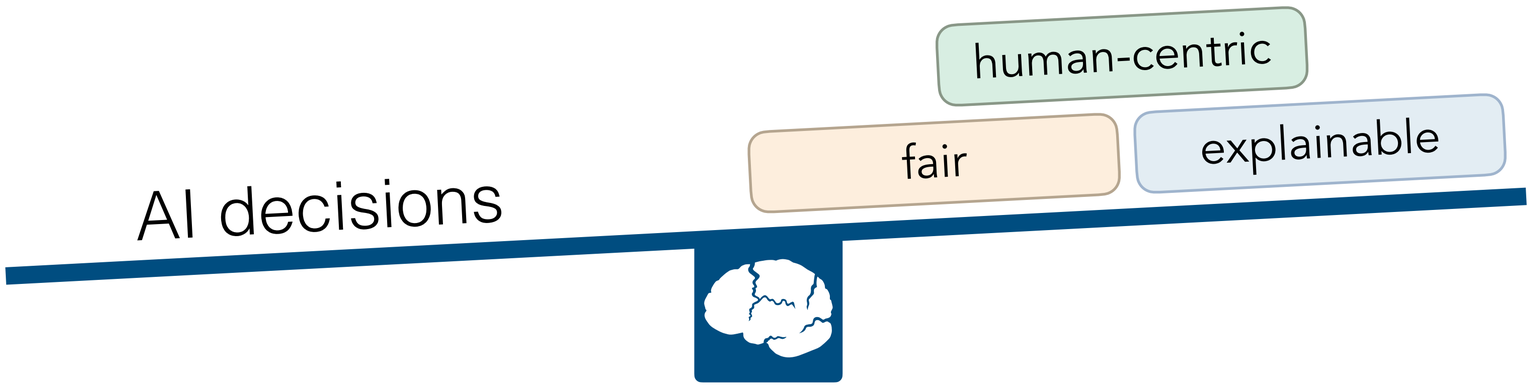
A Brief Account of Explainability, Interpretability, and Verifiability in AI

What logic and formal methods can offer to the regulation of AI



Benedikt Bollig

Université Paris-Saclay, CNRS, ENS Paris-Saclay, LMF, Gif-sur-Yvette, France

Credits

- Based on a talk at the GT DAAL Annual Meeting, May 2025, LIGM, Champs-sur-Marne.
- Please send comments or questions to lastname@lmf.cnrs.fr
- Version as of: May 2025.

Outline of talk

behavioral/functional		structural/epistemic		normative	
fair	robust	explainable	interpretable	ethical	human-centric
safe	optimal	verifiable	transparent	trustworthy	accountable

what is computed? — *how is it computed?* — *why does it matter?*

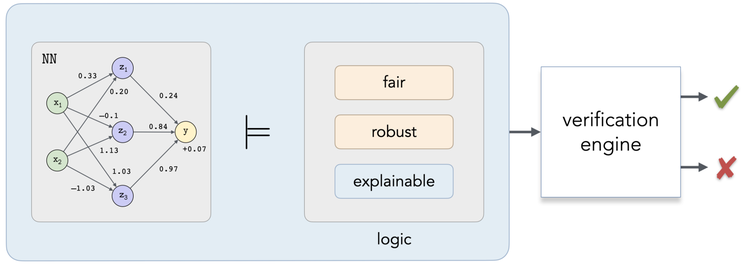
classification of (AI) systems

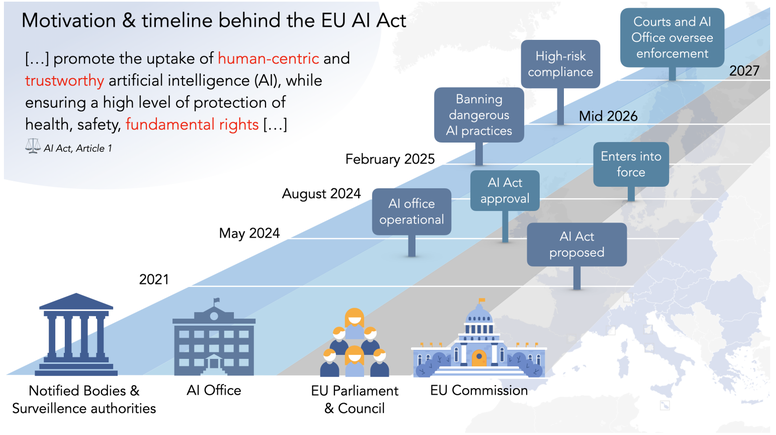
Such properties are (directly or indirectly) reflected in the EU AI Act.

explainable	Decisions are human-comprehensible.
interpretable	The functioning of a model can be understood.
transparent	Internals and design are accessible.
verifiable	Formally provable against specifications.

ethical	Aligned with human values, rights, and societal norms.
trustworthy	Consistently reliable, safe, and worthy of confidence.
accountable	Responsibility is clear and traceable for system outcomes.
human-centric	Puts people in control and respects their rights.

Verifiability

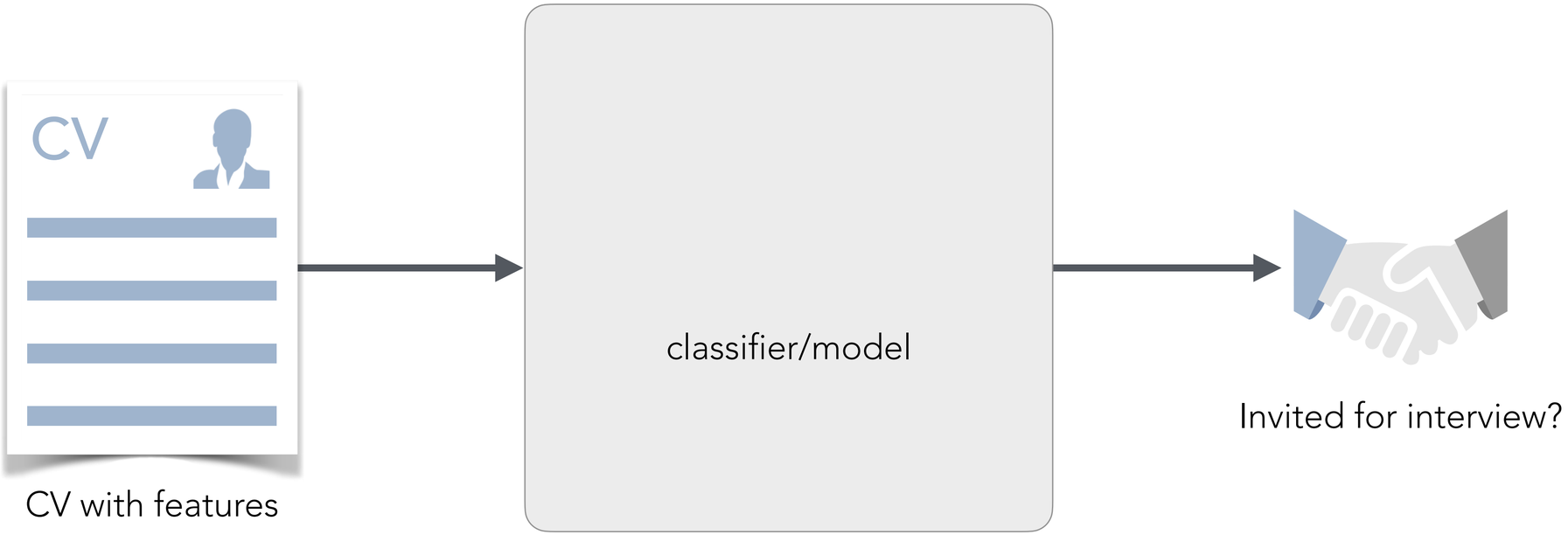




Classification of AI systems → Logic & Formal Methods → AI Act

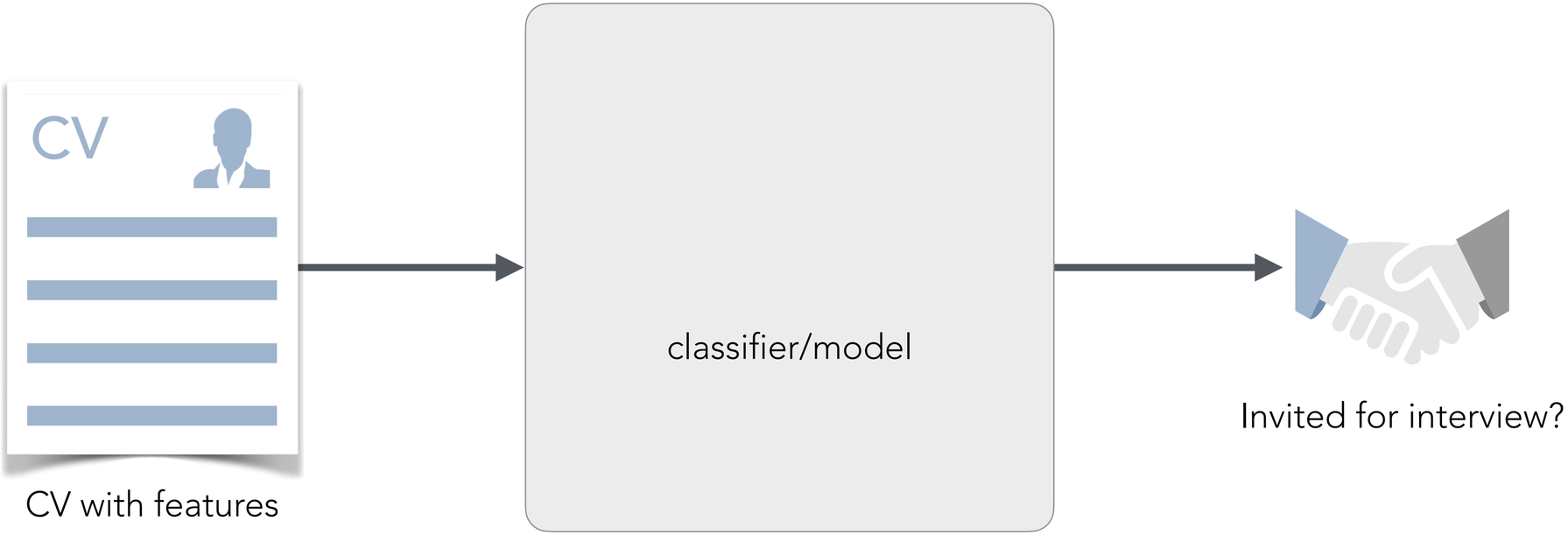
Towards a Classification of (AI) Systems

fair	robust	explainable	interpretable	ethical	trustworthy
safe	optimal	transparent	verifiable	accountable	human-centric


CV
CV with features
classifier/model
Invited for interview?

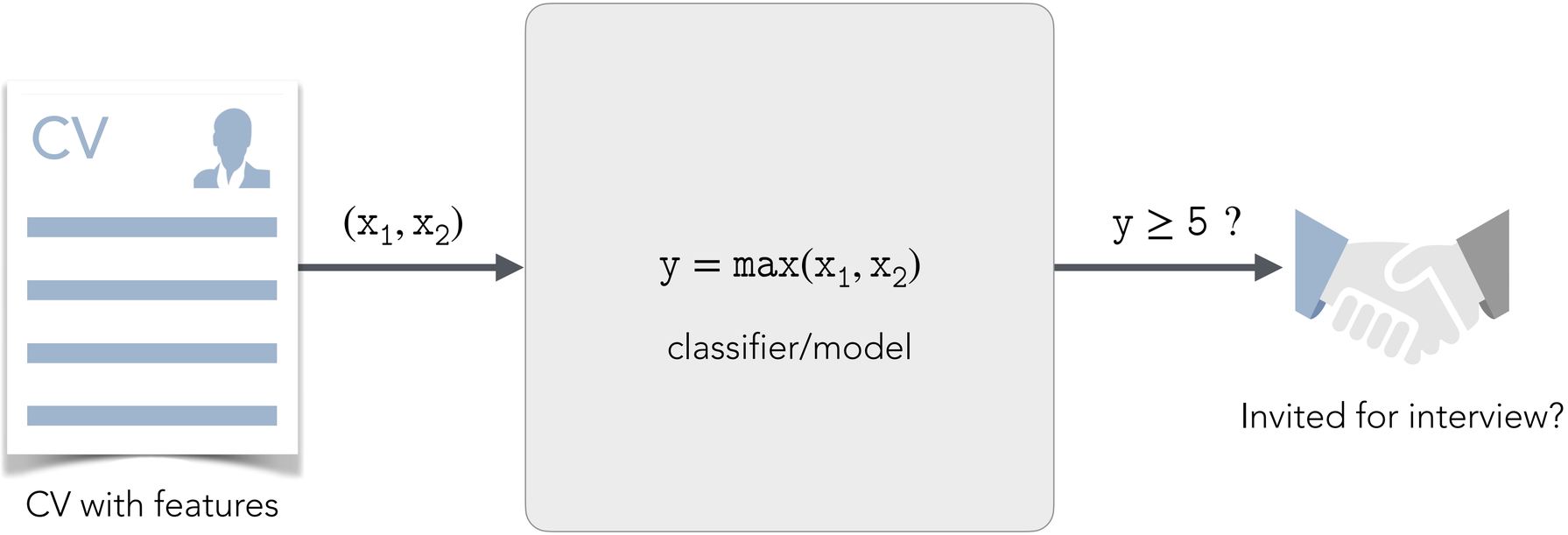
fair

Is classification independent of sensitive features?



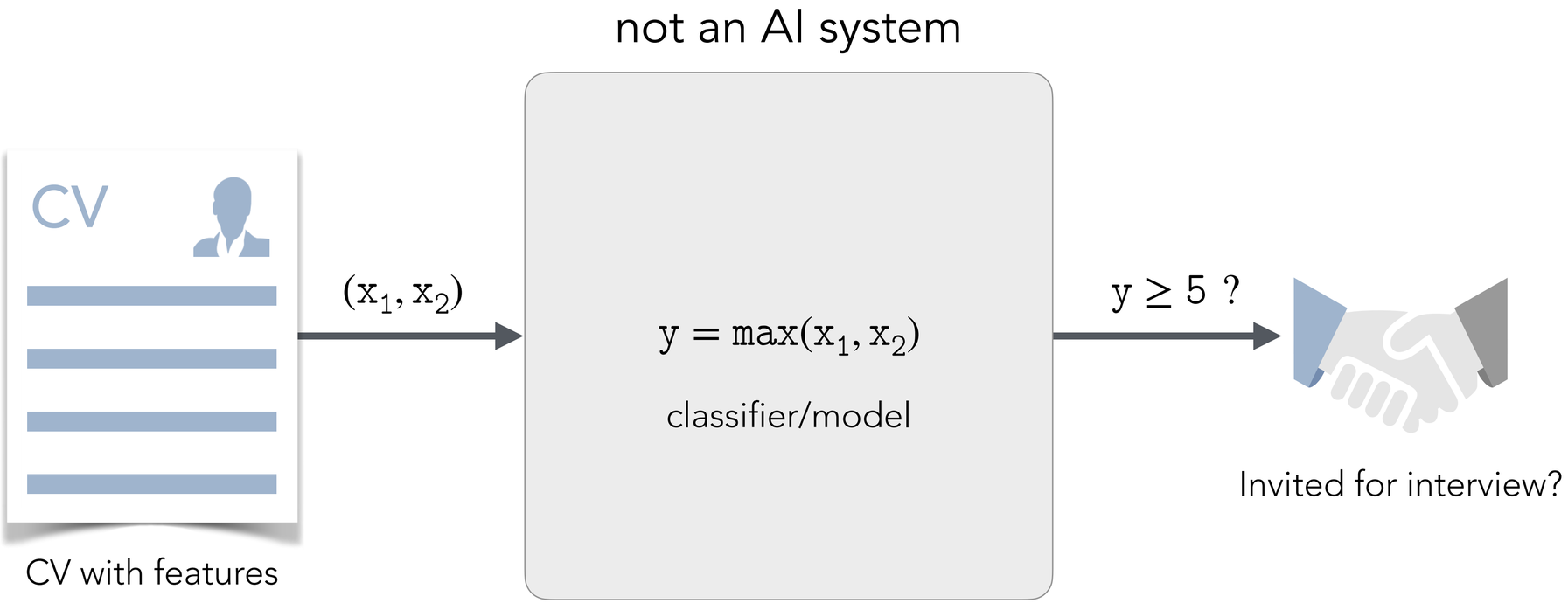
fair

Is classification independent of sensitive features?



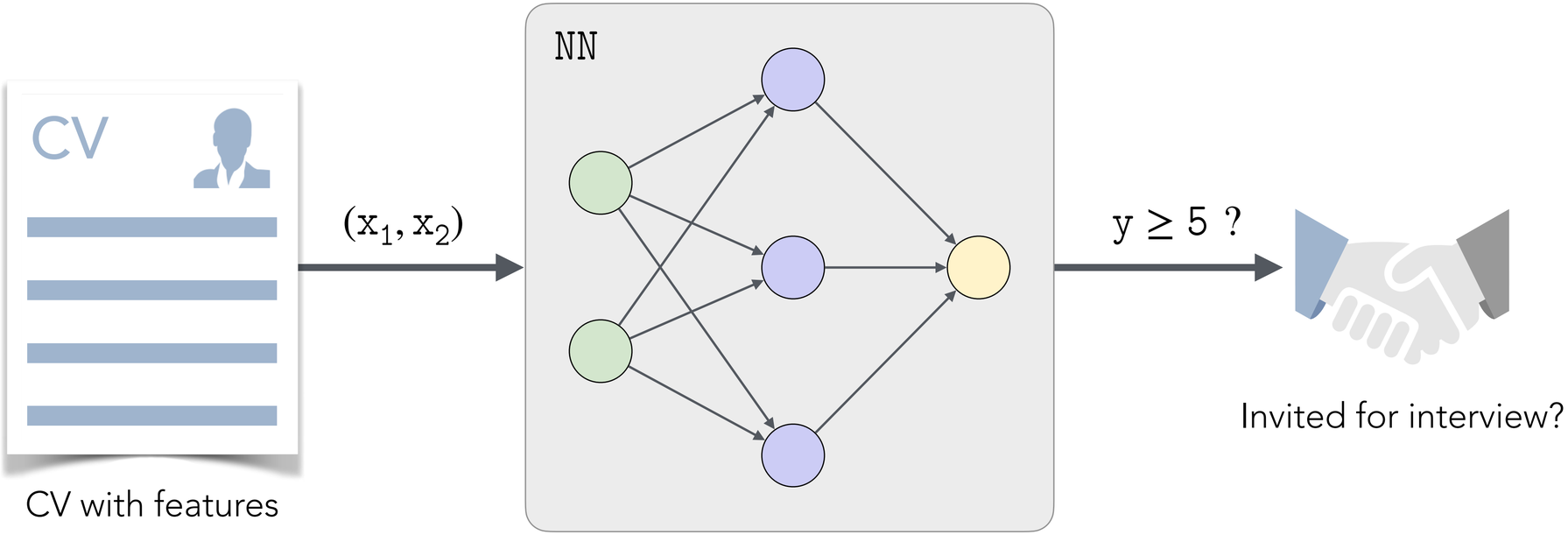
fair

Is classification independent of sensitive features?



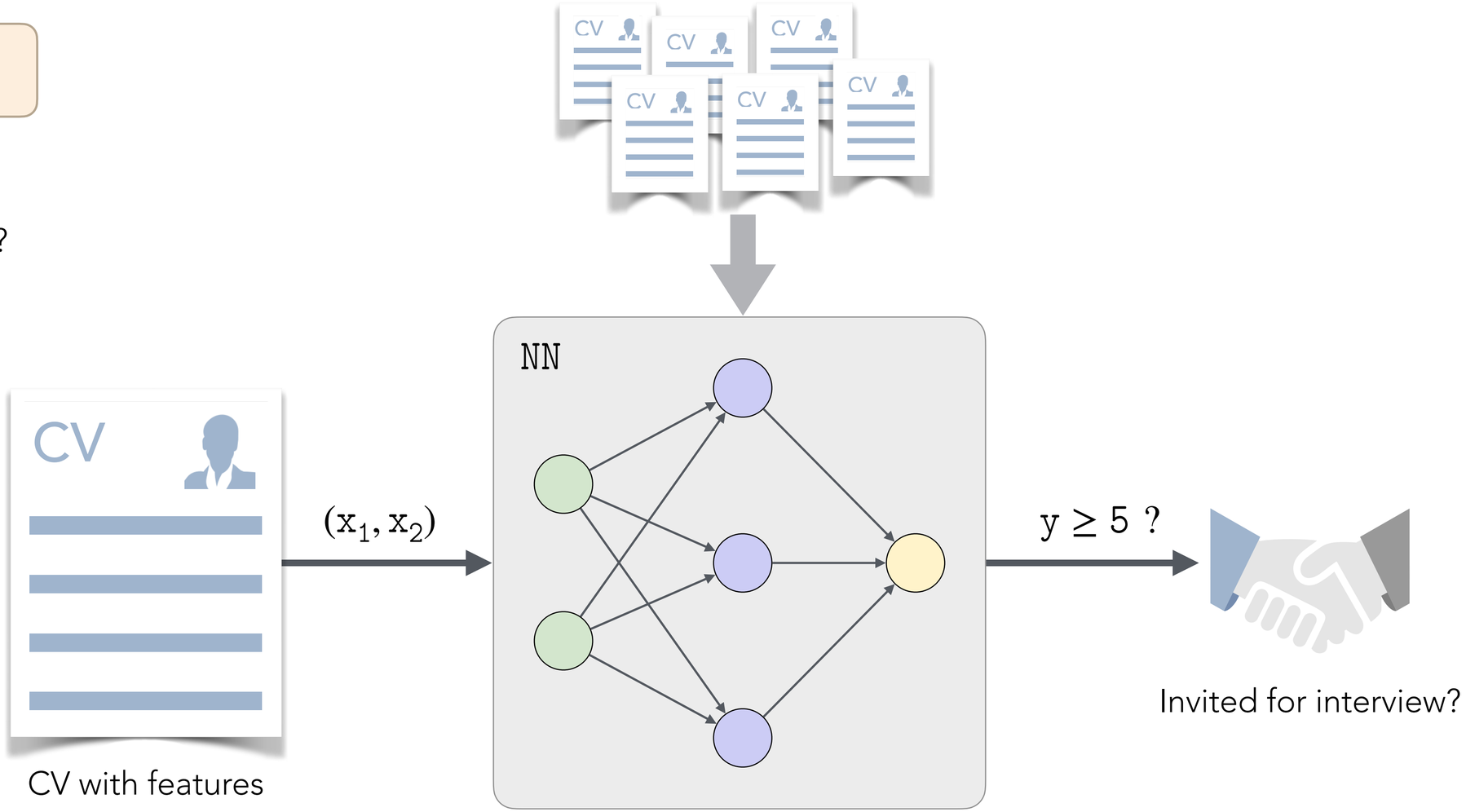
fair

Is classification independent of sensitive features?



