AI systems you're using, your processes and policies, your technical capabilities, and how your organization is structured.

WHAT RISKS DO YOUR SYSTEMS POSE?

We analyze the risks your AI systems pose and how they will be classified under the AI Act. This will determine how intensive your compliance track will be.

HOW ARE YOU MANAGING RISK?

We scope out how ready your organization is to handle the risks posed by your AI, focusing on operations (processes and policies), technology, and organizational structures.

HOW CAN YOU DO (EVEN) BETTER?

We deliver a report on your risk rating & maturity score. We also tailor-make a detailed roadmap to improve your AI Act readiness.

Sparkle

Recap

EU AI Act – the 4 W's

WHAT is the focus of AI Act?

- Human-centric AI
- Risk-based classification of AI systems
- Trustworthy, non-discriminating solutions

WHY do companies need to comply with the AI Act?

- Organizations are using AI systems
  - *The need for trust and transparency is growing*
- AI compliance provides framework that can accelerate innovation safely
- Fines and penalties

Sparkle

EU AI Act – the 4 W's

WHO must comply with the AI Act?

- Providers of AI systems;
- Distributors of AI systems;
- Importers of AI systems;
- Deployers of AI systems;
- Any third parties.

WHEN will the EU AI Act apply?

- Final approval expected Q2 2024
- Time to act:
  - Unacceptable: 6 mo
  - Penalties & GPAI requirements: 12 mo
  - High risk obligations under Annex III: 24 mo
  - High risk obligations under Annex II: 36 mo

Sparkle

Penalties for non-compliance

Non-compliance with prohibitions:
Up to €35M or
7% of global AT

Non-compliance with other obligations:
Up to €15M or
3% of global AT

Supplying incorrect or incomplete information:
Up to €7.5M or
1% of global AT

For **SMEs**:
whichever of the two amounts is **LOWER**

Final amount depends on circumstances of incident

Sparkle

Limiting or accelerating?

Limiting:

- Higher costs (human oversight, reporting, monitoring…)
- More bureaucracy might lead to a longer time-to-market

Accelerating:

- Comprehensive AI framework
- Improved trust, e.g. investors and enterprises
- Capturing a single country's market (in the EU) allows scaling to the whole EU
- Regulatory Sandboxes

All-in-all:

- Compliance is mandatory for most risk categories
- Obligations are shared between providers (e.g. model documentation) and deployers (e.g. human oversight)
- Obligations for generative AI lie primarily with providers of GPAI, not deployers (downstream)

Sparkle

Sparkle

Contact

Martin Karu

Data Expert

+372 5662 4031

martin.karu@sparkle.consulting

https://sparkle.consulting