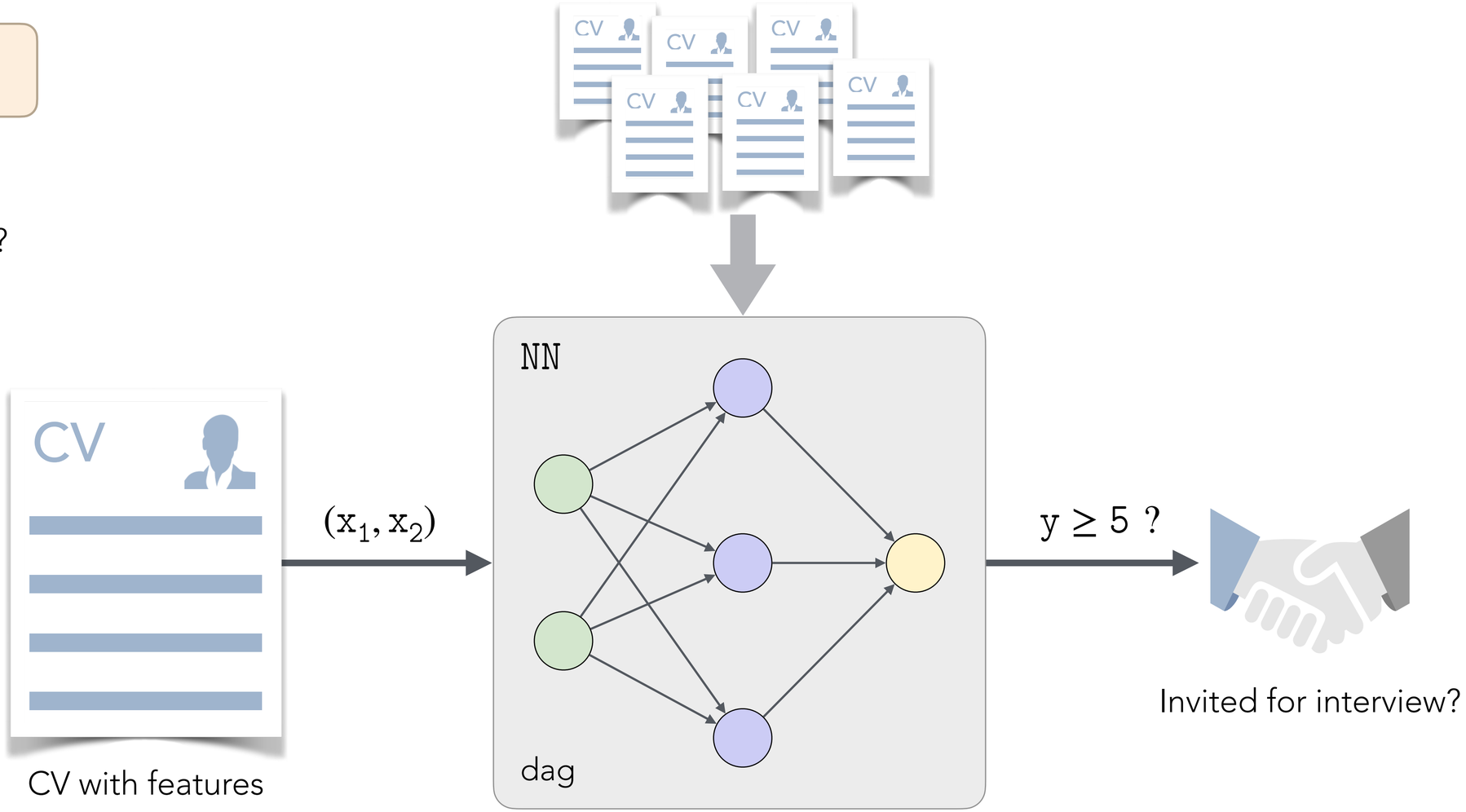
fair

Is classification independent of sensitive features?



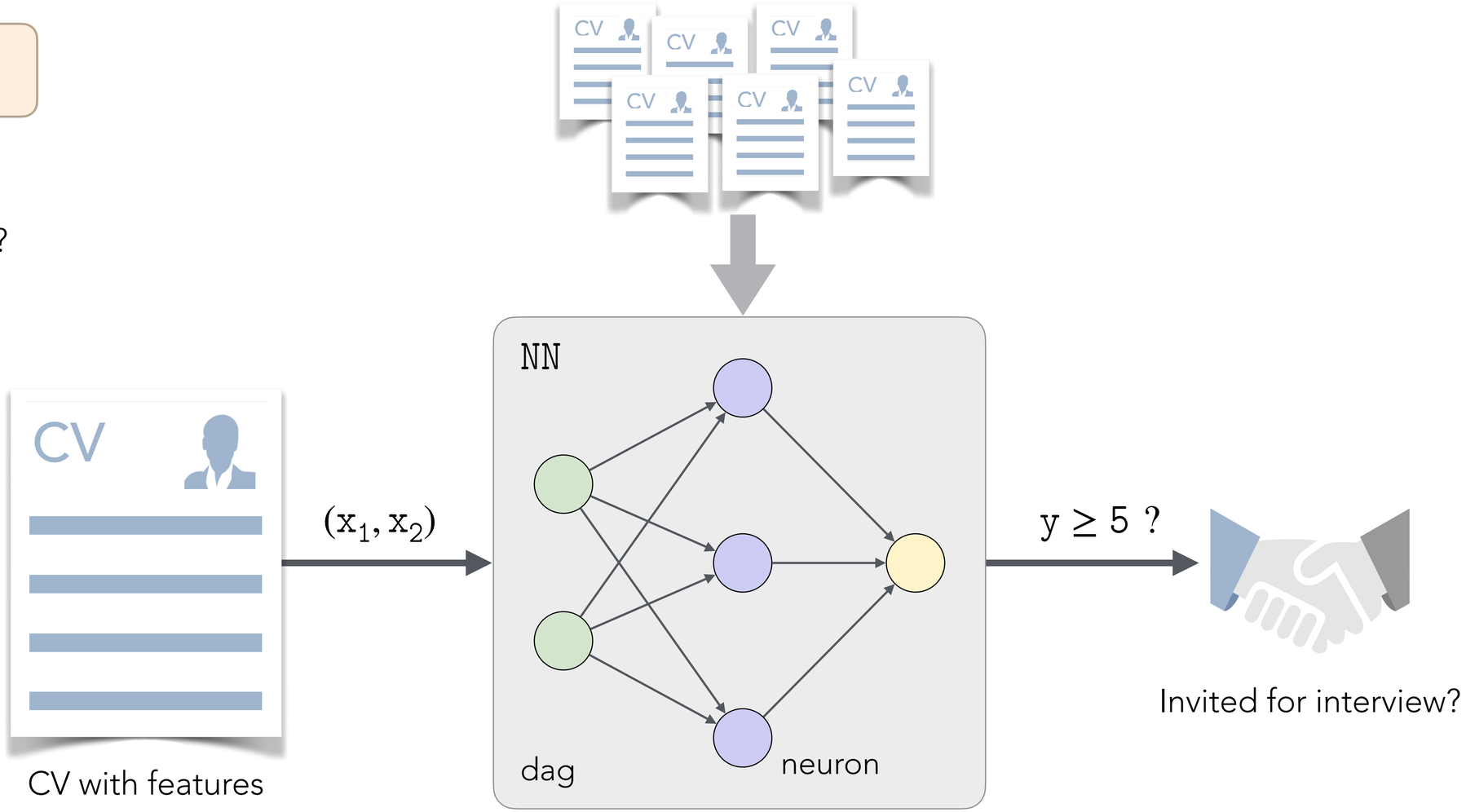
fair

Is classification independent of sensitive features?



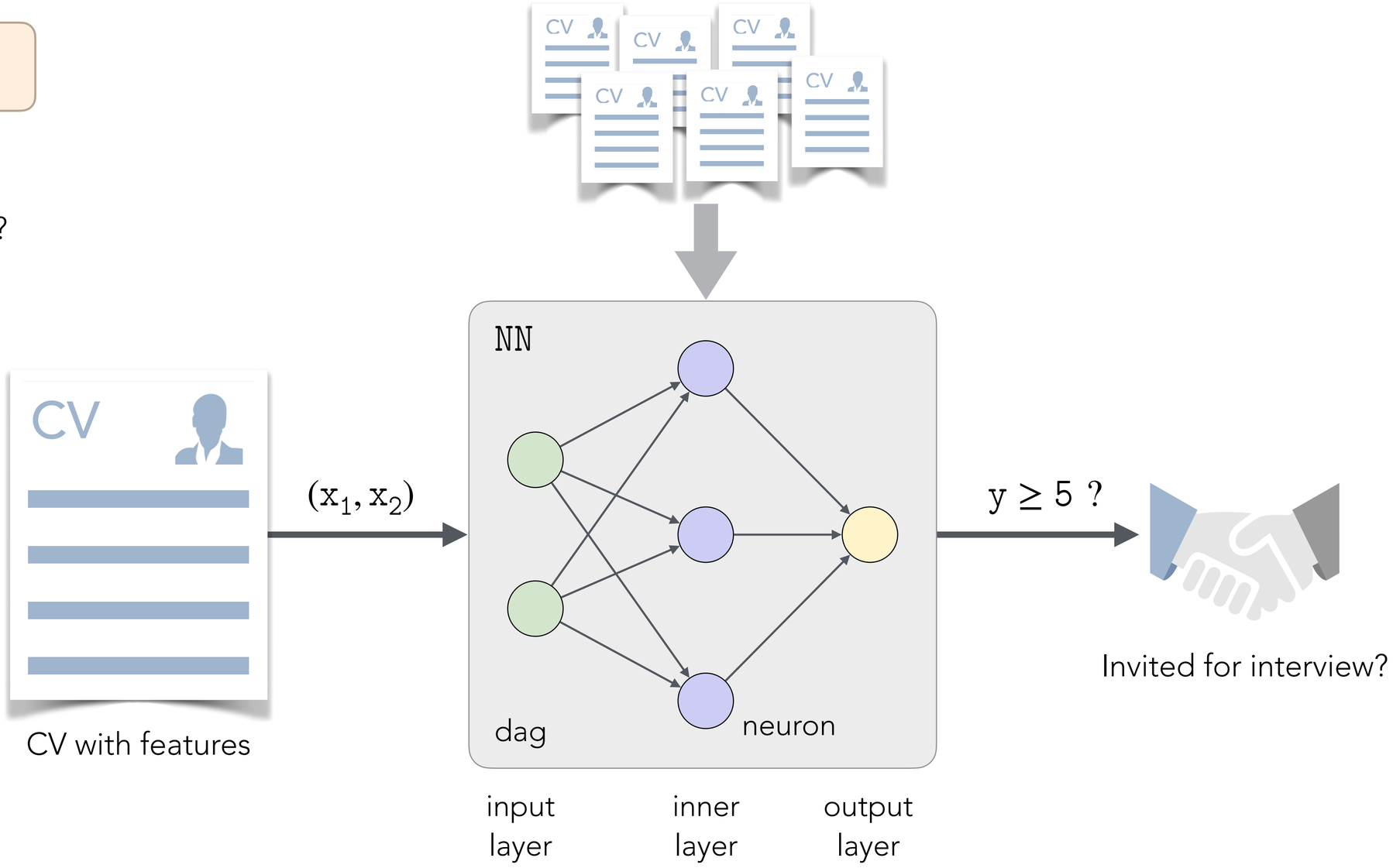
fair

Is classification independent of sensitive features?



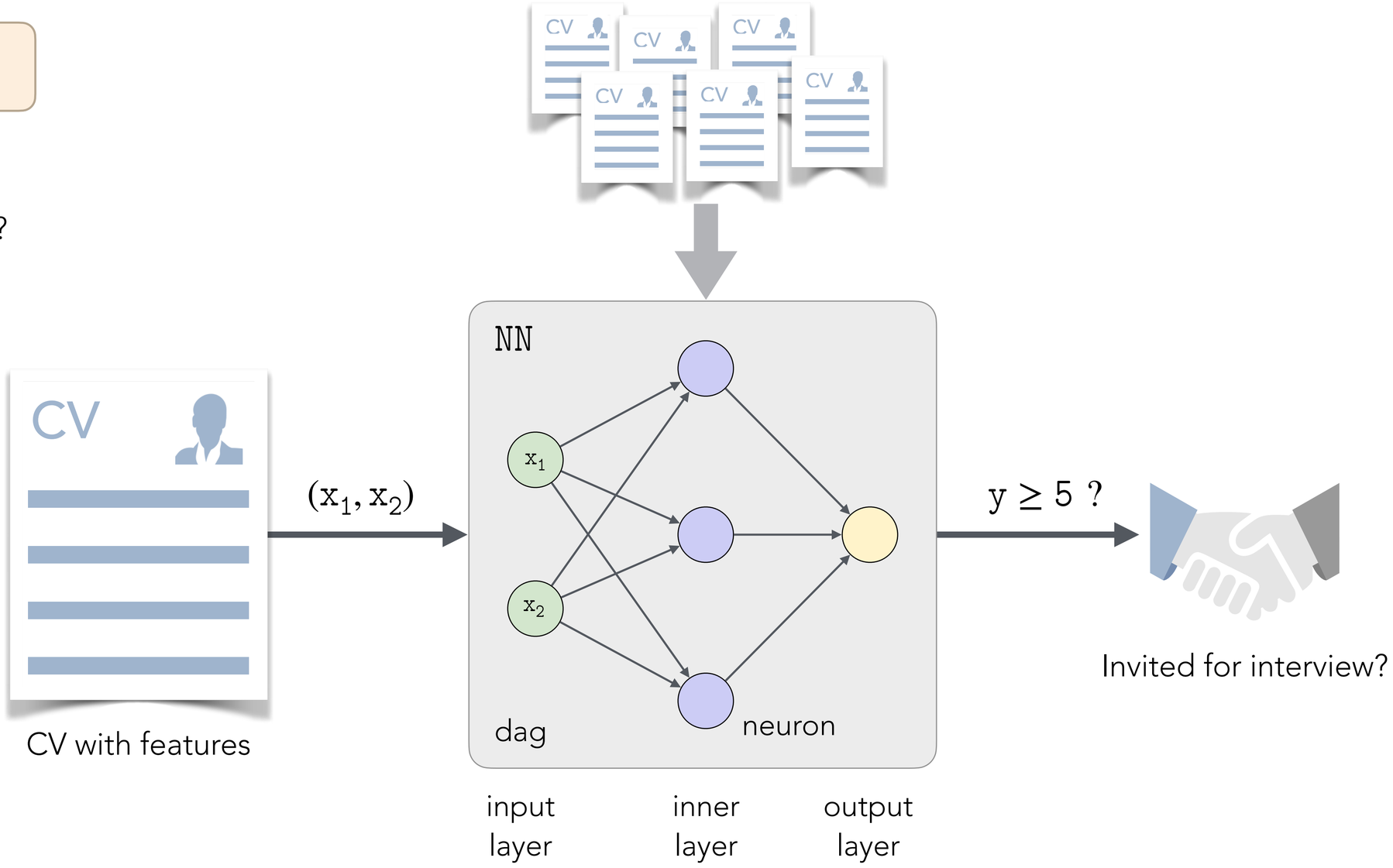
fair

Is classification independent of sensitive features?



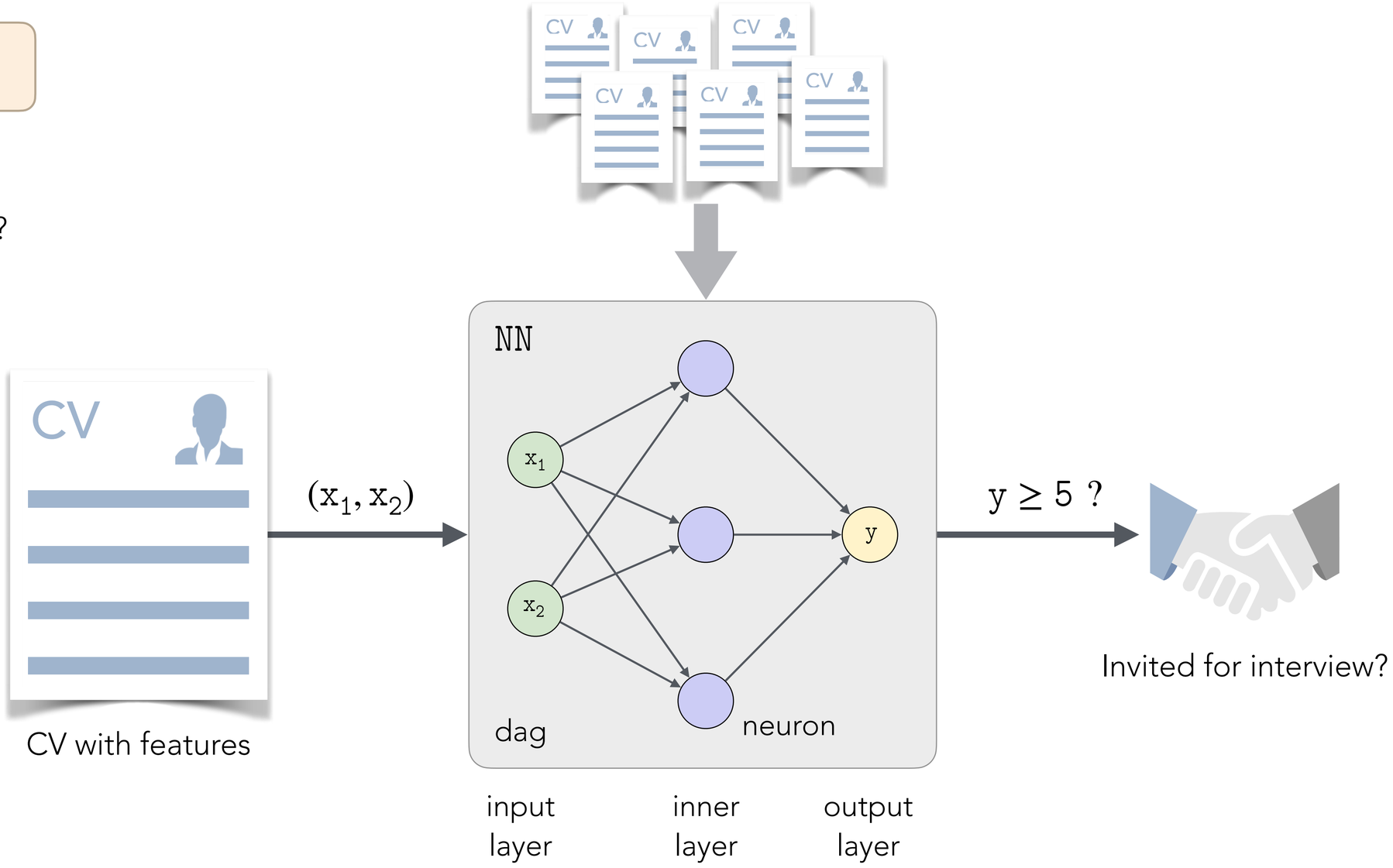
fair

Is classification independent of sensitive features?



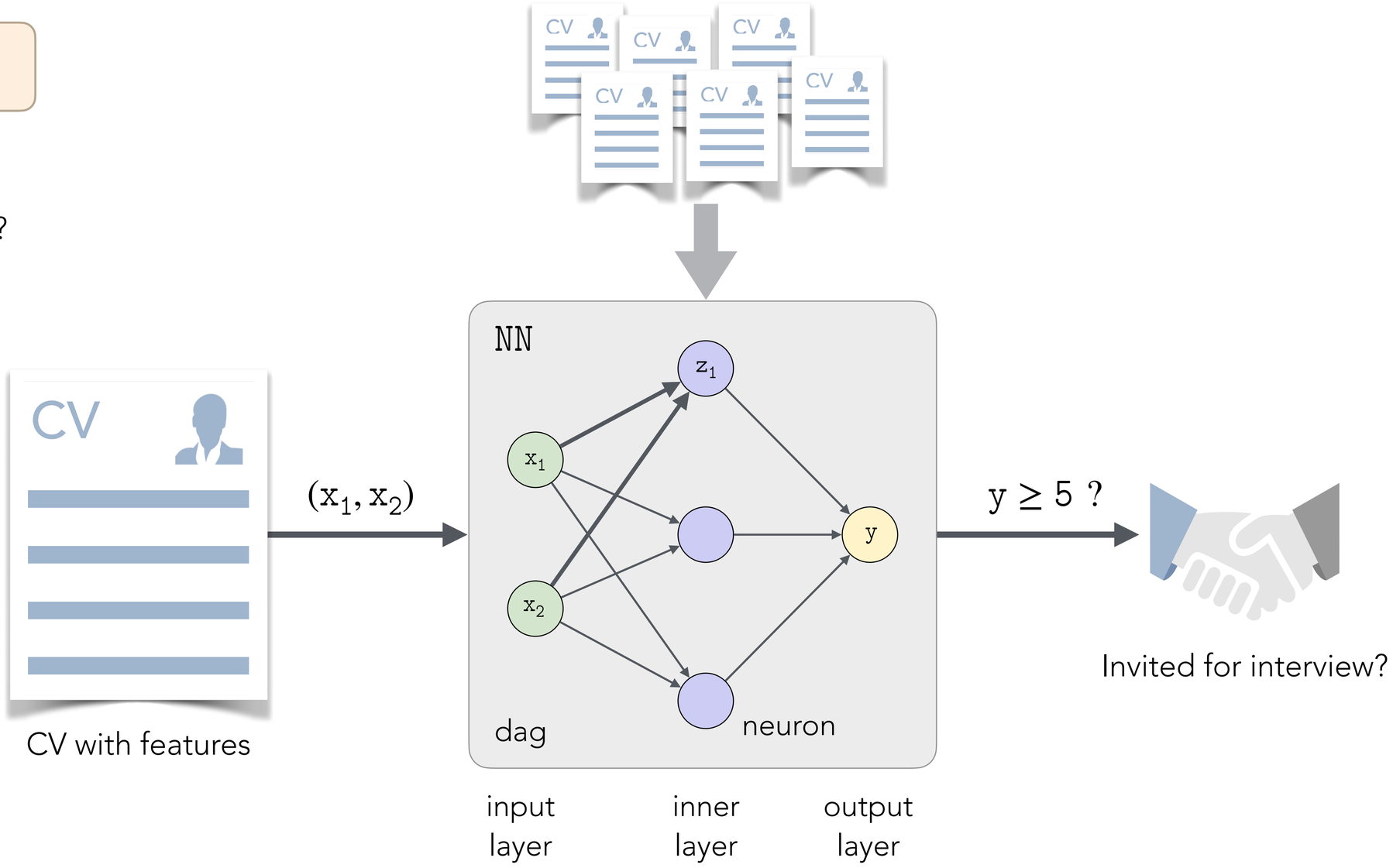
fair

Is classification independent of sensitive features?



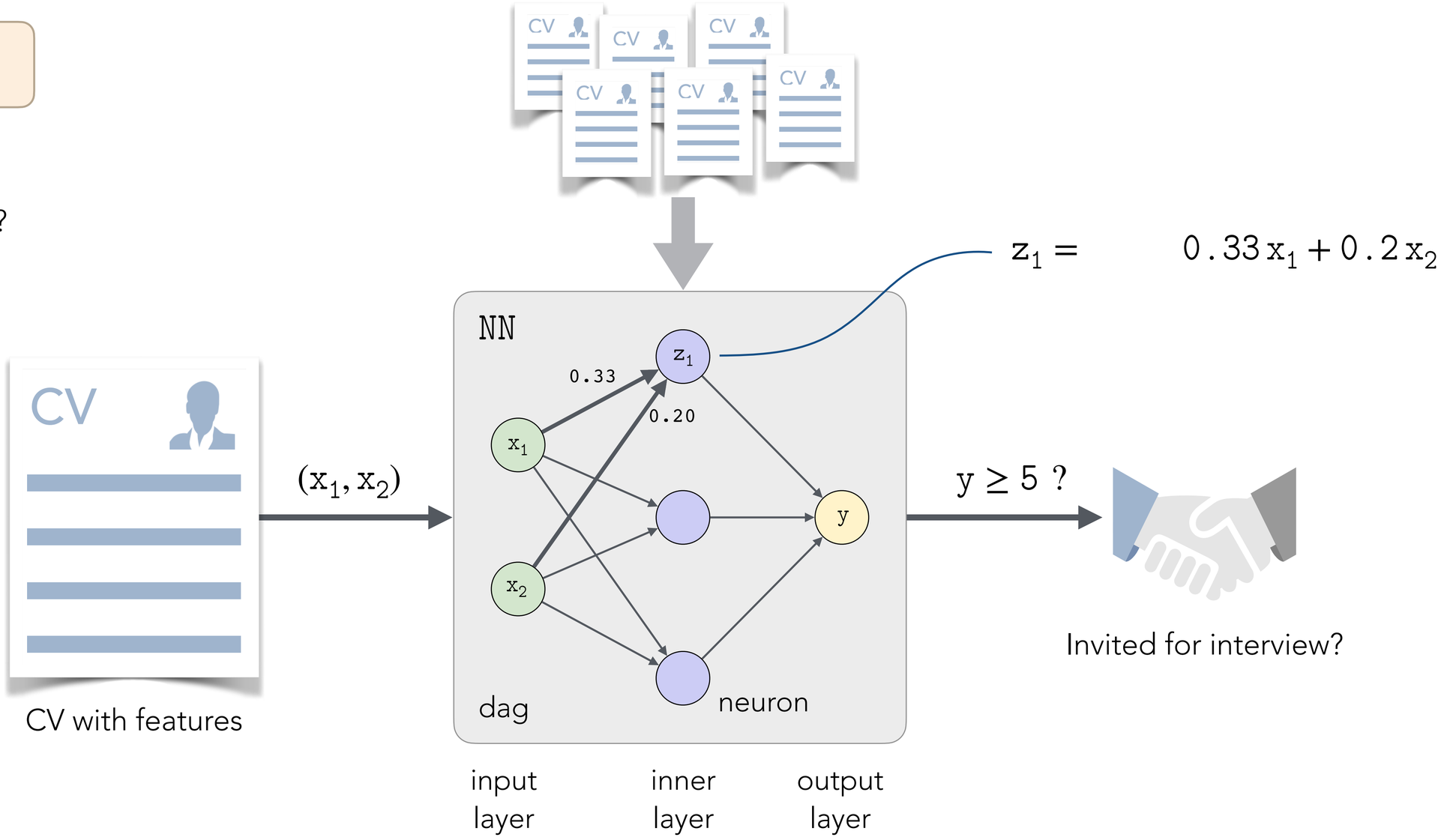
fair

Is classification independent of sensitive features?



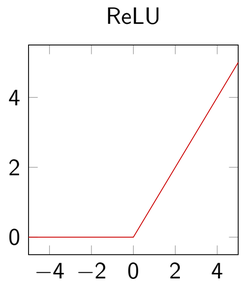
fair

Is classification independent of sensitive features?



fair

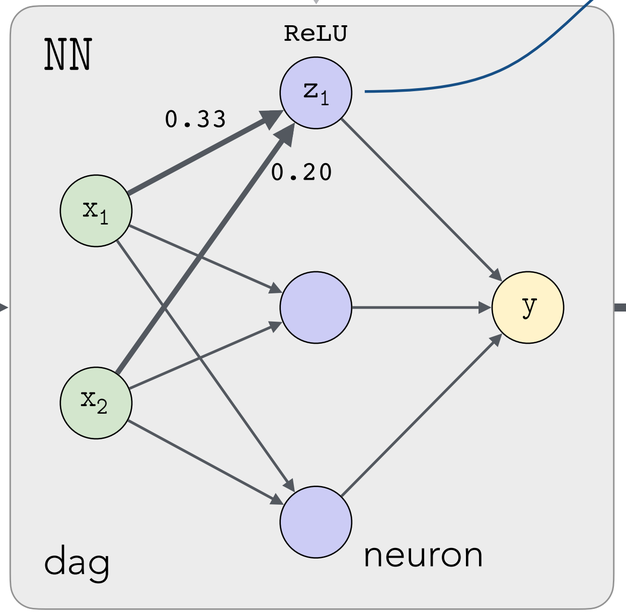
Is classification independent of sensitive features?



$$z_1 = \text{ReLU}(0.33\,x_1 + 0.2\,x_2)$$

CV with features

(x_1, x_2)



input layer

inner layer

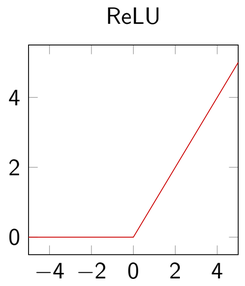
output layer

$y \geq 5$?

Invited for interview?

fair

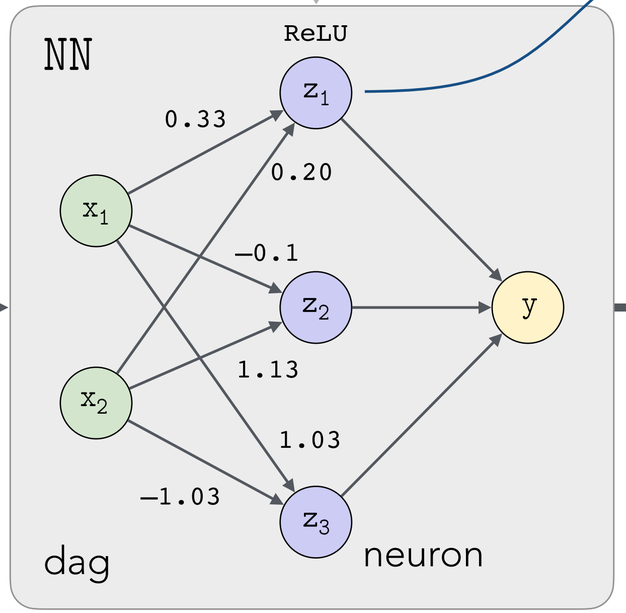
Is classification independent of sensitive features?



$$z_1 = \text{ReLU}(0.33\,x_1 + 0.2\,x_2)$$

CV with features

(x_1, x_2)



input layer

inner layer

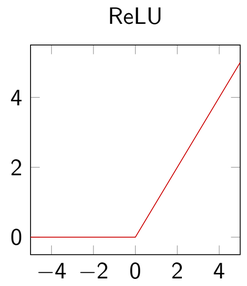
output layer

$y \geq 5$?

Invited for interview?

fair

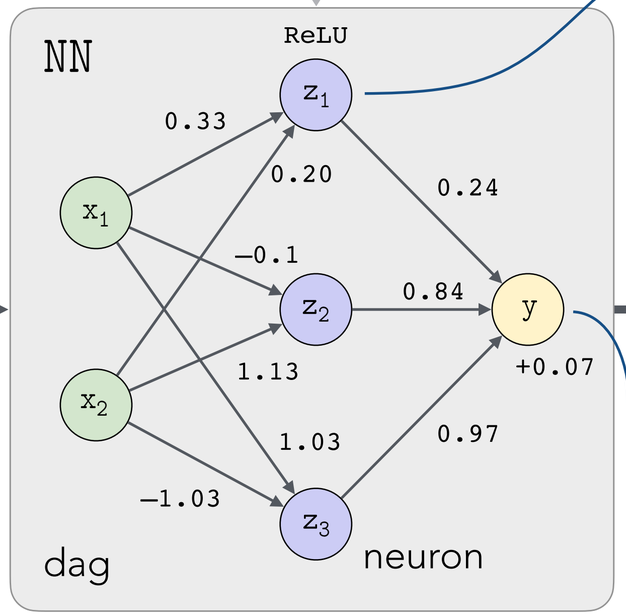
Is classification independent of sensitive features?



$$z_1 = \text{ReLU}(0.33\,x_1 + 0.2\,x_2)$$

CV with features

(x_1, x_2)



input layer

inner layer

output layer

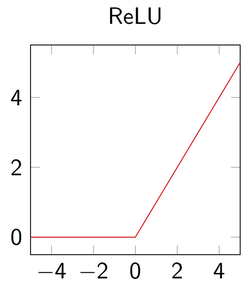
$y \geq 5$?

Invited for interview?

$$y = 0.24\,z_1 + 0.84\,z_2 + 0.97\,z_3 + 0.07$$

fair

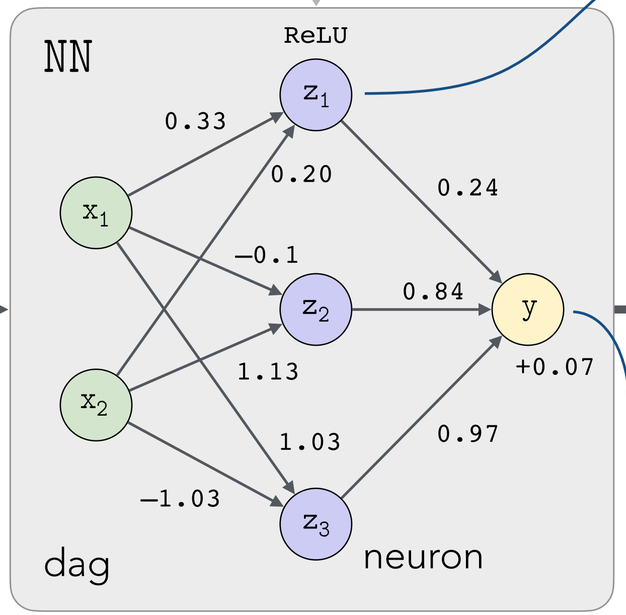
Is classification independent of sensitive features?



$$z_1 = \text{ReLU}(0.33\,x_1 + 0.2\,x_2)$$

CV with features

(x_1, x_2)



input layer

inner layer

output layer

$y \geq 5$?

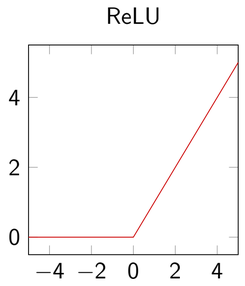
Invited for interview?

$$y = 0.24\,z_1 + 0.84\,z_2 + 0.97\,z_3 + 0.07$$

$$\text{NN}(3, 2) = 3.049$$

fair

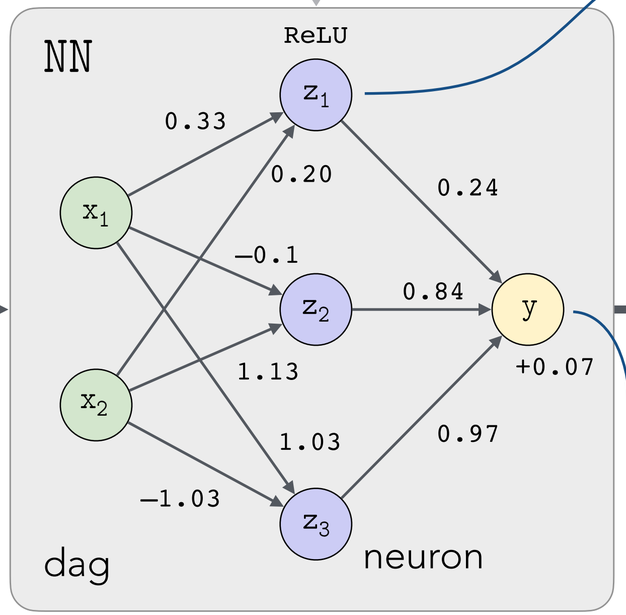
Is classification independent of sensitive features?



$$z_1 = \text{ReLU}(0.33\,x_1 + 0.2\,x_2)$$

CV with features

(x_1, x_2)



input layer

inner layer

output layer

$y \geq 5$?

Invited for interview?

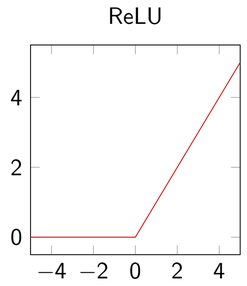
$$y = 0.24\,z_1 + 0.84\,z_2 + 0.97\,z_3 + 0.07$$

$$\text{NN}(3, 2) = 3.049$$

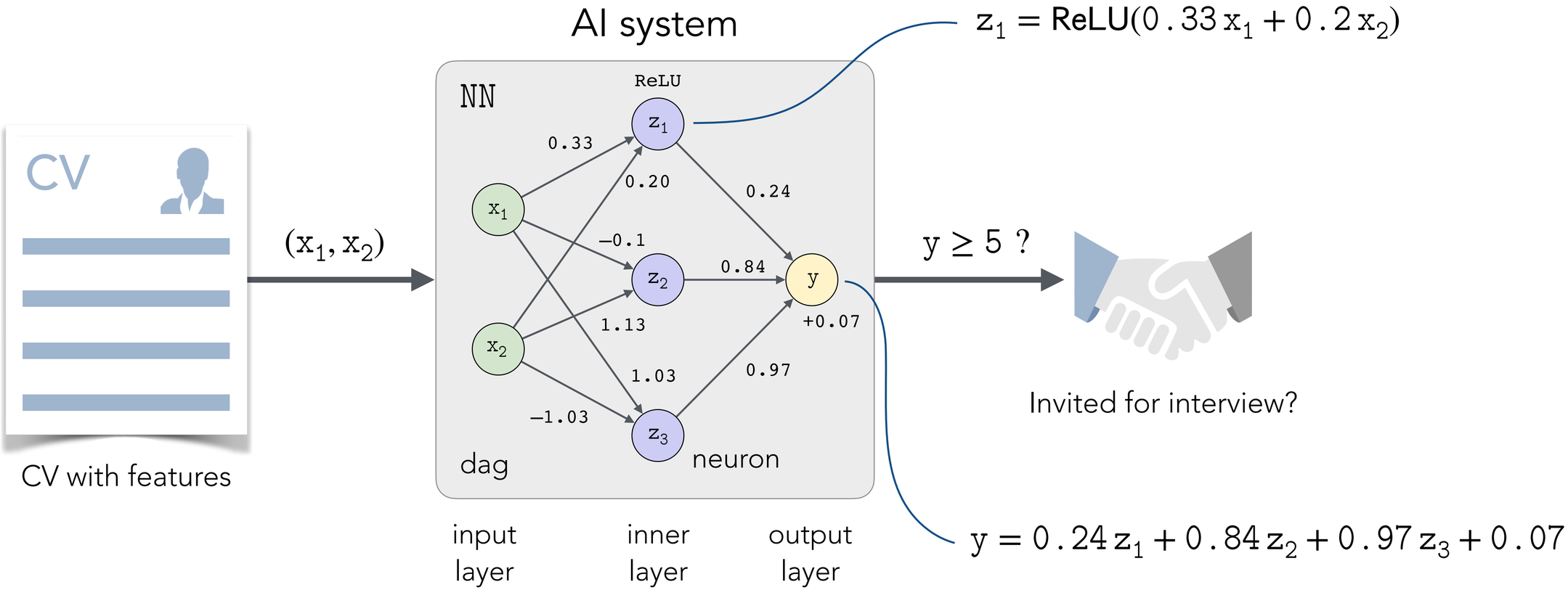
$$\text{NN}(4, 9) = 9.026$$

fair

Is classification independent of sensitive features?



$$z_1 = \text{ReLU}(0.33\,x_1 + 0.2\,x_2)$$



$$y = 0.24\,z_1 + 0.84\,z_2 + 0.97\,z_3 + 0.07$$

$$\text{NN}(3, 2) = 3.049$$

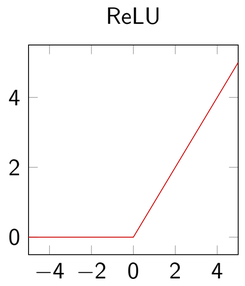
$$\text{NN}(4, 9) = 9.026$$

fair

Is classification independent of sensitive features?

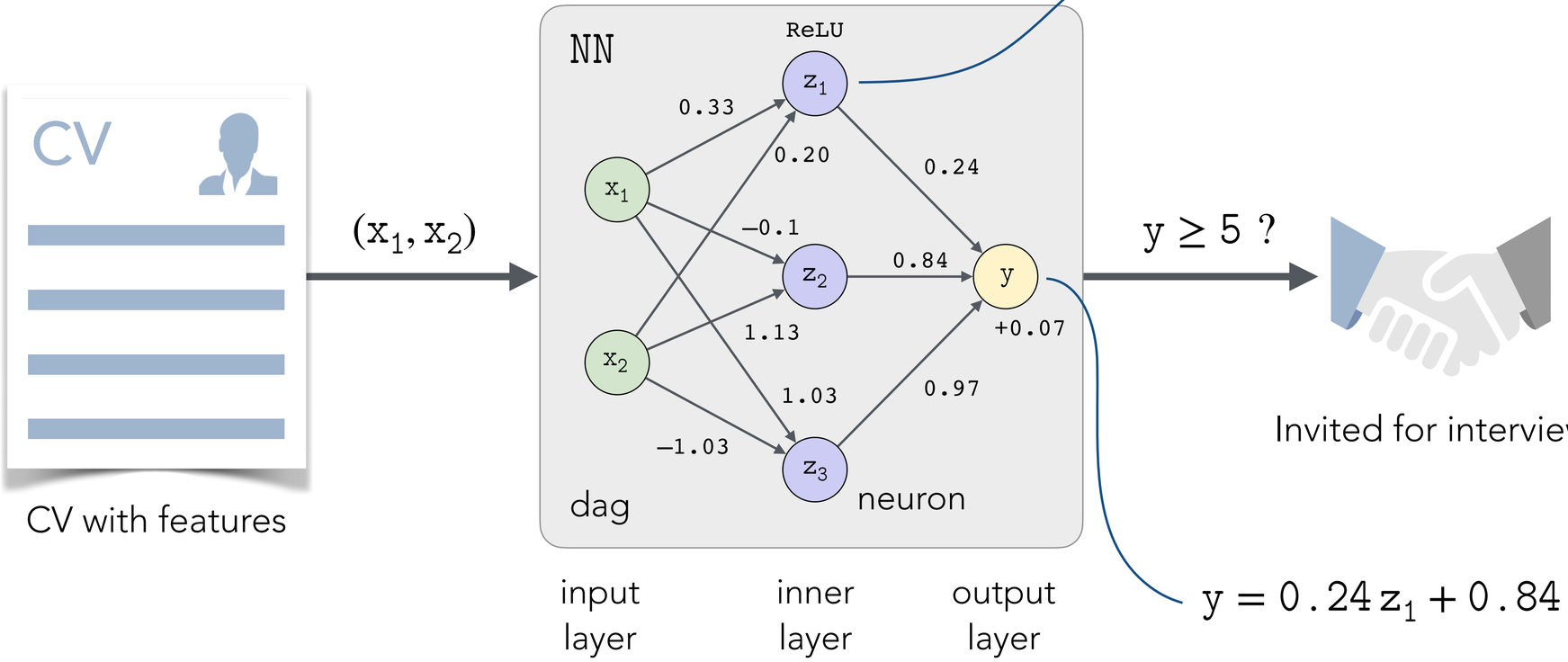
interpretable

Is its inner functioning understandable?



$$z_1 = \text{ReLU}(0.33\,x_1 + 0.2\,x_2)$$

AI system



$$y = 0.24\,z_1 + 0.84\,z_2 + 0.97\,z_3 + 0.07$$

$$\text{NN}(3, 2) = 3.049$$

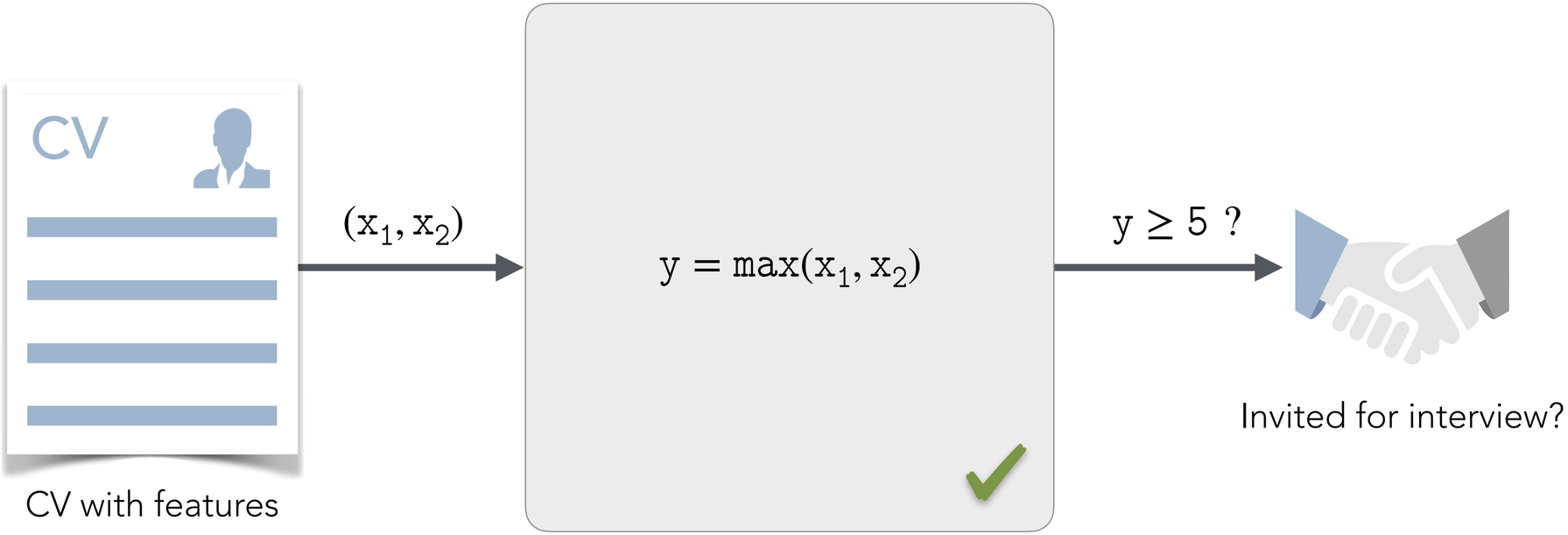
$$\text{NN}(4, 9) = 9.026$$

fair

Is classification independent of sensitive features?

interpretable

Is its inner functioning understandable?

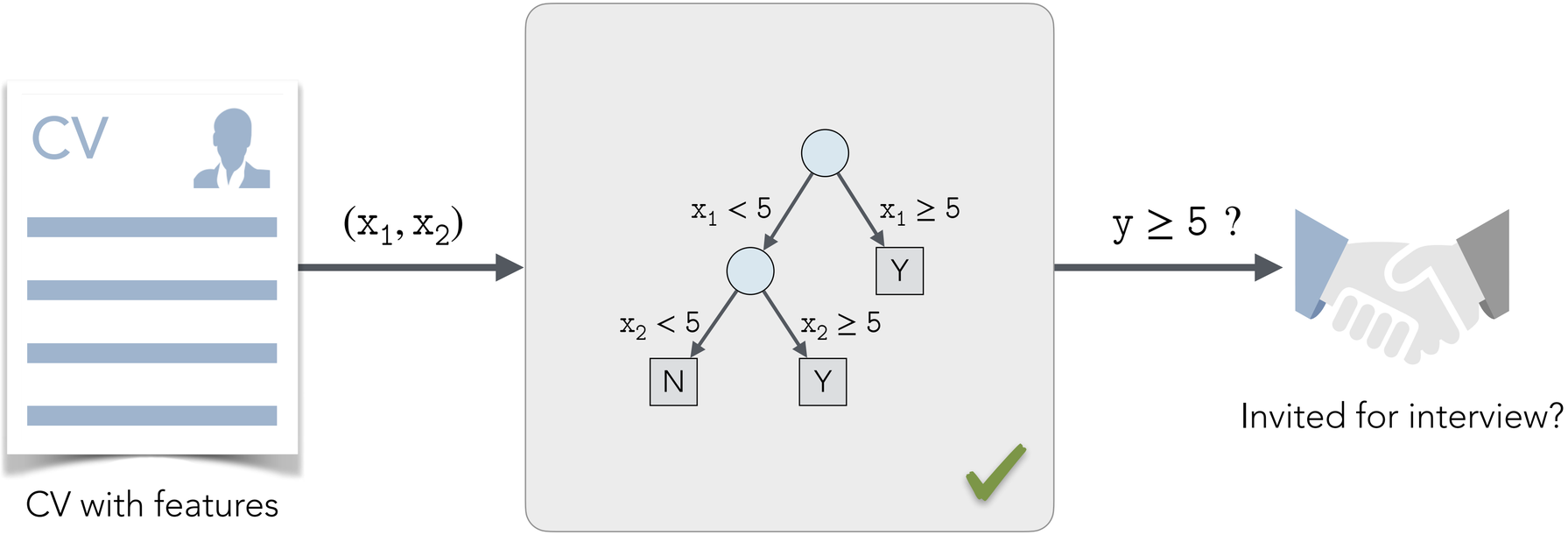


fair

Is classification independent of sensitive features?

interpretable

Is its inner functioning understandable?

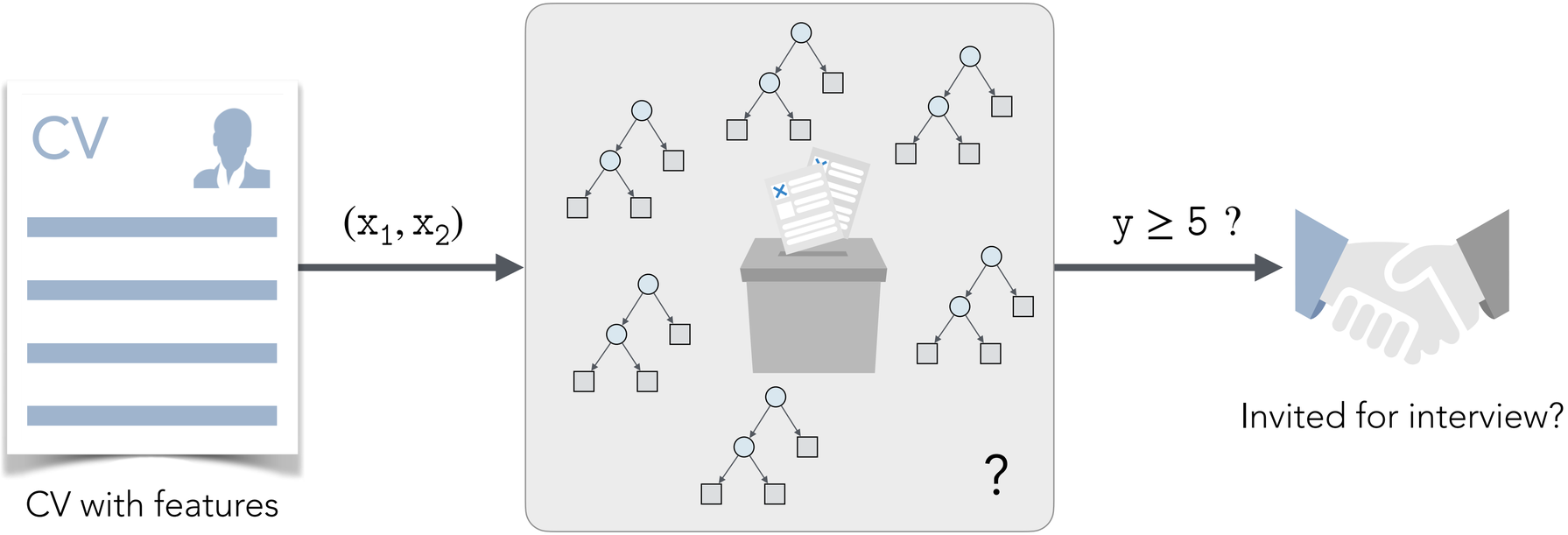


fair

Is classification independent of sensitive features?

interpretable

Is its inner functioning understandable?

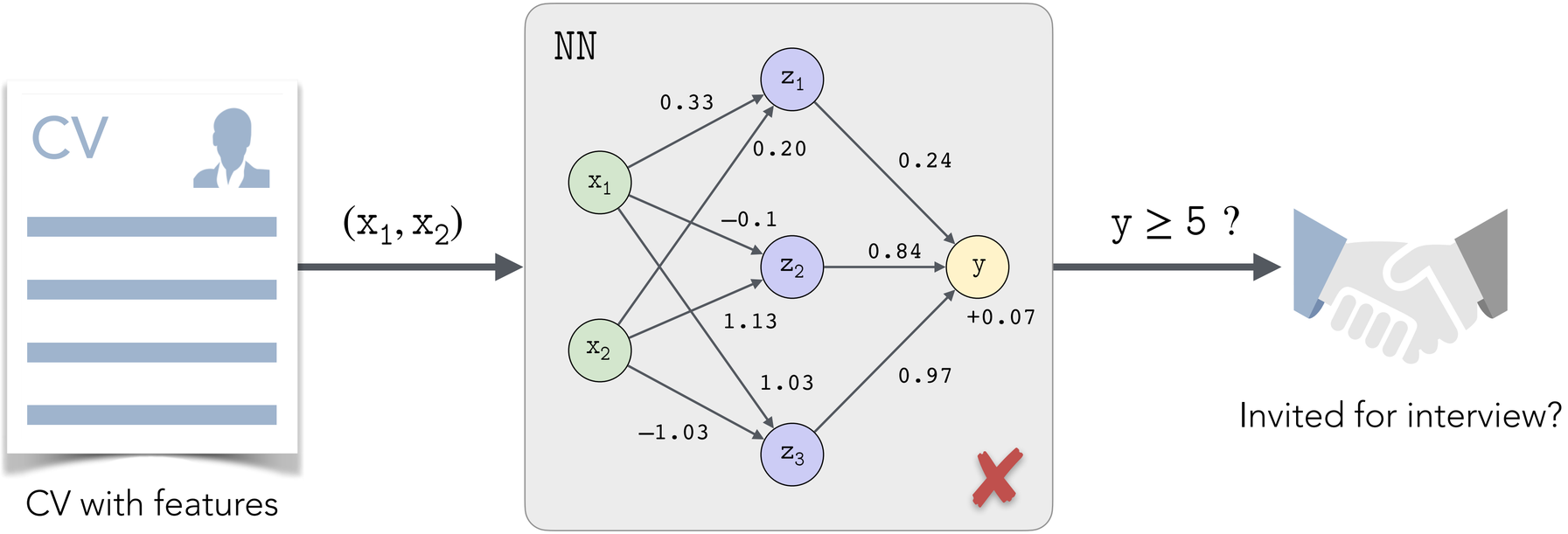


fair

Is classification independent of sensitive features?

interpretable

Is its inner functioning understandable?



fair

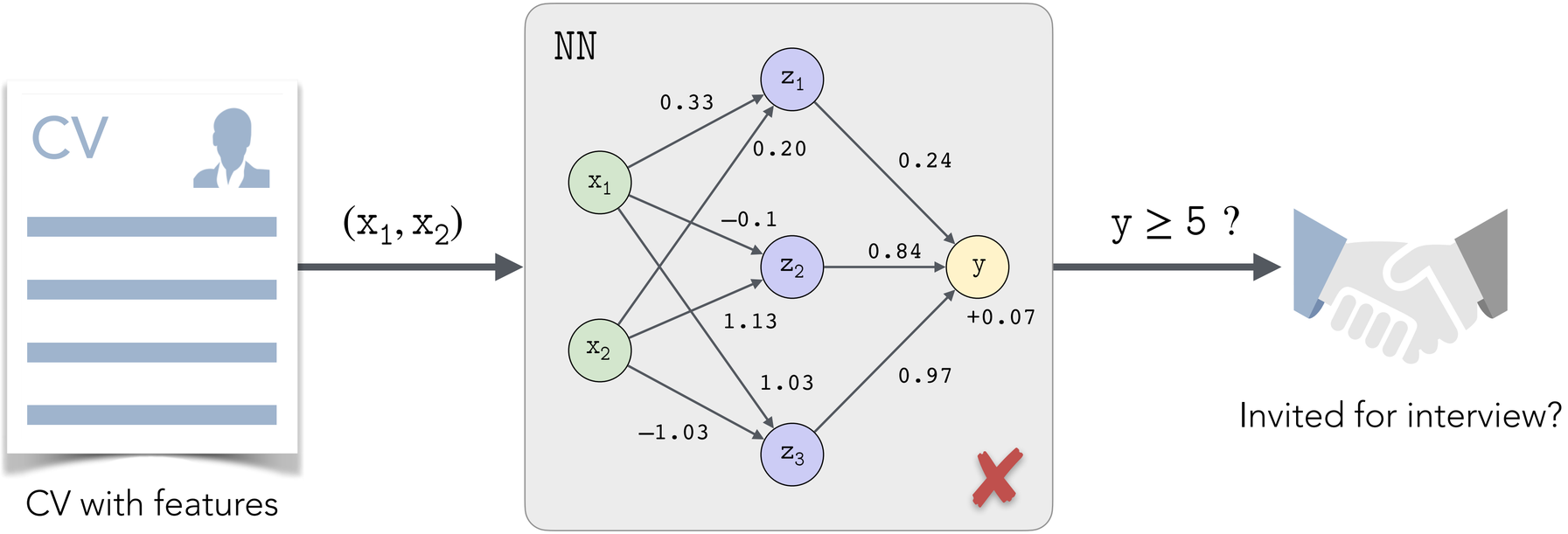
Is classification independent of sensitive features?

explainable

Is decision human-understandable?

interpretable

Is its inner functioning understandable?



fair

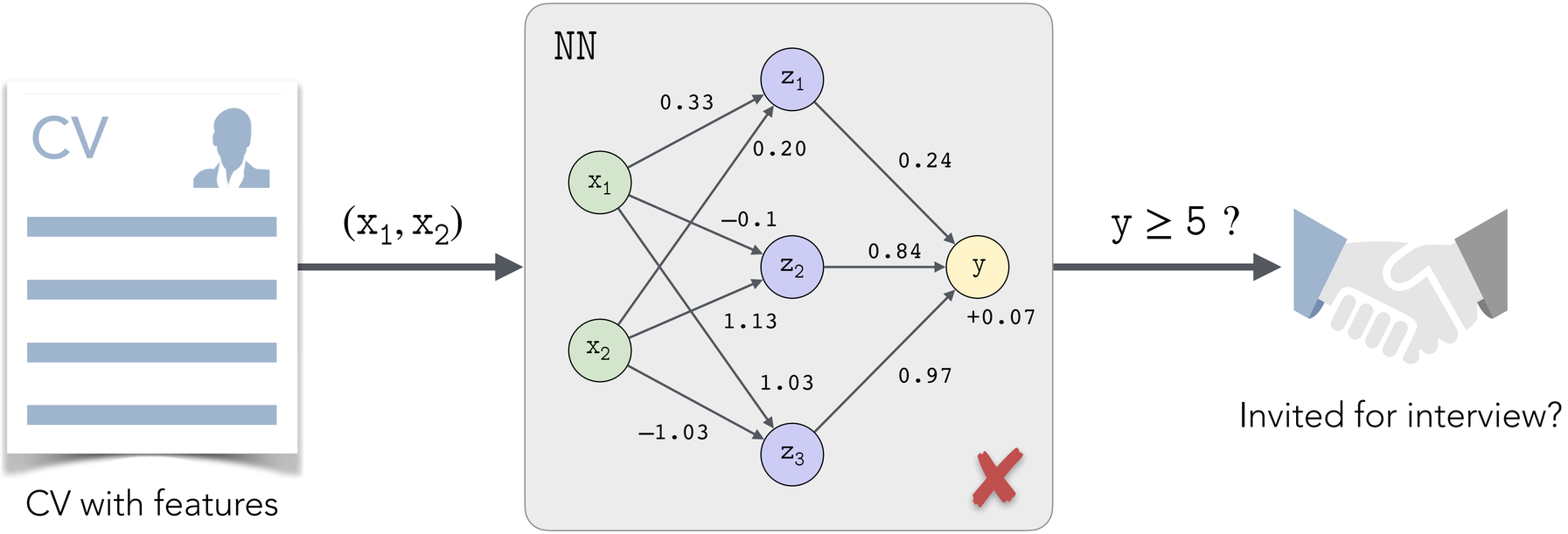
Is classification independent of sensitive features?

explainable

interpretable

verifiable

Is its inner functioning understandable?



fair

Is classification independent of sensitive features?

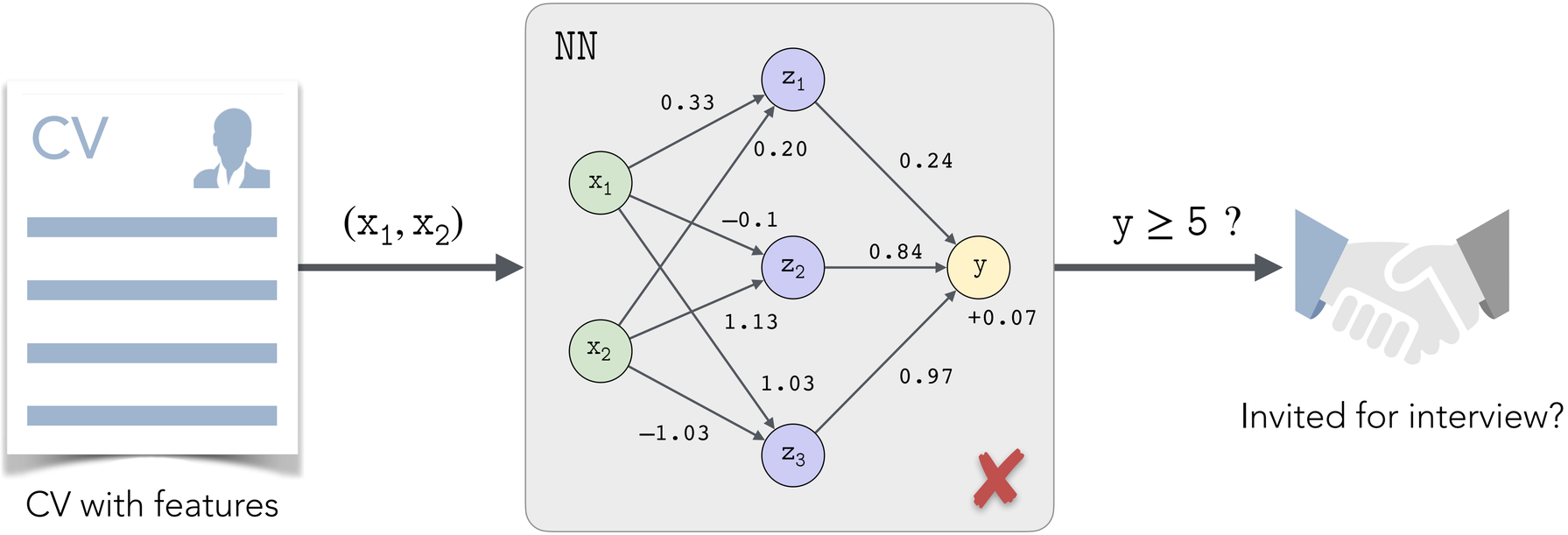
robust

explainable

verifiable

interpretable

Is its inner functioning understandable?



fair

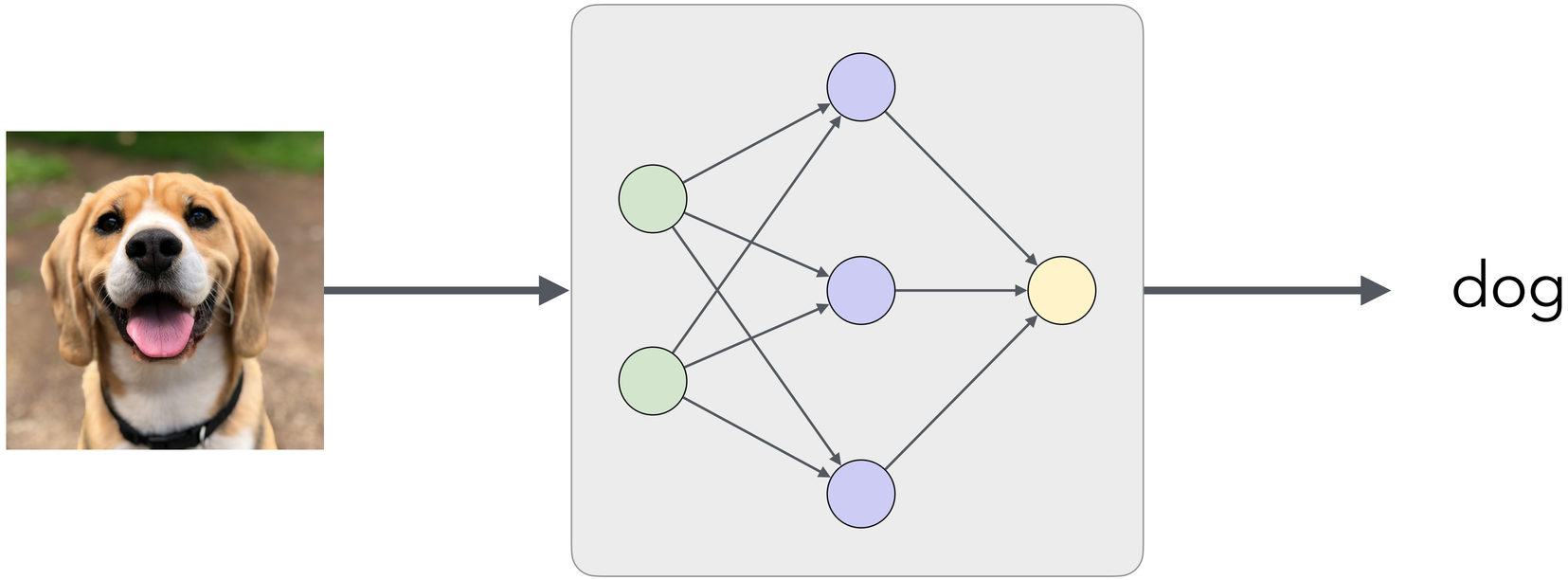
robust

Outcome is stable in presence of small perturbations.

explainable

verifiable

interpretable



fair

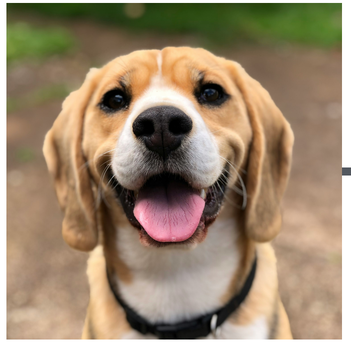
robust

Outcome is stable in presence of small perturbations.

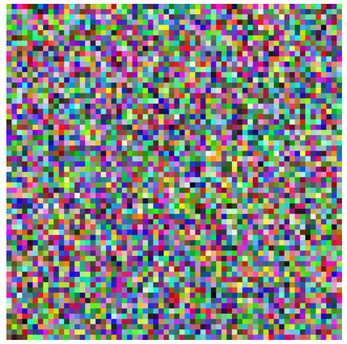
explainable

verifiable

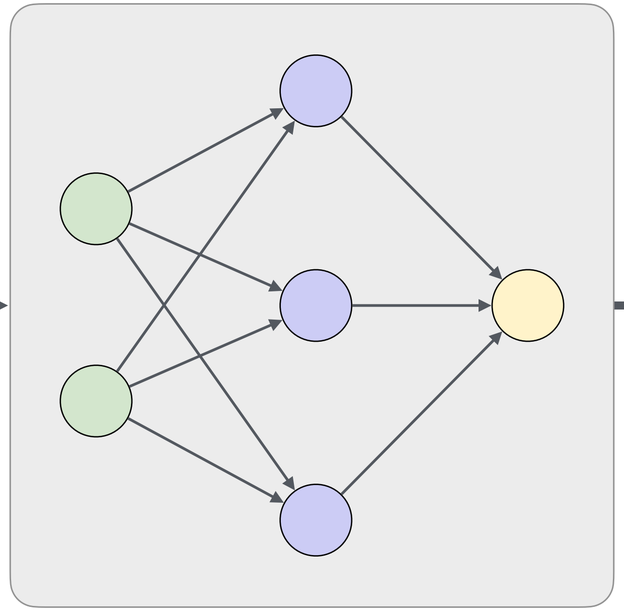
interpretable



+



noise



cat

fair

robust

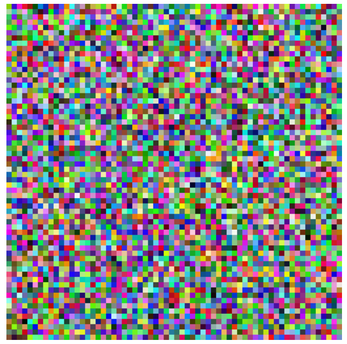
Outcome is stable in presence of small perturbations.

explainable

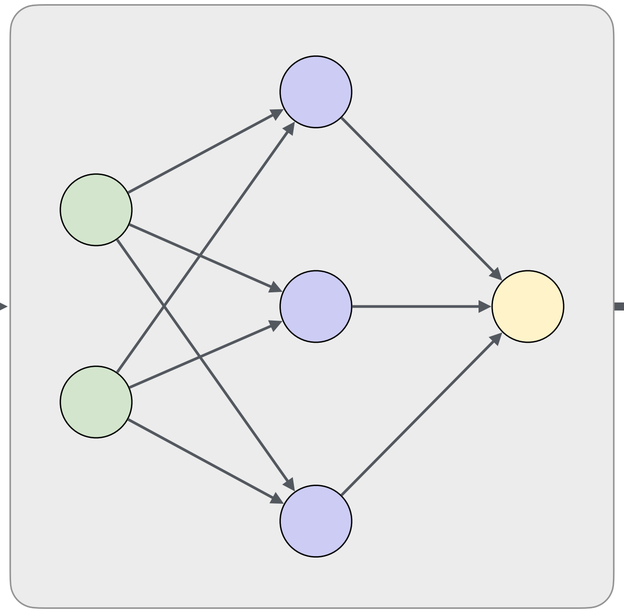
verifiable

interpretable

+



noise



green

fair

robust

Outcome is stable in presence of small perturbations.

explainable

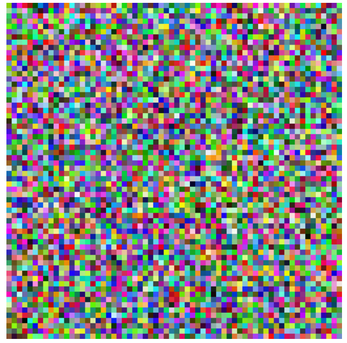
interpretable

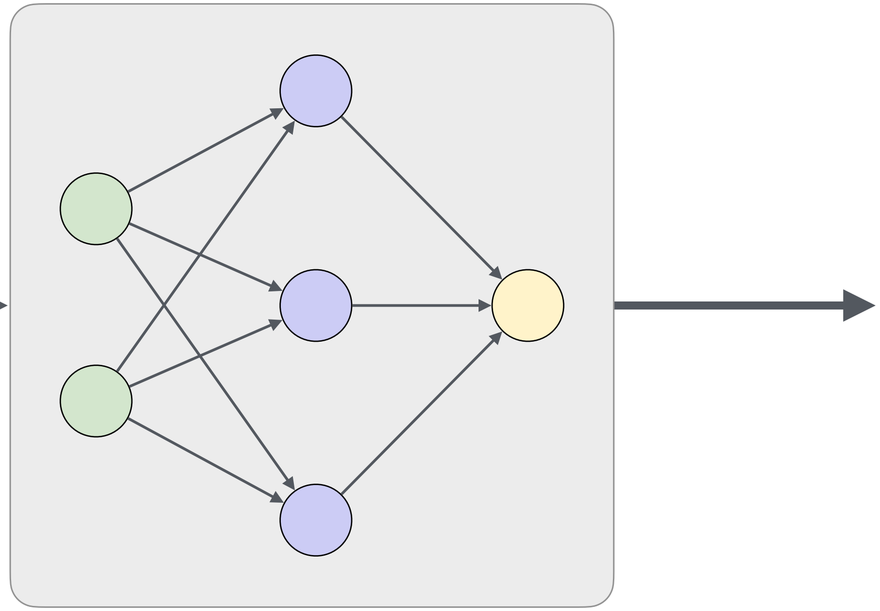
verifiable

synthetic example of a skin lesion (for illustration)

+

noise





benign → malignant

Small perturbations can fool medical AI in high-stakes applications.

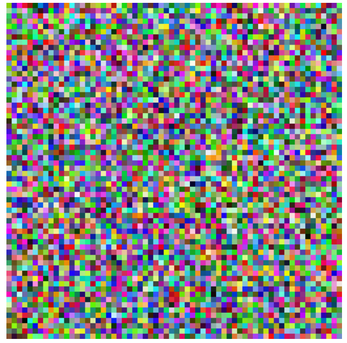
Source: Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS: *Adversarial attacks on medical machine learning: Emerging vulnerabilities demand new conversations.* Science. 2019;364(6439):128–130. https://www.science.org/doi/10.1126/science.aaw4399

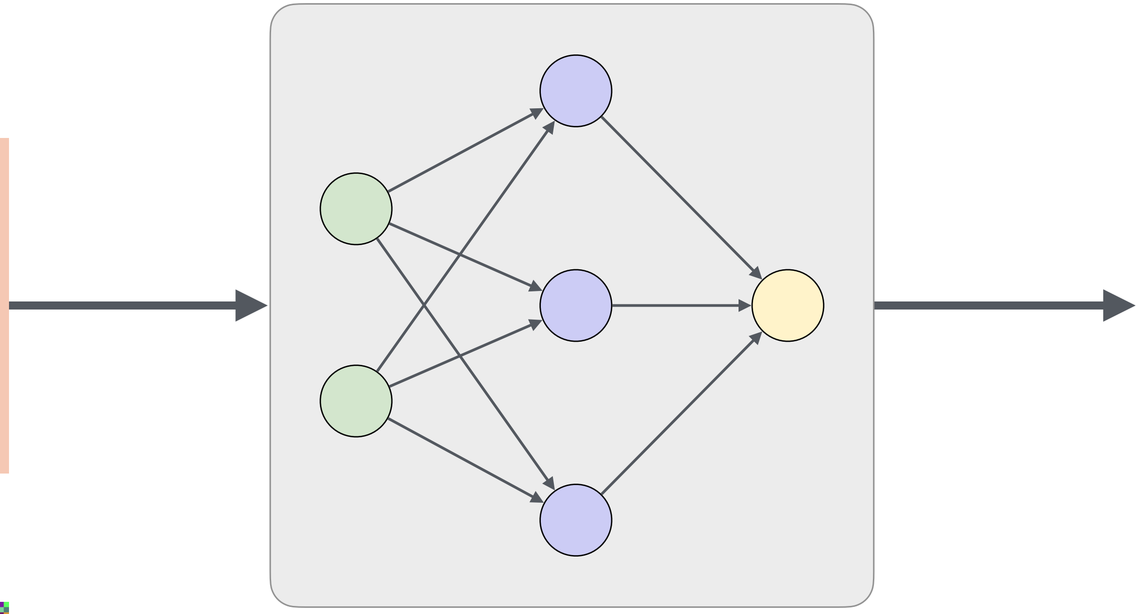
fair	robust	explainable	interpretable	ethical	human-centric
	Outcome is stable in presence of small perturbations.	verifiable	transparent	trustworthy	accountable

synthetic example of a skin lesion (for illustration)

+

noise





benign → malignant

Small perturbations can fool medical AI in high-stakes applications.

Source: Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS: *Adversarial attacks on medical machine learning: Emerging vulnerabilities demand new conversations.* Science. 2019;364(6439):128–130. https://www.science.org/doi/10.1126/science.aaw4399

fair	robust	explainable	interpretable	ethical	human-centric
safe	optimal	verifiable	transparent	trustworthy	accountable

fair	robust	explainable	interpretable	ethical	human-centric
safe	optimal	verifiable	transparent	trustworthy	accountable

explainable	Decisions are human-comprehensible.
interpretable	The functioning of a model can be understood.
verifiable	Formally provable against specifications.
transparent	Internals, design, and limitations are accessible.

ethical	Aligned with human values, rights, and societal norms.
human-centric	Puts people in control and respects their rights.
trustworthy	Consistently reliable, safe, and worthy of confidence.
accountable	Responsibility is clear and traceable for system outcomes.

behavioral/functional		structural/epistemic		normative	
fair	robust	explainable	interpretable	ethical	human-centric
safe	optimal	verifiable	transparent	trustworthy	accountable

what is computed? — *how is it computed?* — *why does it matter?*

classification of (AI) systems

explainable	Decisions are human-comprehensible.
interpretable	The functioning of a model can be understood.
verifiable	Formally provable against specifications.
transparent	Internals, design, and limitations are accessible.

ethical	Aligned with human values, rights, and societal norms.
human-centric	Puts people in control and respects their rights.
trustworthy	Consistently reliable, safe, and worthy of confidence.
accountable	Responsibility is clear and traceable for system outcomes.

fair	robust	explainable	interpretable	ethical	human-centric
safe	optimal	verifiable	transparent	trustworthy	accountable

behavioral/functional	structural/epistemic	normative
what is computed?	*how is it computed?*	*why does it matter?*

classification of (AI) systems

Such properties are (directly or indirectly) reflected in the EU AI Act.

explainable	Decisions are human-comprehensible.
interpretable	The functioning of a model can be understood.
verifiable	Formally provable against specifications.
transparent	Internals, design, and limitations are accessible.

ethical	Aligned with human values, rights, and societal norms.
human-centric	Puts people in control and respects their rights.
trustworthy	Consistently reliable, safe, and worthy of confidence.
accountable	Responsibility is clear and traceable for system outcomes.

behavioral/functional		structural/epistemic		normative	
fair	robust	explainable	interpretable	ethical	human-centric
safe	optimal	verifiable	transparent	trustworthy	accountable
what is computed?		*how is it computed?*		*why does it matter?*	

classification of (AI) systems

Such properties are (directly or indirectly) reflected in the EU AI Act.

- **Fair:** "[…] data sets shall be […] sufficiently representative." (Article 10(3))
- **Interpretable:** "[…] operation is sufficiently transparent […]" (Article 13(1))
- **Human-centric:** "[…] be effectively overseen by natural persons […]" (Article 14(1))

behavioral/functional		structural/epistemic		normative	
fair	robust	explainable	interpretable	ethical	human-centric
safe	optimal	verifiable	transparent	trustworthy	accountable

what is computed? | *how is it computed?* | *why does it matter?*

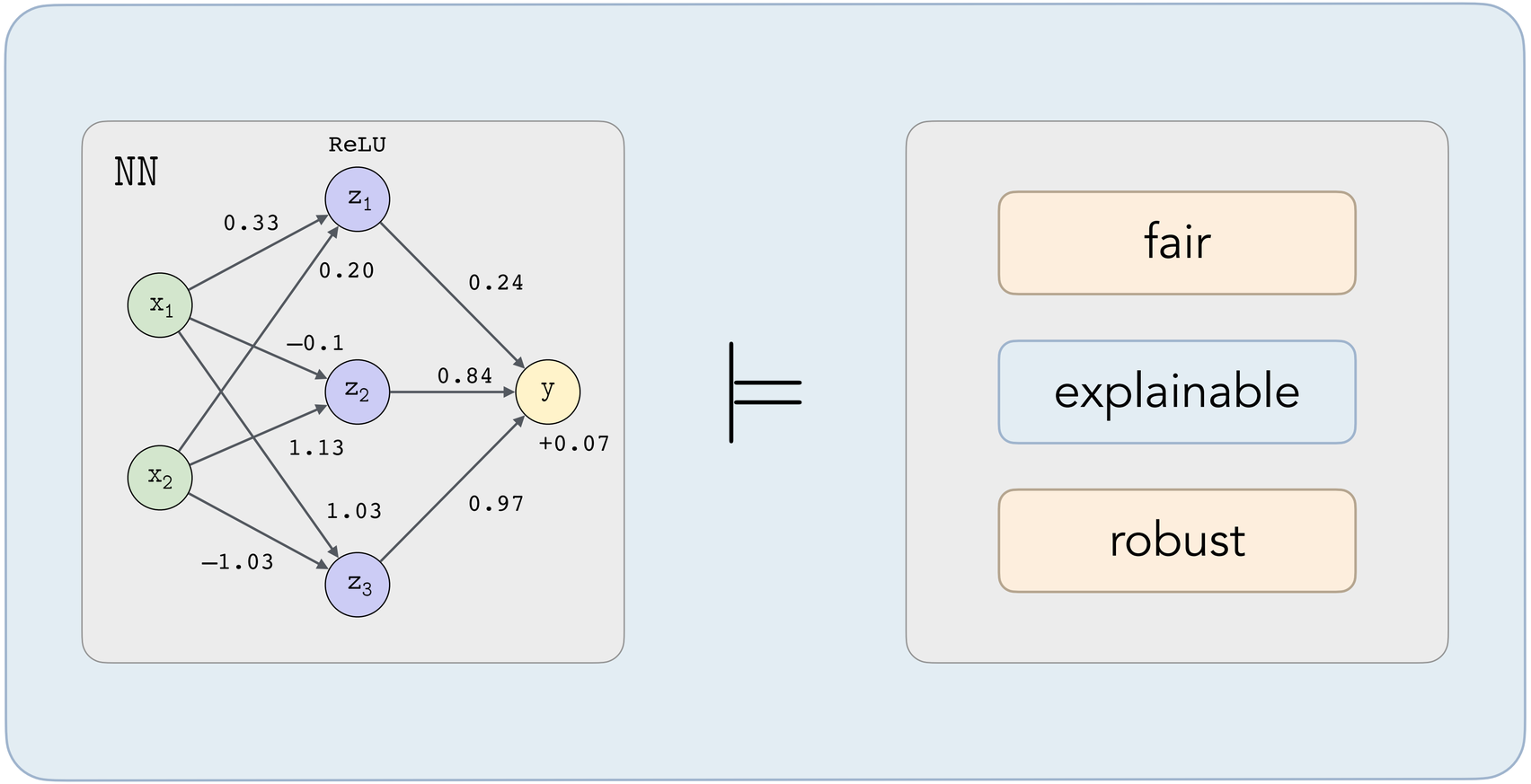
classification of (AI) systems

Such properties are (directly or indirectly) reflected in the EU AI Act.

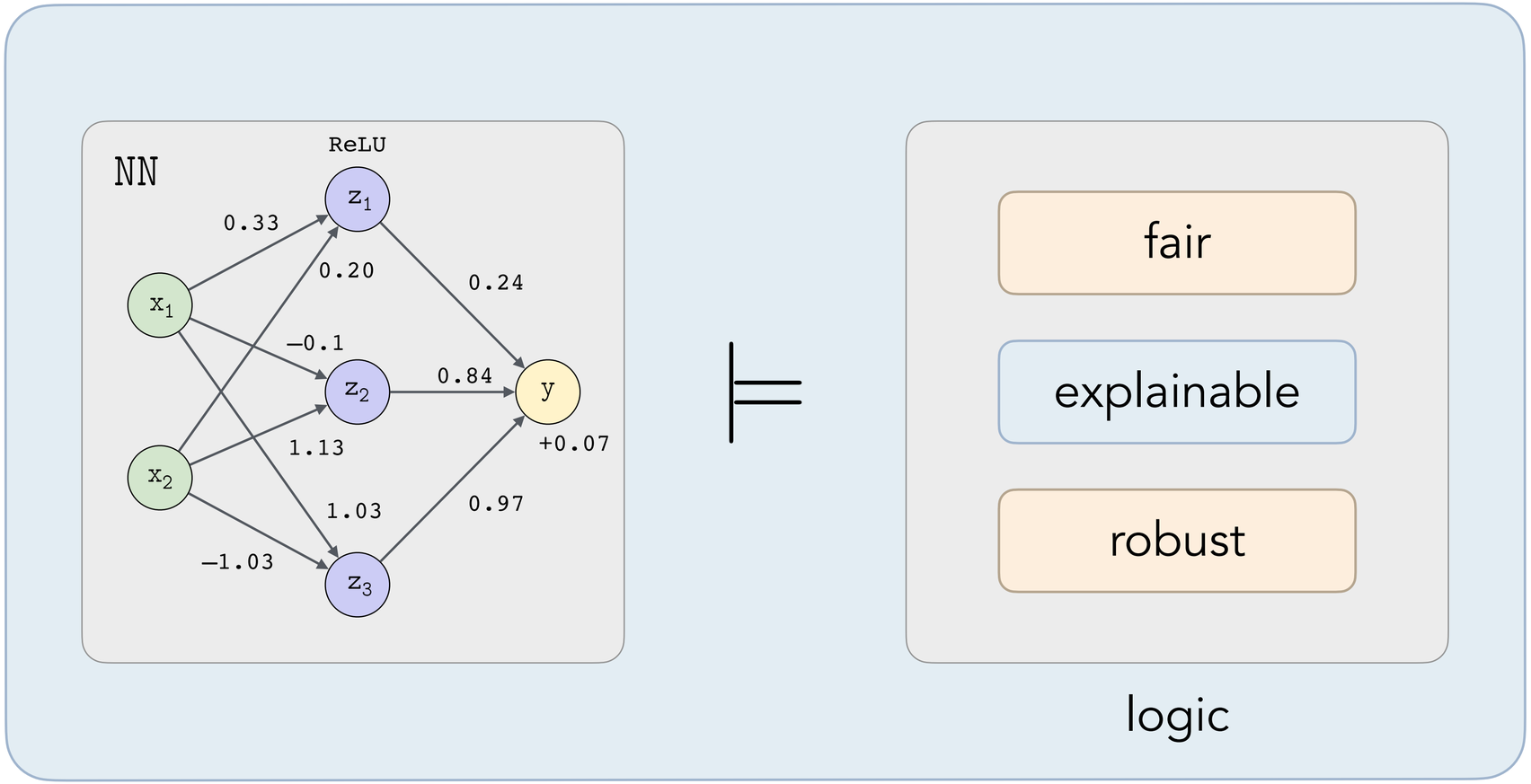
- Provide technical criteria and methods.
- If satisfied/applied sucessfully, creates *presumption of compliance*.
- But: Even standards rarely contain formal definitions.

The Role of Logic and Formal Methods

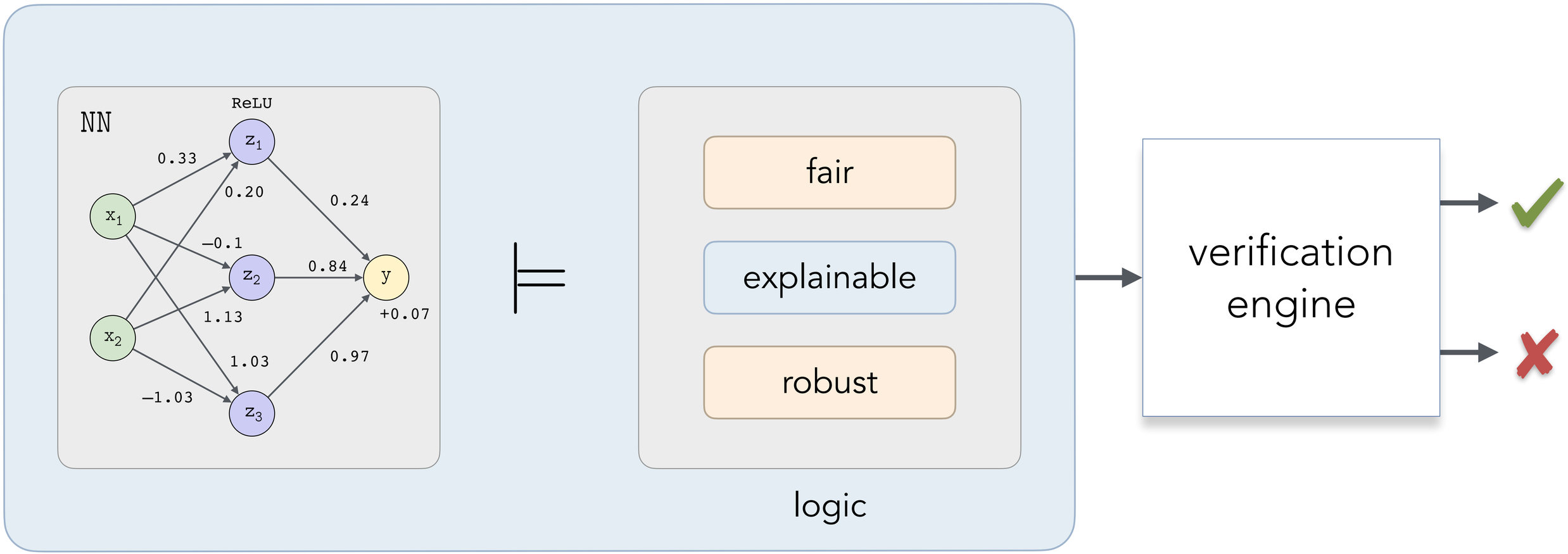
Verifiability



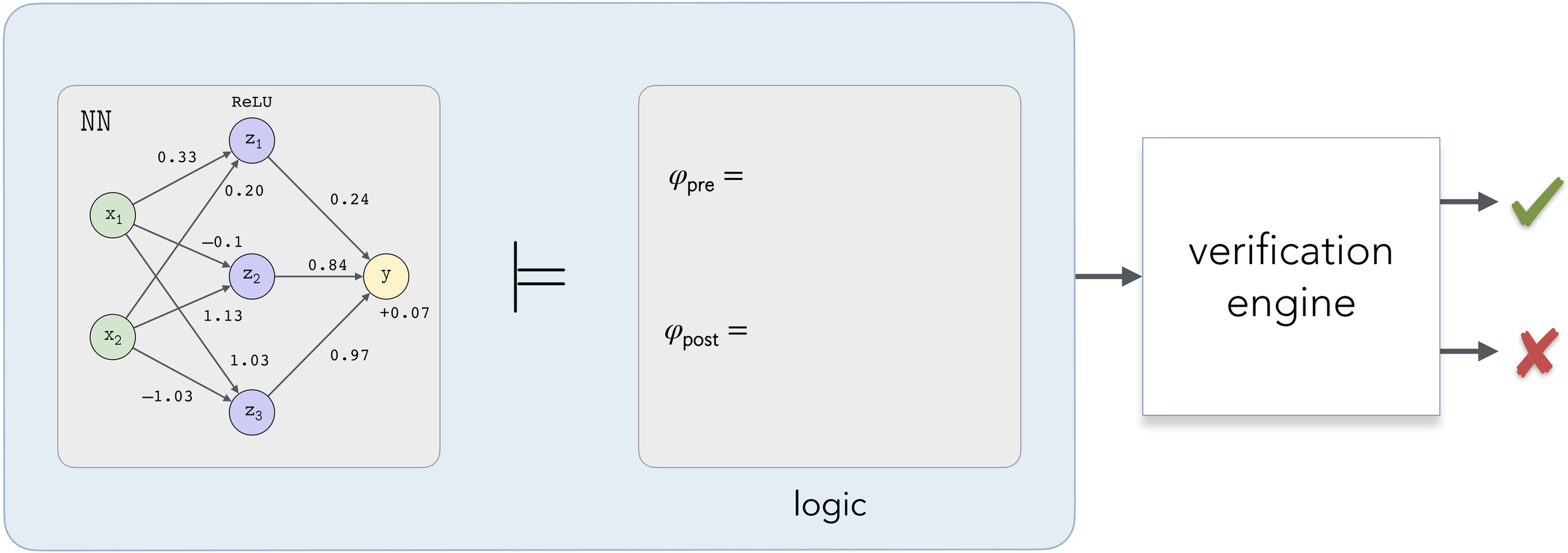
Verifiability



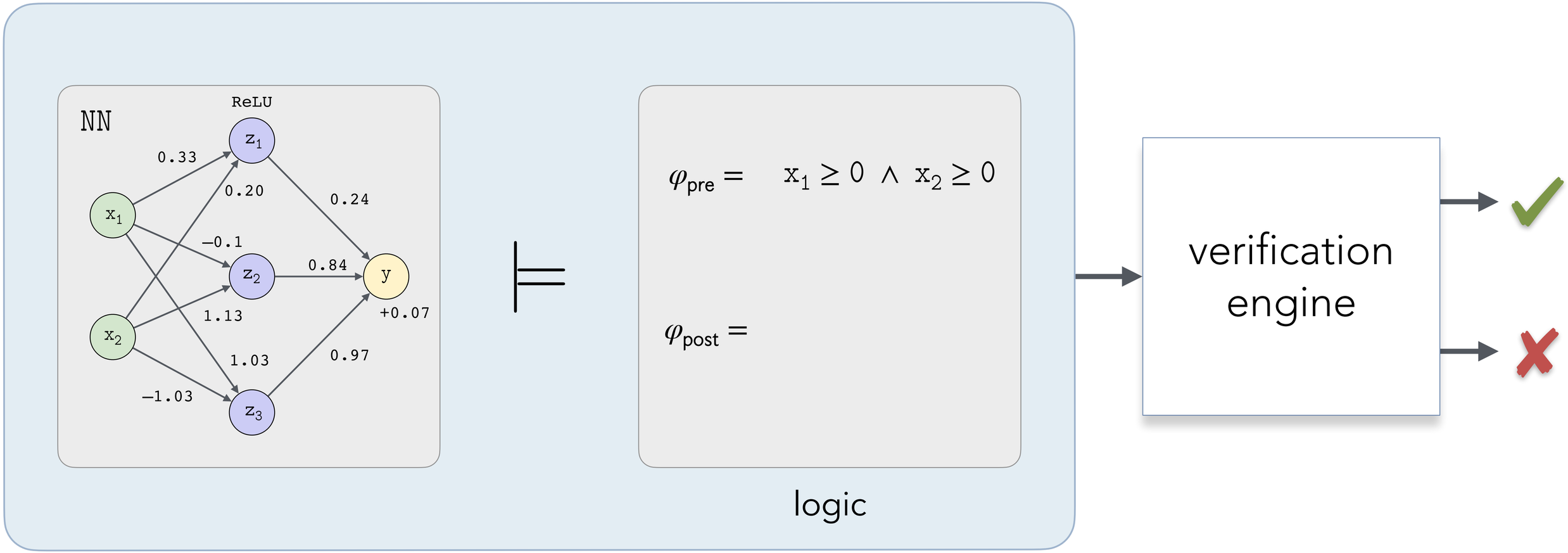
Verifiability



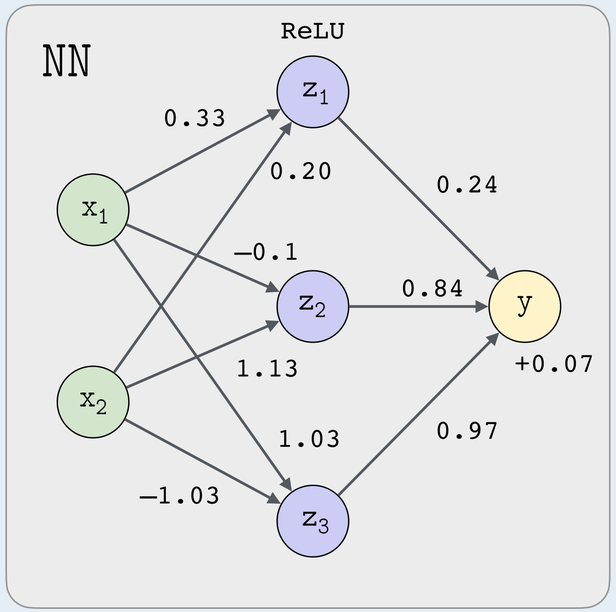
Verifiability



Verifiability



Verifiability



$\models$

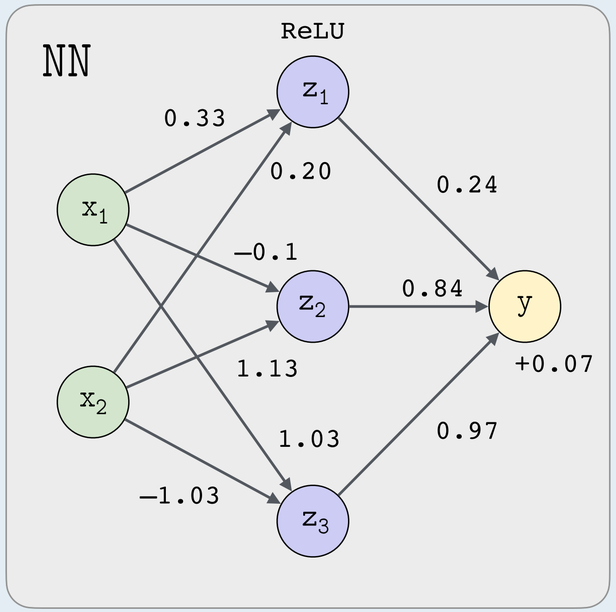
$$\varphi_{\text{pre}} = \quad x_1 \geq 0 \ \wedge \ x_2 \geq 0$$

$$\varphi_{\text{post}} = \quad \mathrm{NN}(x_1, x_2) \geq x_1 \ \wedge \ \mathrm{NN}(x_1, x_2) \geq x_2$$

logic

verification engine

Verifiability



max ?

$\models$

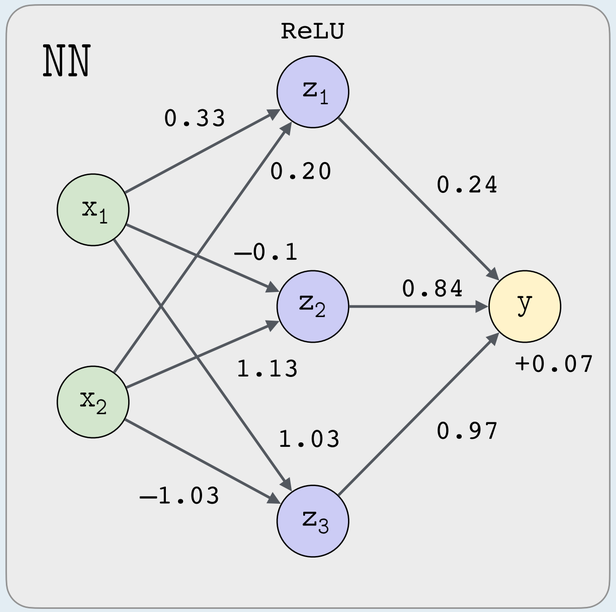
$$\varphi_{\text{pre}} = \quad x_1 \geq 0 \ \wedge \ x_2 \geq 0$$

$$\varphi_{\text{post}} = \quad \mathrm{NN}(x_1, x_2) \geq x_1 \ \wedge \ \mathrm{NN}(x_1, x_2) \geq x_2$$

logic

verification engine

Verifiability



max ?

$\models$

$$\varphi_{\text{pre}} = \quad x_1 \geq 0 \ \wedge \ x_2 \geq 0$$

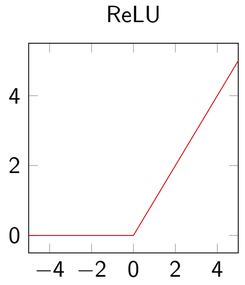
$$\varphi_{\text{post}} = \quad \mathrm{NN}(x_1, x_2) \geq x_1 \ \wedge \ \mathrm{NN}(x_1, x_2) \geq x_2$$

linear real arithmetic

verification engine

Verifiability

$$z_1 = \text{ReLU}(\underbrace{0.33\,x_1 + 0.2\,x_2}_{z_1'})$$



φ_{NN}

$$z_1' = 0.33\,x_1 + 0.2\,x_2$$

$\models$

$$\varphi_{\text{pre}} = \quad x_1 \geq 0 \;\wedge\; x_2 \geq 0$$

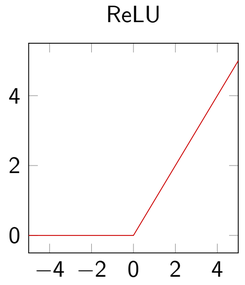
$$\varphi_{\text{post}} = \quad \text{NN}(x_1, x_2) \geq x_1 \;\wedge\; \text{NN}(x_1, x_2) \geq x_2$$

linear real arithmetic

verification engine

Verifiability

$$z_1 = \text{ReLU}(\underbrace{0.33\,x_1 + 0.2\,x_2}_{z_1'})$$



φ_{NN}

$$z_1' = 0.33\,x_1 + 0.2\,x_2$$
$$\wedge\ z_1' < 0 \ \Rightarrow\ z_1 = 0$$

$\models$

$$\varphi_{\text{pre}} = \quad x_1 \geq 0 \ \wedge\ x_2 \geq 0$$

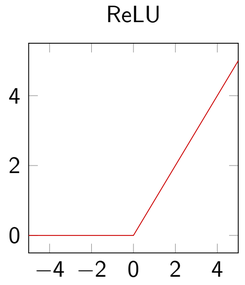
$$\varphi_{\text{post}} = \quad \text{NN}(x_1, x_2) \geq x_1$$
$$\wedge\ \text{NN}(x_1, x_2) \geq x_2$$

linear real arithmetic

verification engine

Verifiability

$$z_1 = \text{ReLU}(\underbrace{0.33\,x_1 + 0.2\,x_2}_{z_1'})$$



φ_{NN}

$$\begin{aligned} & z_1' = 0.33\,x_1 + 0.2\,x_2 \\ \wedge\ & z_1' < 0 \ \Rightarrow\ z_1 = 0 \\ \wedge\ & z_1' \geq 0 \ \Rightarrow\ z_1 = z_1' \end{aligned}$$

$\models$

$$\varphi_{\text{pre}} = \quad x_1 \geq 0 \ \wedge\ x_2 \geq 0$$

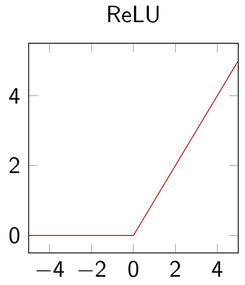
$$\varphi_{\text{post}} = \quad \text{NN}(x_1, x_2) \geq x_1 \ \wedge\ \text{NN}(x_1, x_2) \geq x_2$$

linear real arithmetic

verification engine

Verifiability

$$z_1 = \text{ReLU}(\underbrace{0.33\,x_1 + 0.2\,x_2}_{z_1'})$$



φ_{NN}

$$\begin{aligned} & z_1' = 0.33\,x_1 + 0.2\,x_2 \\ \wedge\ & z_1' < 0 \Rightarrow z_1 = 0 \\ \wedge\ & z_1' \geq 0 \Rightarrow z_1 = z_1' \\ & \vdots \\ \wedge\ & y = 0.24\,z_1 + 0.84\,z_2 \\ & \quad + 0.97\,z_3 + 0.07 \end{aligned}$$

$\models$

$$\varphi_{\text{pre}} = x_1 \geq 0 \wedge x_2 \geq 0$$

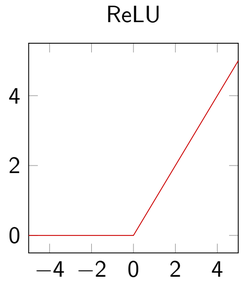
$$\varphi_{\text{post}} = \text{NN}(x_1, x_2) \geq x_1 \wedge \text{NN}(x_1, x_2) \geq x_2$$

linear real arithmetic

verification engine

Verifiability

$$z_1 = \text{ReLU}(\underbrace{0.33\,x_1 + 0.2\,x_2}_{z_1'})$$



φ_{NN}

$$\begin{aligned} & z_1' = 0.33\,x_1 + 0.2\,x_2 \\ \wedge\ & z_1' < 0 \Rightarrow z_1 = 0 \\ \wedge\ & z_1' \geq 0 \Rightarrow z_1 = z_1' \\ & \vdots \\ \wedge\ & y = 0.24\,z_1 + 0.84\,z_2 \\ & \quad + 0.97\,z_3 + 0.07 \end{aligned}$$

$\models$

$$\varphi_{\text{pre}} = x_1 \geq 0 \wedge x_2 \geq 0$$

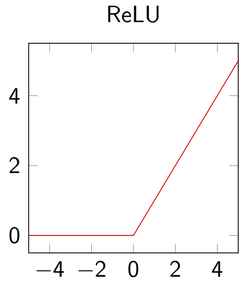
$$\varphi_{\text{post}} = y \geq x_1 \wedge y \geq x_2$$

linear real arithmetic

verification engine

Verifiability

$$z_1 = \text{ReLU}(\underbrace{0.33\,x_1 + 0.2\,x_2}_{z_1'})$$



$\varphi_{\mathtt{NN}}$

$$\begin{aligned} & z_1' = 0.33\,x_1 + 0.2\,x_2 \\ \wedge\ & z_1' < 0 \Rightarrow z_1 = 0 \\ \wedge\ & z_1' \geq 0 \Rightarrow z_1 = z_1' \\ & \vdots \\ \wedge\ & y = 0.24\,z_1 + 0.84\,z_2 \\ & \quad + 0.97\,z_3 + 0.07 \end{aligned}$$

$\models$

$$\varphi_{\text{pre}} = x_1 \geq 0 \wedge x_2 \geq 0$$

$$\varphi_{\text{post}} = y \geq x_1 \wedge y \geq x_2$$

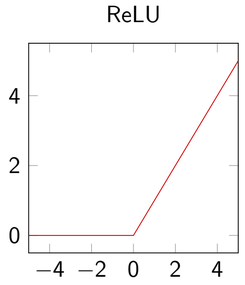
linear real arithmetic

$$\forall x_1, x_2, y$$
$$\forall (z_i, z_i')_{i=1,2,3}$$
$$(\varphi_{\text{pre}} \wedge \varphi_{\mathtt{NN}}) \Rightarrow \varphi_{\text{post}}$$

true?

Verifiability

$$z_1 = \text{ReLU}(\underbrace{0.33\,x_1 + 0.2\,x_2}_{z_1'})$$



φ_{NN}

$$\begin{aligned} & z_1' = 0.33\,x_1 + 0.2\,x_2 \\ \wedge\ & z_1' < 0 \Rightarrow z_1 = 0 \\ \wedge\ & z_1' \geq 0 \Rightarrow z_1 = z_1' \\ & \vdots \\ \wedge\ & y = 0.24\,z_1 + 0.84\,z_2 \\ & \quad + 0.97\,z_3 + 0.07 \end{aligned}$$

$\models$

$$\varphi_{\text{pre}} = x_1 \geq 0 \wedge x_2 \geq 0$$

$$\varphi_{\text{post}} = y \geq x_1 \wedge y \geq x_2$$

linear real arithmetic

$\varphi_{\text{pre}} \wedge \varphi_{\text{NN}} \wedge \neg\varphi_{\text{post}}$

satisfiable?

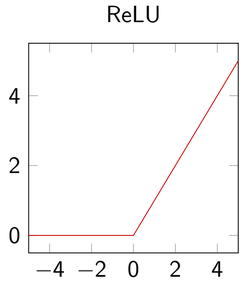
✓

✗

(x_1, x_2)

Verifiability

$$z_1 = \text{ReLU}(\underbrace{0.33\,x_1 + 0.2\,x_2}_{z_1'})$$



φ_{NN}

$$\begin{aligned} & z_1' = 0.33\,x_1 + 0.2\,x_2 \\ \wedge\ & z_1' < 0 \Rightarrow z_1 = 0 \\ \wedge\ & z_1' \geq 0 \Rightarrow z_1 = z_1' \\ & \vdots \\ \wedge\ & y = 0.24\,z_1 + 0.84\,z_2 \\ & \quad + 0.97\,z_3 + 0.07 \end{aligned}$$

$\text{NN}(4, 93) = 92.79$

$\models$

$$\varphi_{\text{pre}} = x_1 \geq 0 \wedge x_2 \geq 0$$

$$\varphi_{\text{post}} = y \geq x_1 \wedge y \geq x_2$$

linear real arithmetic

$\varphi_{\text{pre}} \wedge \varphi_{\text{NN}} \wedge \neg\varphi_{\text{post}}$

satisfiable?

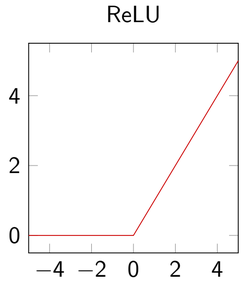
✓

✗

(x_1, x_2)

Verifiability

$$z_1 = \text{ReLU}(\underbrace{0.33\,x_1 + 0.2\,x_2}_{z_1'})$$



φ_{NN}

$$\begin{aligned} & z_1' = 0.33\,x_1 + 0.2\,x_2 \\ \wedge\ & z_1' < 0 \Rightarrow z_1 = 0 \\ \wedge\ & z_1' \geq 0 \Rightarrow z_1 = z_1' \\ & \vdots \\ \wedge\ & y = 0.24\,z_1 + 0.84\,z_2 \\ & \quad + 0.97\,z_3 + 0.07 \end{aligned}$$

$\text{NN}(4, 93) = 92.79$

$\models$

$$\varphi_{\text{pre}} = x_1 \geq 0 \wedge x_2 \geq 0$$

$$\varphi_{\text{post}} = y \geq x_1 \wedge y \geq x_2$$

linear real arithmetic

satisfiability modulo theories

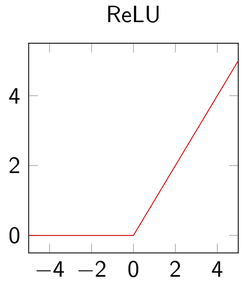
✓

✗

(x_1, x_2)

Verifiability

$$z_1 = \text{ReLU}(\underbrace{0.33\,x_1 + 0.2\,x_2}_{z_1'})$$



φ_{NN}

$$\begin{aligned} & z_1' = 0.33\,x_1 + 0.2\,x_2 \\ \wedge\ & z_1' < 0 \Rightarrow z_1 = 0 \\ \wedge\ & z_1' \geq 0 \Rightarrow z_1 = z_1' \\ & \vdots \\ \wedge\ & y = 0.24\,z_1 + 0.84\,z_2 \\ & \quad + 0.97\,z_3 + 0.07 \end{aligned}$$

$\text{NN}(4, 93) = 92.79$

$\models$

$$\varphi_{\text{pre}} = x_1 \geq 0 \wedge x_2 \geq 0$$

$$\varphi_{\text{post}} = y \geq x_1 \wedge y \geq x_2$$

linear real arithmetic

versatile language

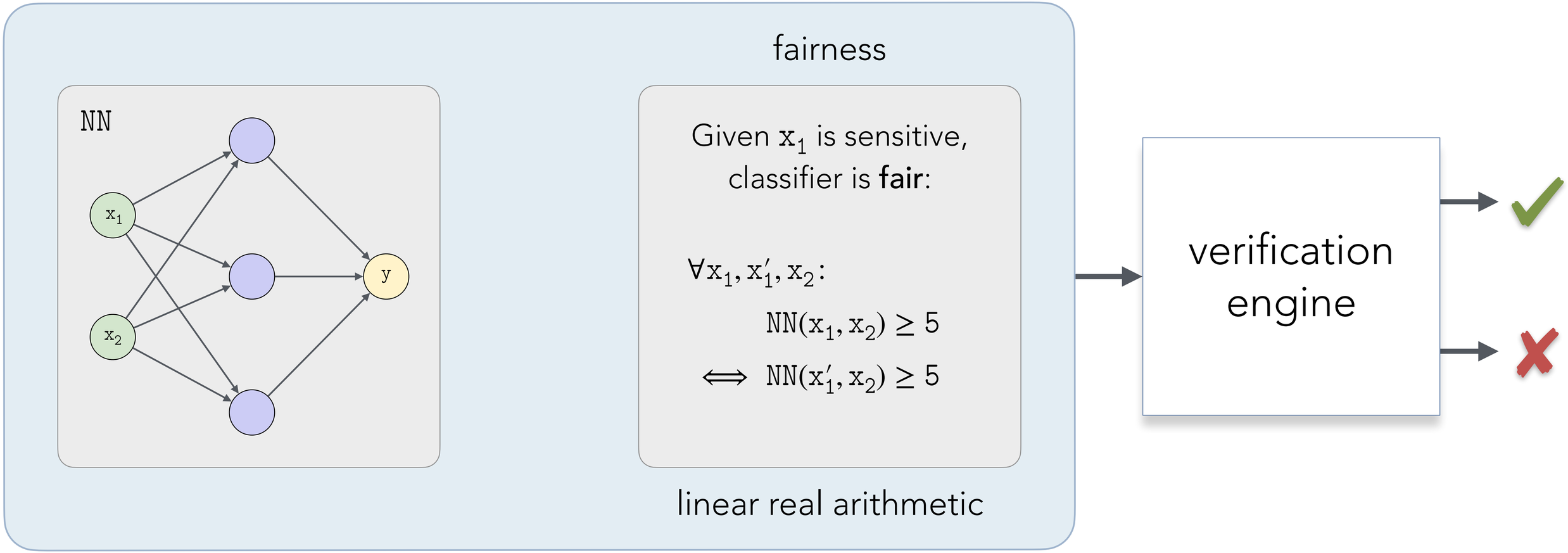
satisfiability modulo theories

✓

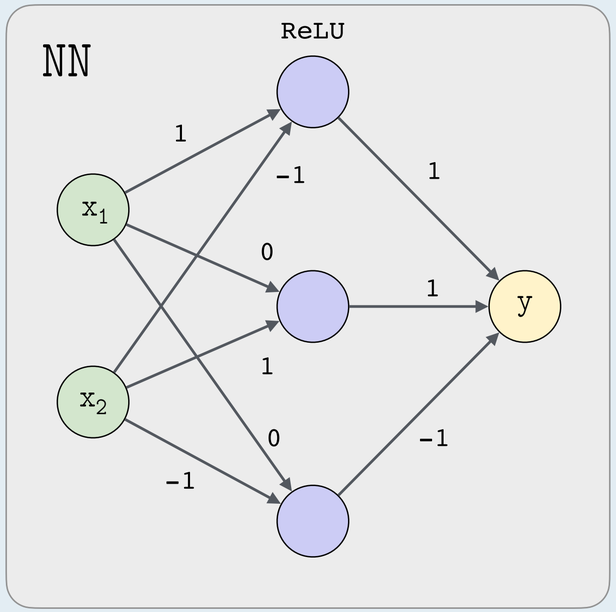
✗

(x_1, x_2)

Verifiability



Verifiability



max

fairness

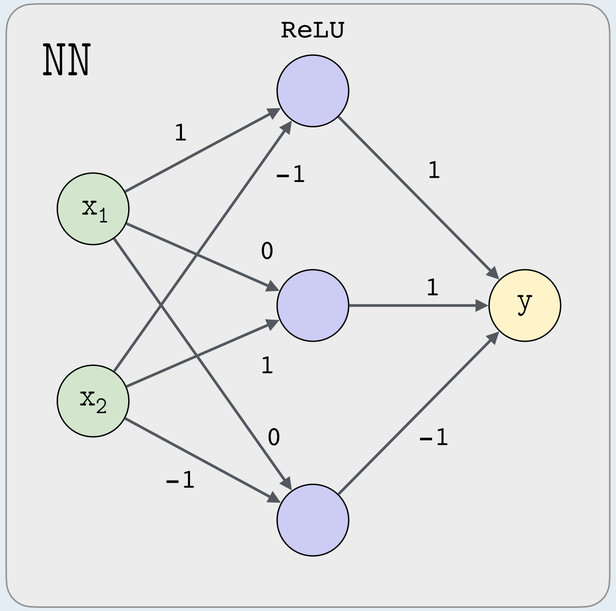
Given x_1 is sensitive, classifier is **fair**:

$$\forall x_1, x_1', x_2:$$
$$\mathrm{NN}(x_1, x_2) \geq 5$$
$$\iff \mathrm{NN}(x_1', x_2) \geq 5$$

linear real arithmetic

verification engine

Verifiability



max

$\not\models$

fairness

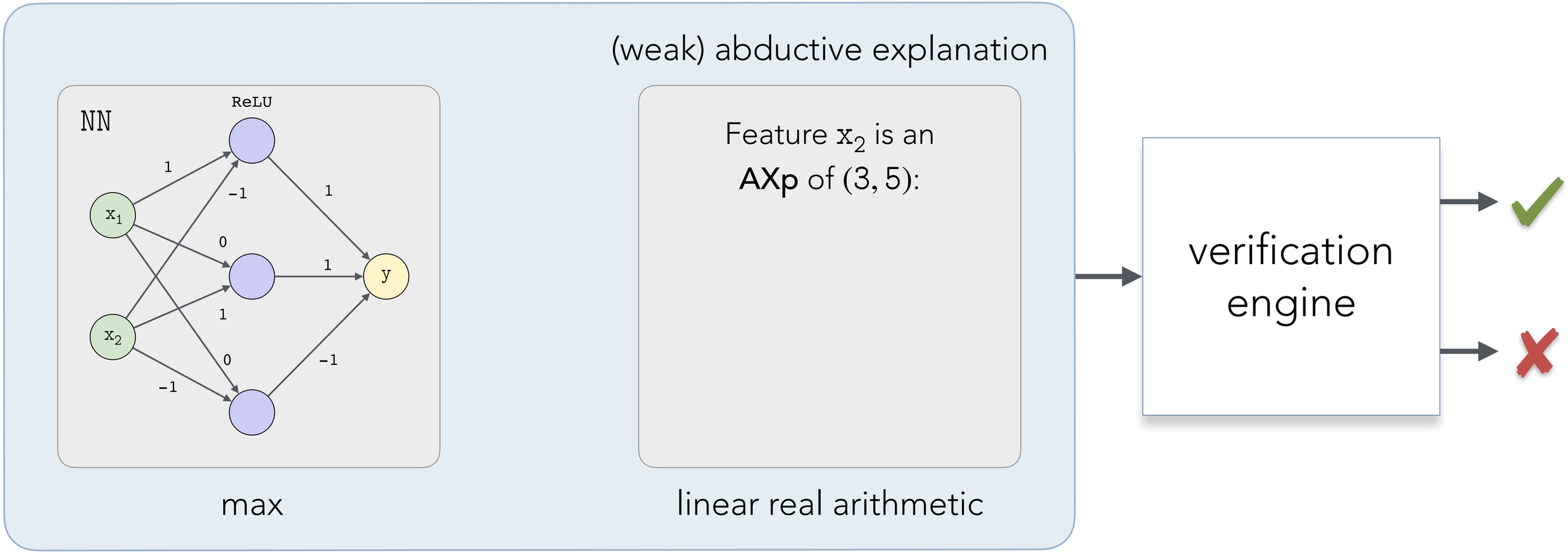
Given x_1 is sensitive, classifier is **fair**:

$$\forall x_1, x_1', x_2:$$
$$\mathrm{NN}(x_1, x_2) \geq 5$$
$$\iff \mathrm{NN}(x_1', x_2) \geq 5$$

linear real arithmetic

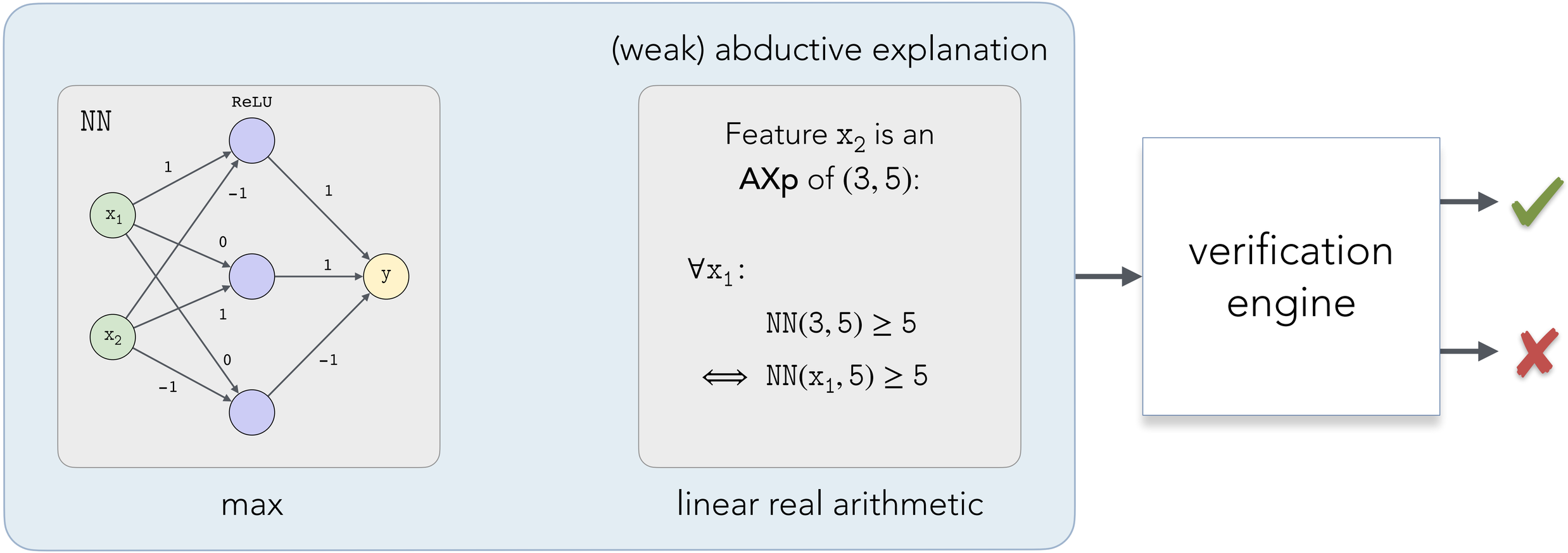
verification engine

Verifiability



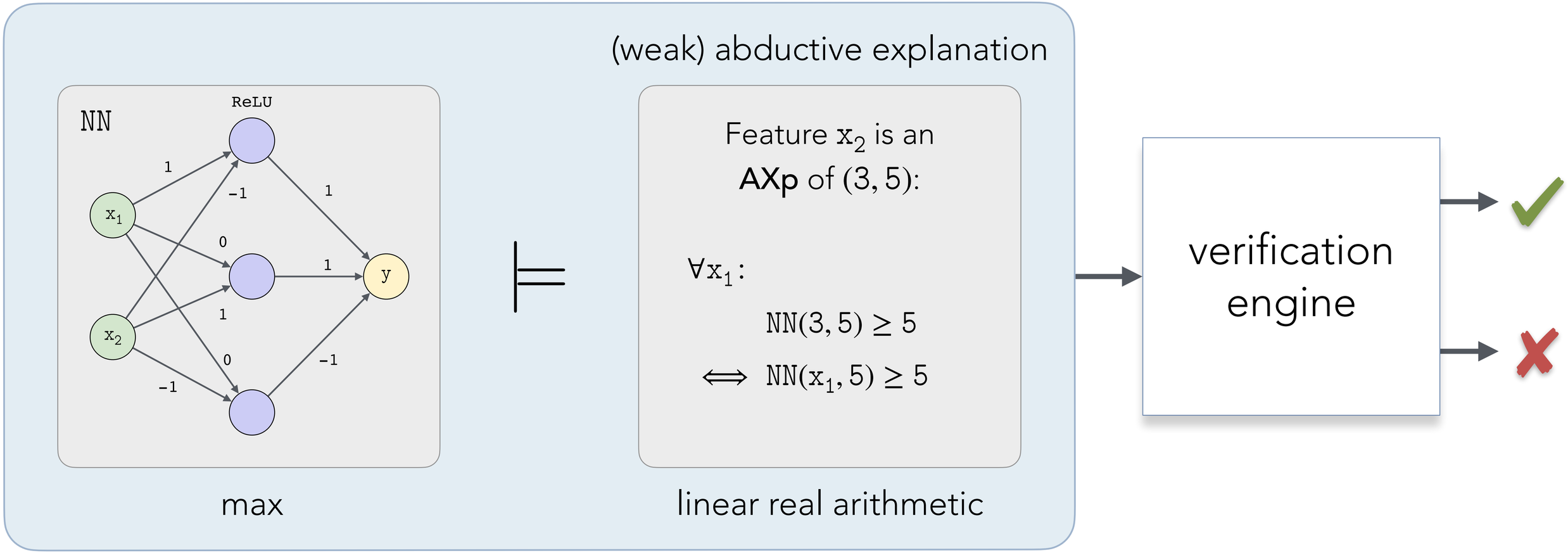
[Marques-Silva, Ignatiev: *Delivering Trustworthy AI through Formal XAI.* AAAI 2022]

Verifiability



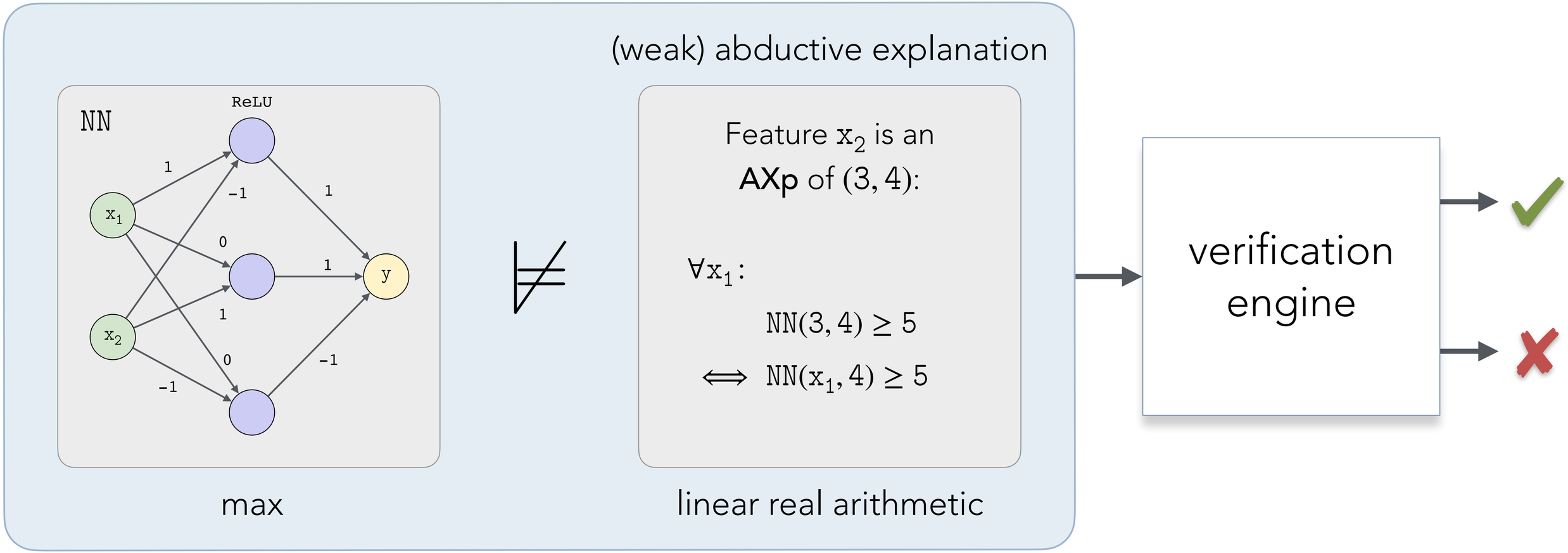
[Marques-Silva, Ignatiev: *Delivering Trustworthy AI through Formal XAI.* AAAI 2022]

Verifiability



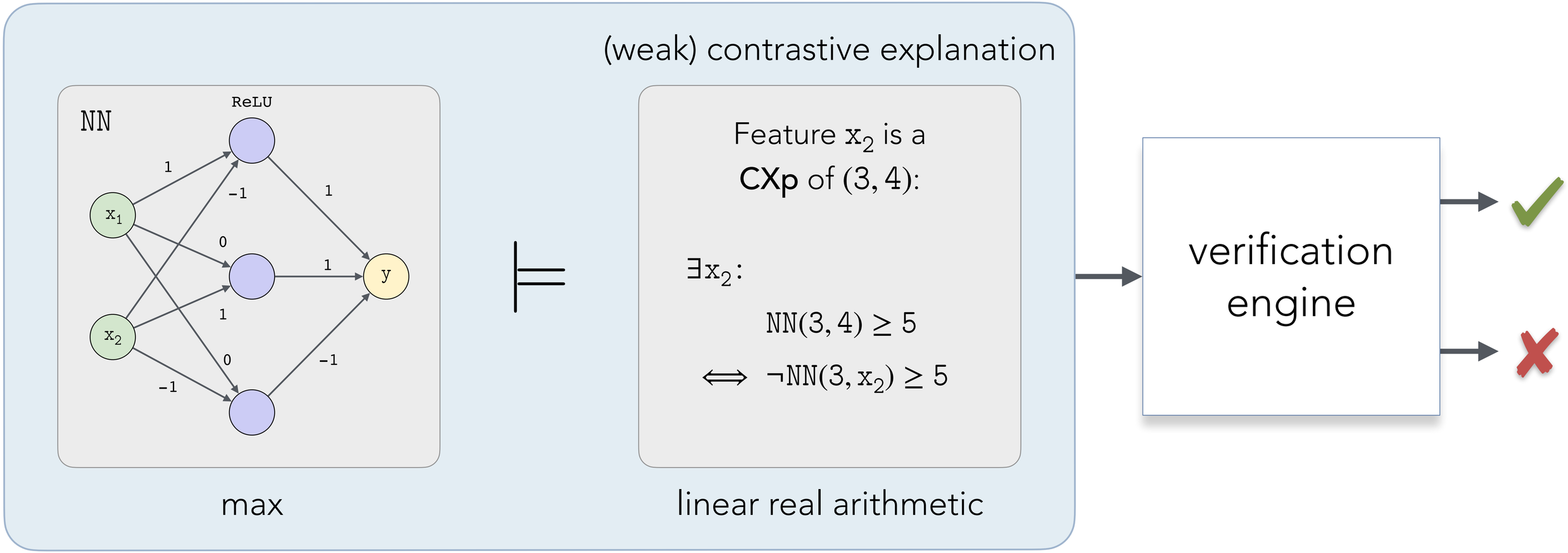
[Marques-Silva, Ignatiev: *Delivering Trustworthy AI through Formal XAI.* AAAI 2022]

Verifiability



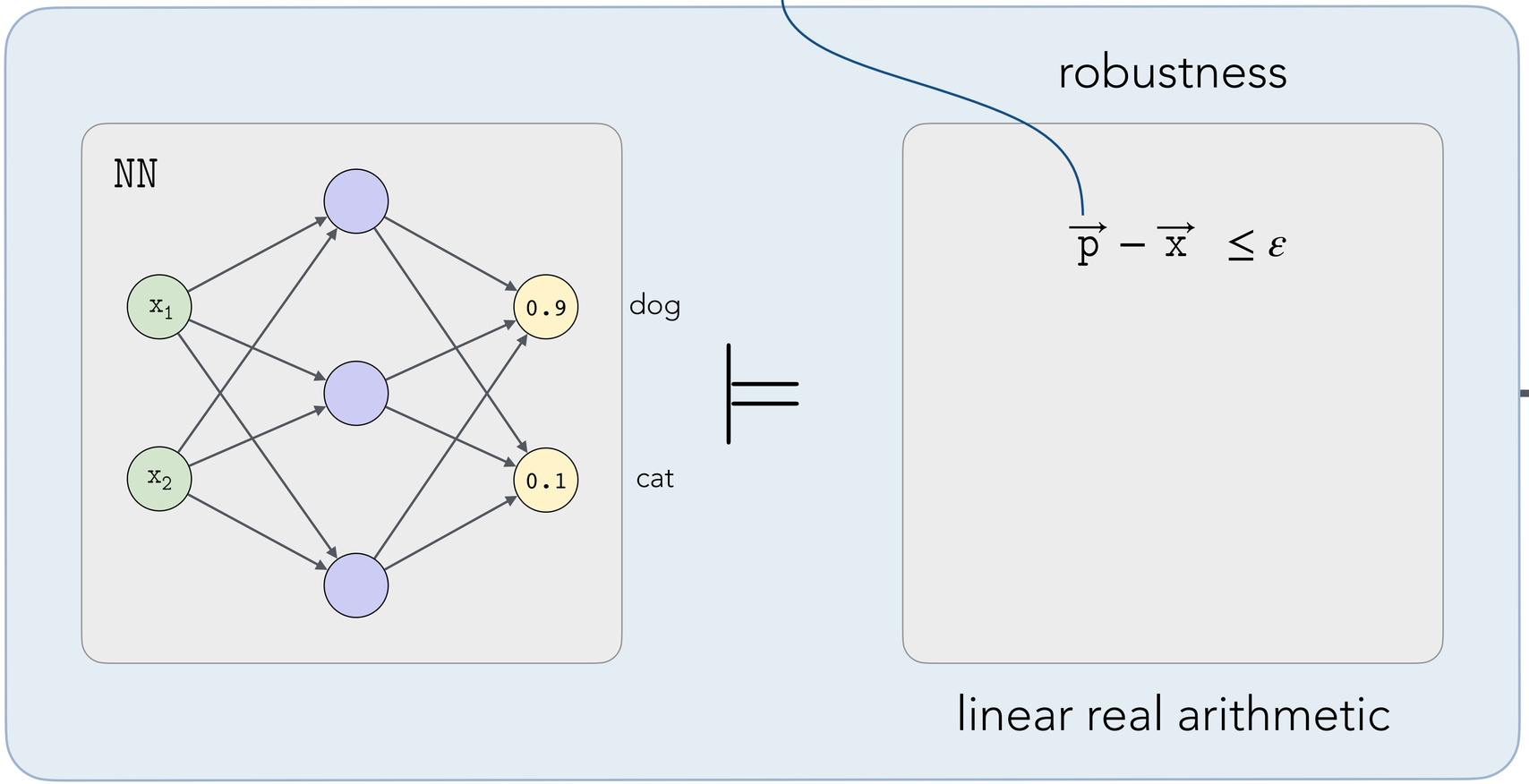
[Marques-Silva, Ignatiev: *Delivering Trustworthy AI through Formal XAI.* AAAI 2022]

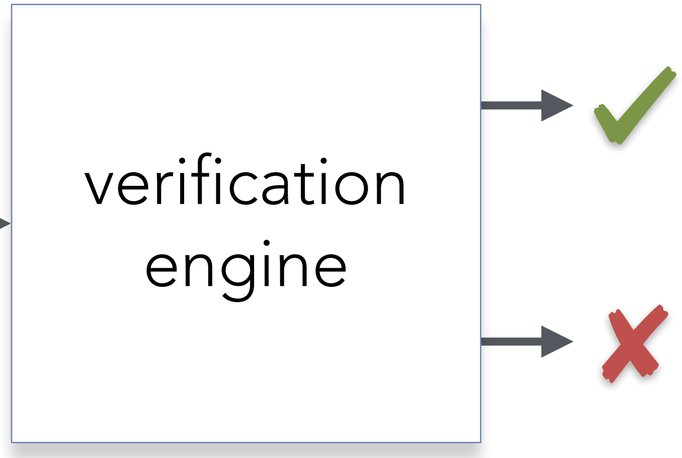
Verifiability



[Marques-Silva, Ignatiev: *Delivering Trustworthy AI through Formal XAI.* AAAI 2022]

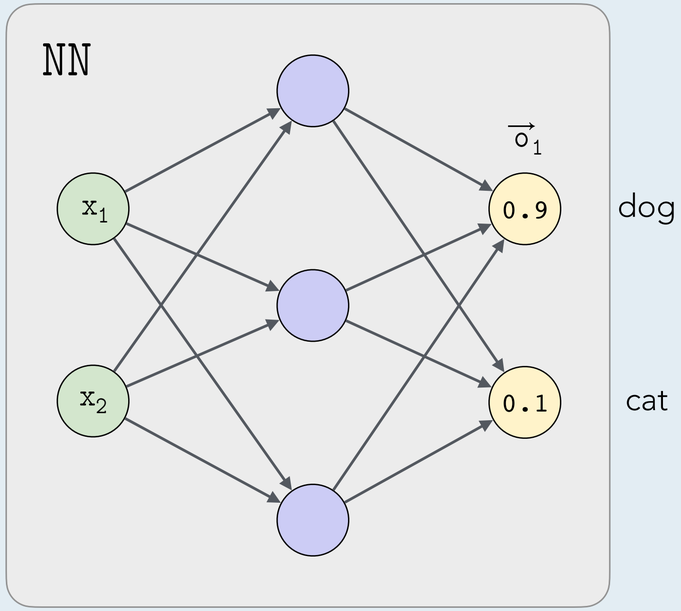
Verifiability





Verifiability

robustness



$\models$

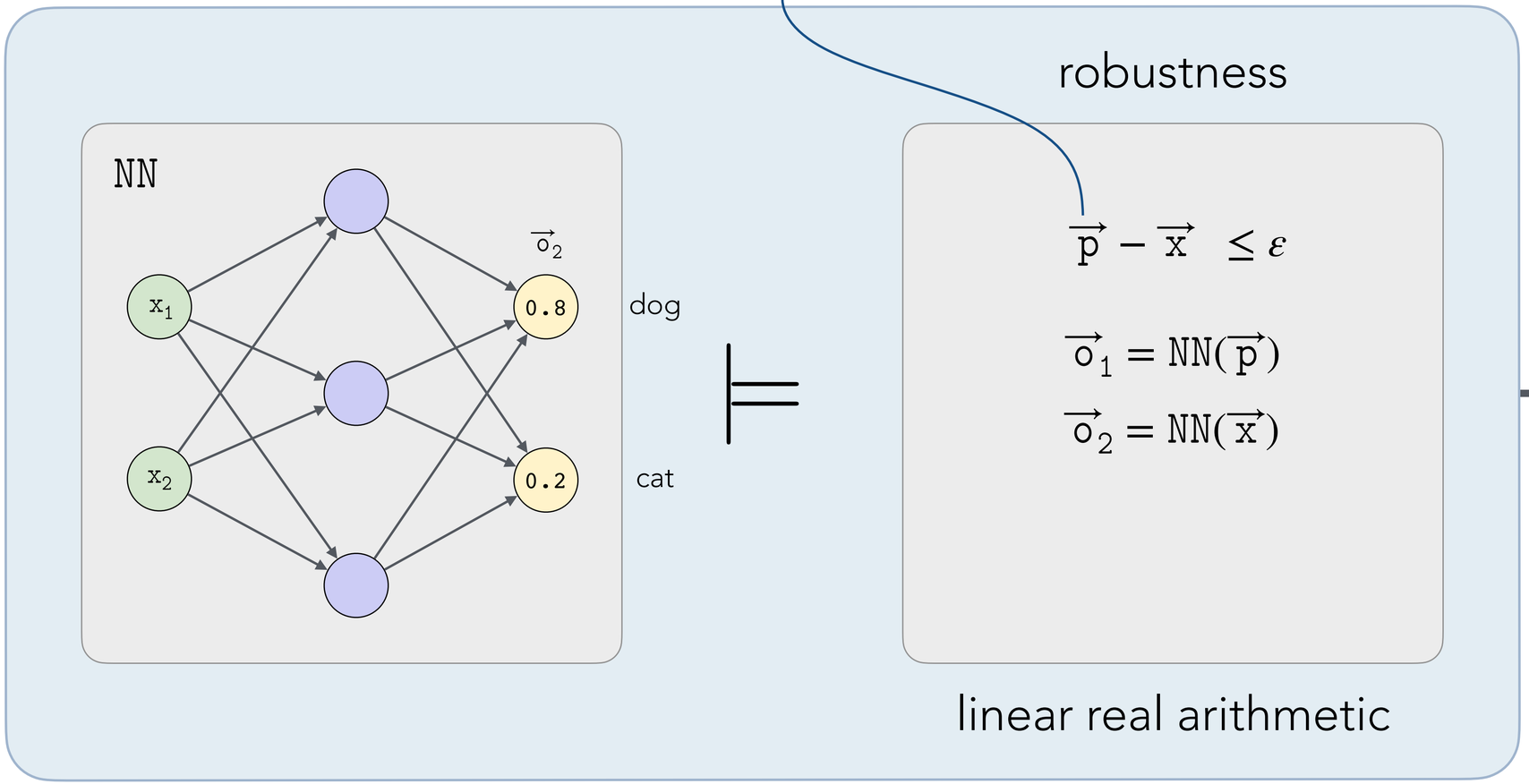
$$\vec{p} - \vec{x} \leq \varepsilon$$

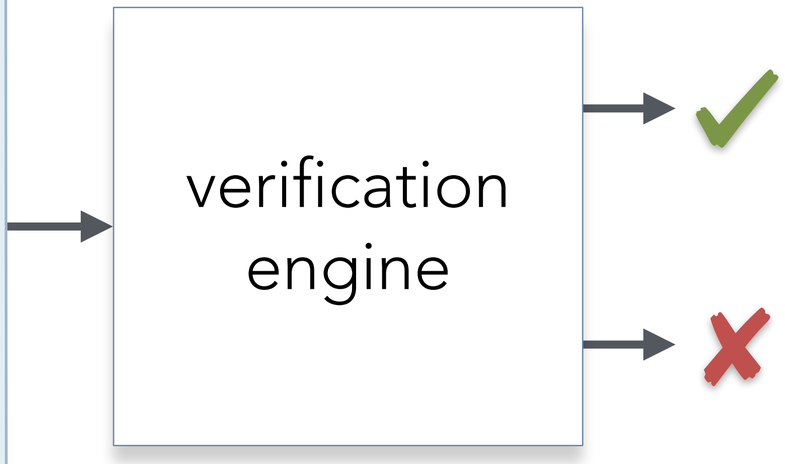
$$\vec{o}_1 = \mathrm{NN}(\vec{p})$$

linear real arithmetic

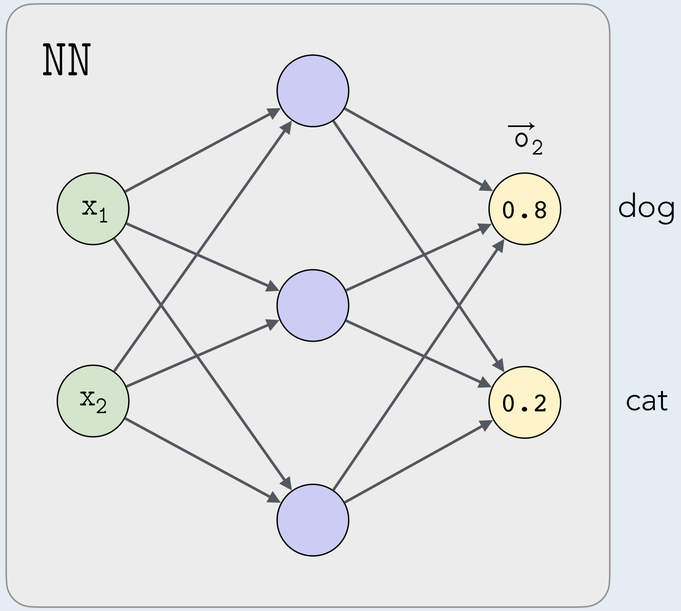
verification engine

Verifiability





Verifiability



$\models$

robustness

$$\vec{p} - \vec{x} \leq \varepsilon$$

$$\vec{o}_1 = \mathrm{NN}(\vec{p})$$

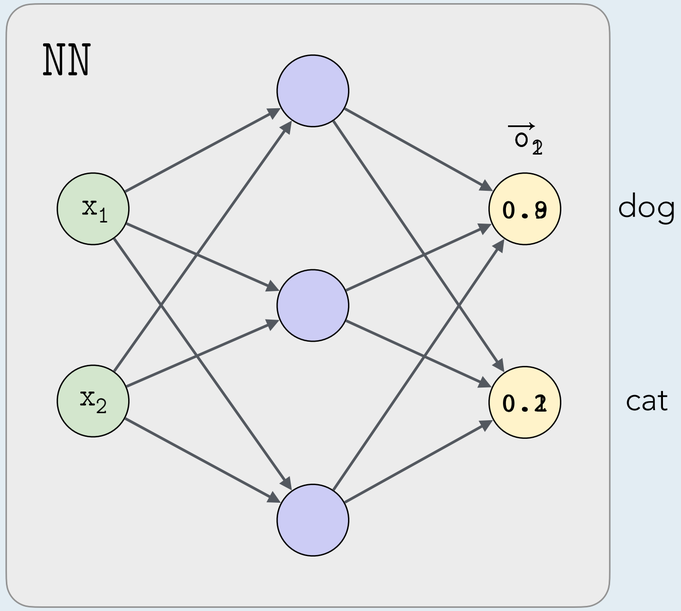
$$\vec{o}_2 = \mathrm{NN}(\vec{x})$$

$$\mathrm{class}(\vec{o}_1) = \mathrm{class}(\vec{o}_2)$$

linear real arithmetic

verification engine

Verifiability



$\models$

robustness

$$\vec{p} - \vec{x} \leq \varepsilon$$

$$\vec{o}_1 = \mathrm{NN}(\vec{p})$$

$$\vec{o}_2 = \mathrm{NN}(\vec{x})$$

$$\mathrm{argmax}(\vec{o}_1) = \mathrm{argmax}(\vec{o}_2)$$

linear real arithmetic

verification engine

EU Artificial Intelligence Act

Disclaimer: This is not legal advice.

Motivation & timeline behind the EU AI Act

[…] promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights […]

⚖️ *AI Act, Article 1*

Motivation & timeline behind the EU AI Act

[…] promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights […]

⚖️ *AI Act, Article 1*

Motivation & timeline behind the EU AI Act

[…] promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights […]

⚖️ *AI Act, Article 1*

2021

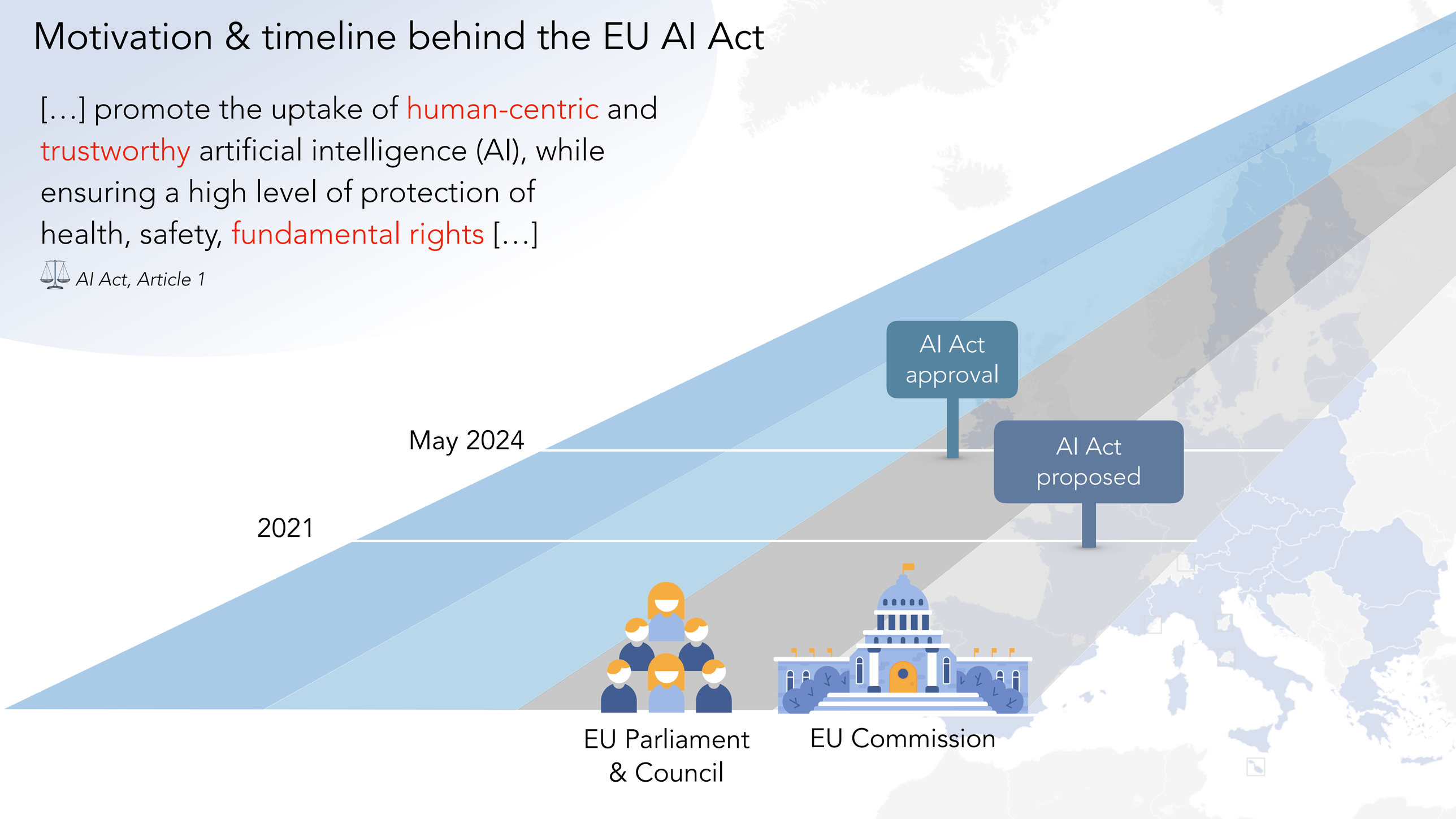
AI Act proposed

EU Commission

Motivation & timeline behind the EU AI Act

[…] promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights […]

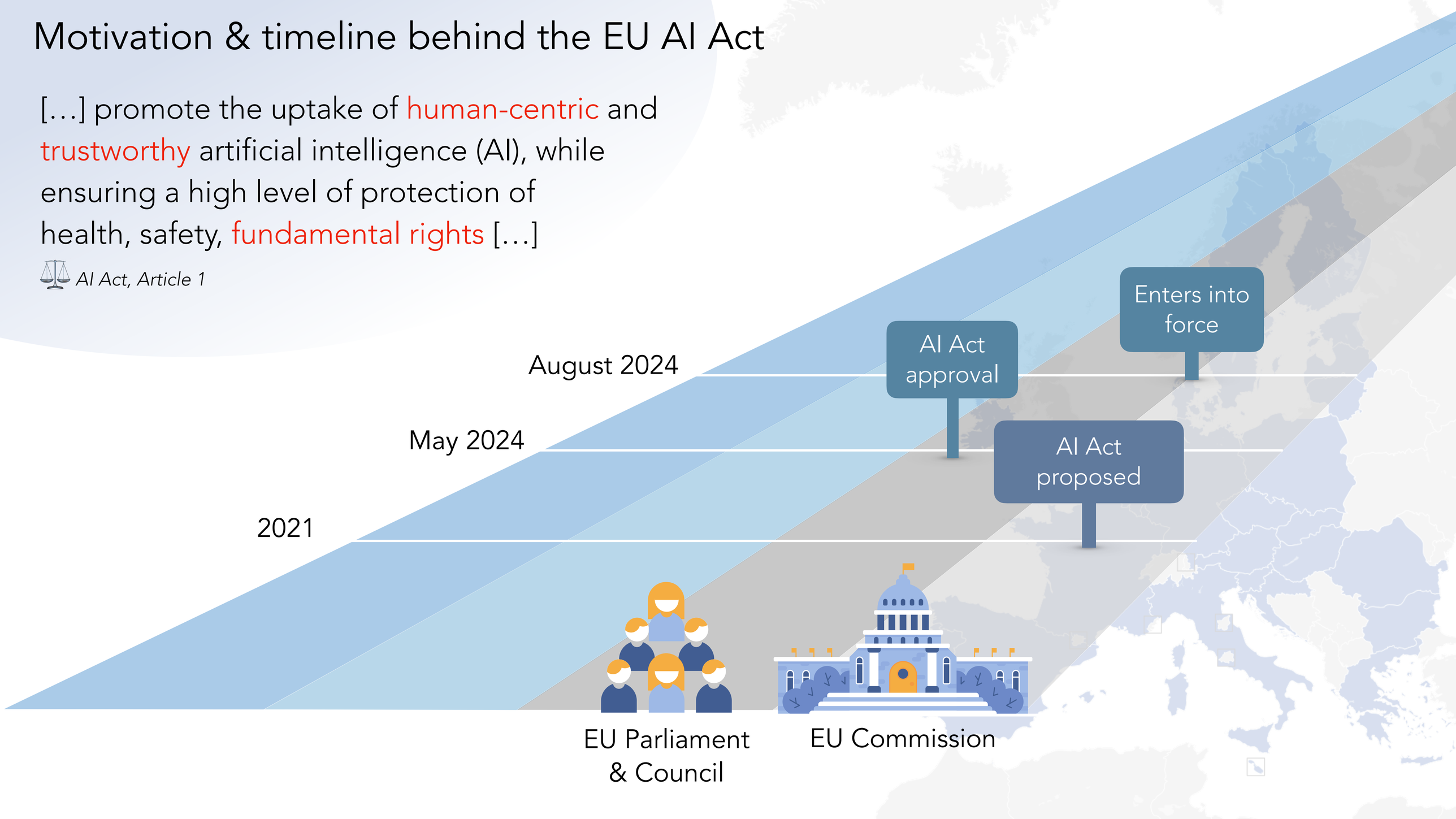
⚖️ *AI Act, Article 1*



Motivation & timeline behind the EU AI Act

[…] promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights […]

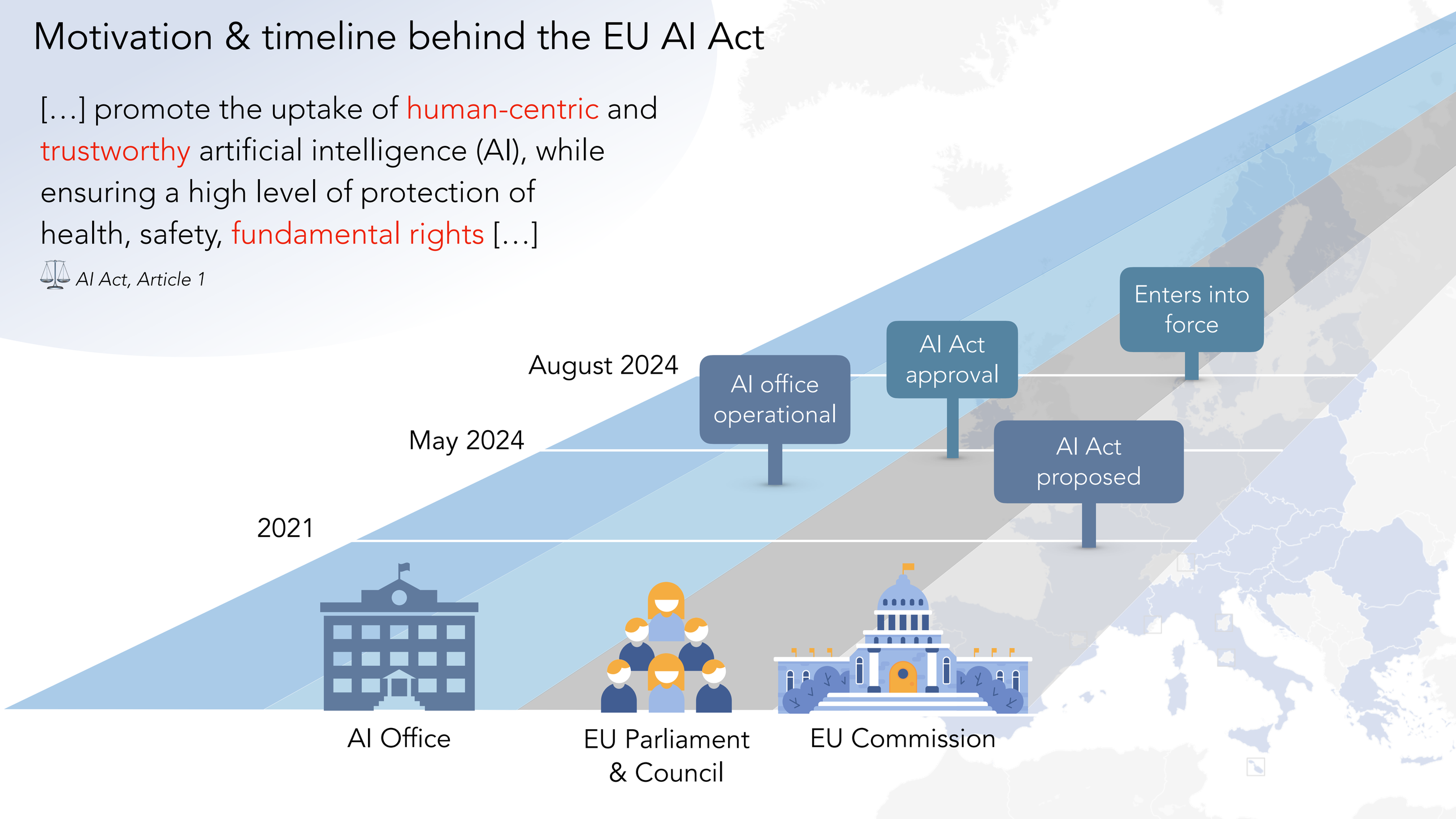
⚖️ *AI Act, Article 1*



Motivation & timeline behind the EU AI Act

[…] promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights […]

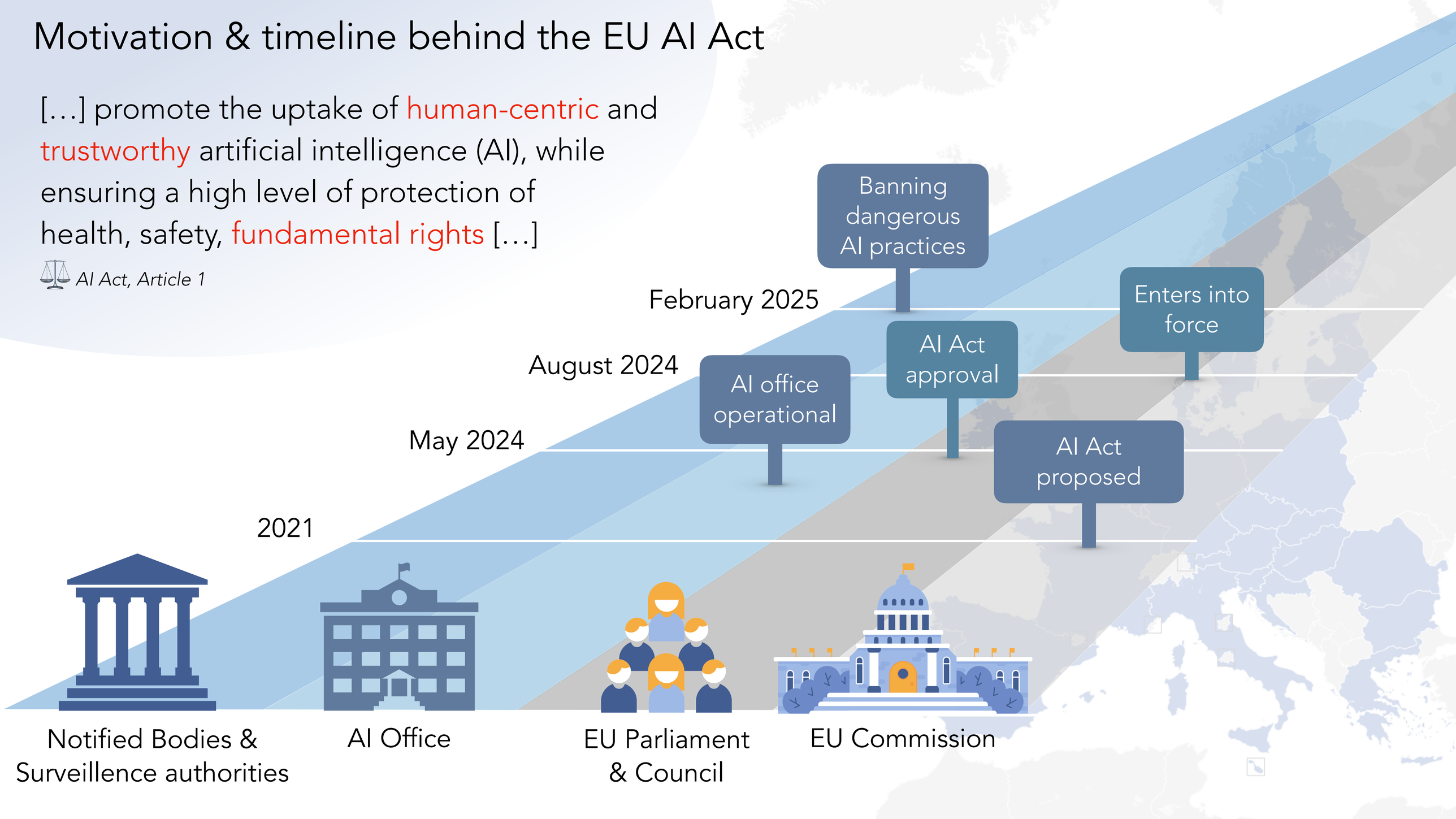
AI Act, Article 1



Motivation & timeline behind the EU AI Act

[…] promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights […]

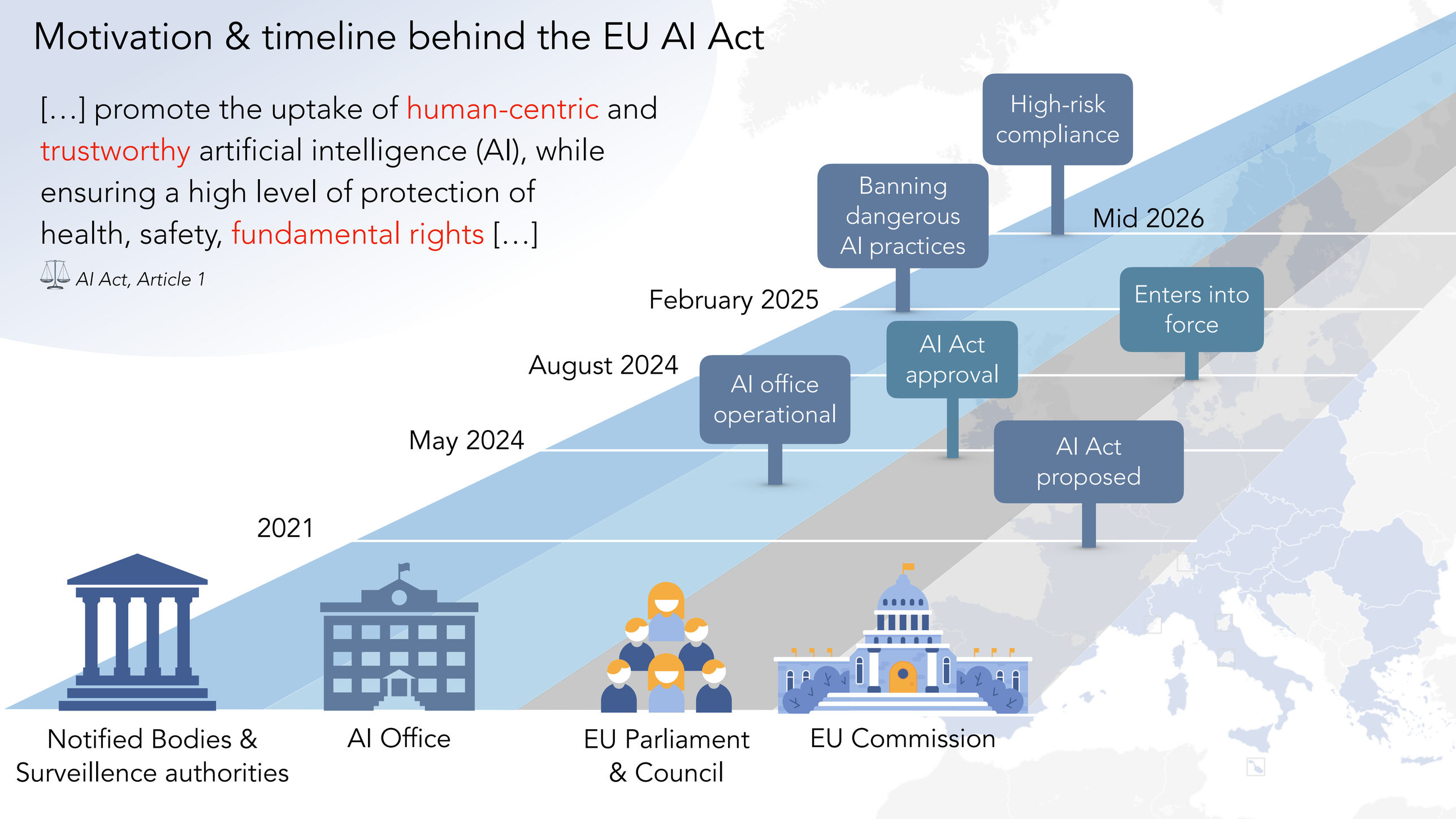
AI Act, Article 1



Motivation & timeline behind the EU AI Act

[…] promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights […]

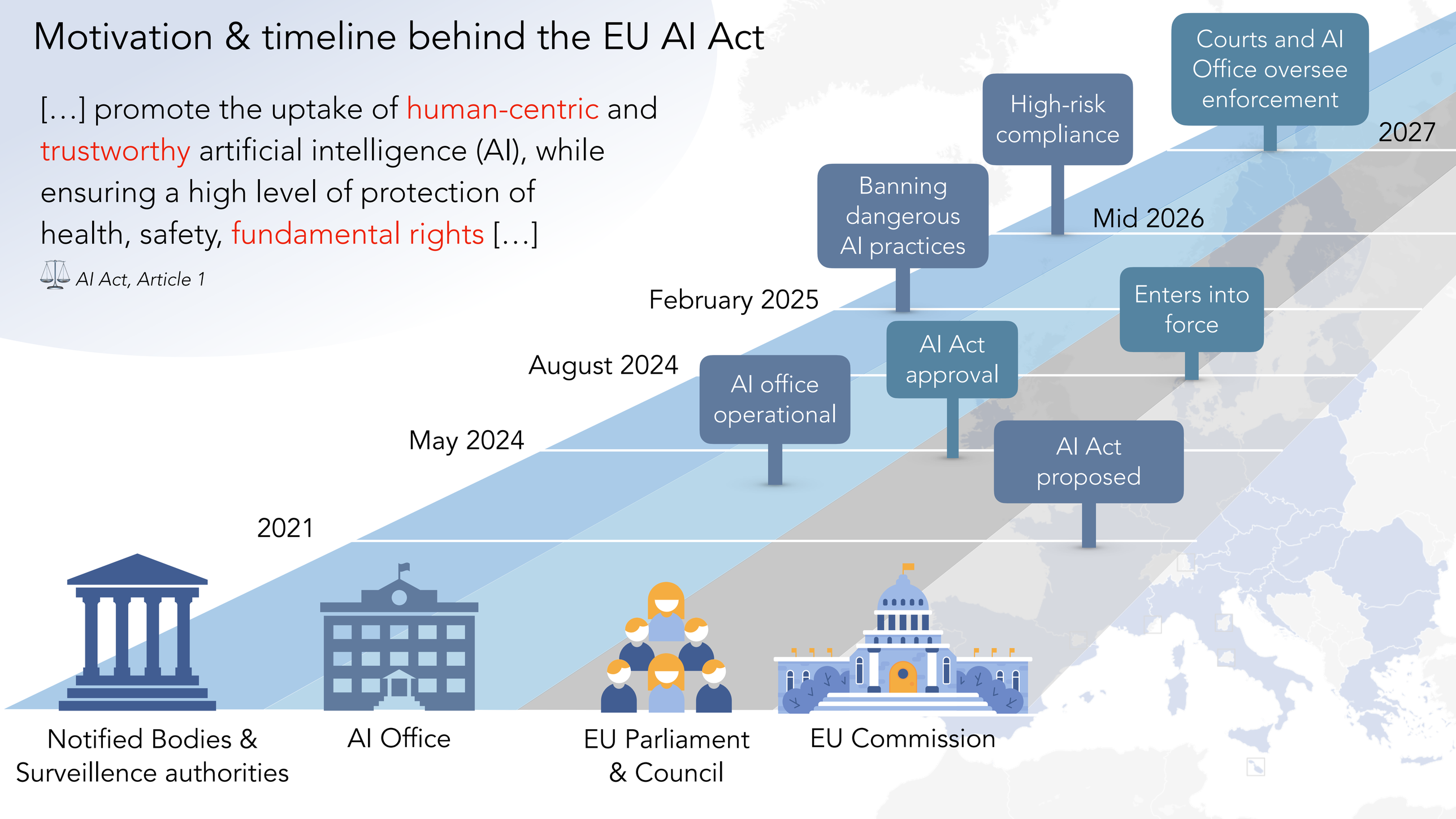
AI Act, Article 1



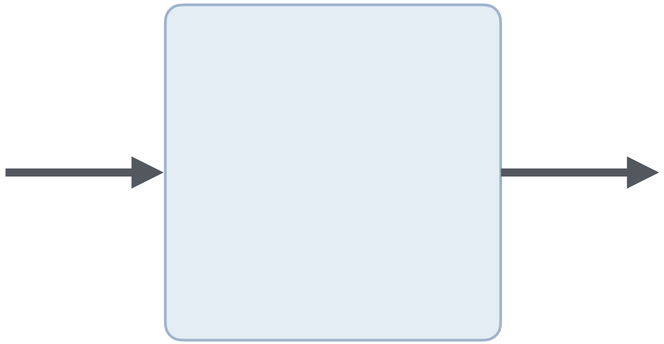
Motivation & timeline behind the EU AI Act

[…] promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights […]

AI Act, Article 1



AI systems



AI systems

When does the AI Act apply?

AI systems

Article 1

Article 2

Article 3

Article 4

⋮

[…] promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights […]

AI systems

Article 1

Article 2

Article 3

Article 4

⋮

(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means […]

(3) **provider** means […]

AI systems

Article 1

Article 2

Article 3

Article 4

⋮

(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means […]

(3) **provider** means […]

AI systems

Article 1

Article 2

Article 3

Article 4

⋮

(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means […]

(3) **provider** means […]

Recital 1

⋮

Recital 12

⋮

Recital 97

[…] the definition should be based on key characteristics of AI systems that distinguish it from simpler traditional software systems or programming approaches and should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations. A key characteristic of AI systems is their capability to infer. […]

AI systems

Article 1

Article 2

Article 3

Article 4

⋮

(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means […]

(3) **provider** means […]

Recital 1

⋮

Recital 12

⋮

Recital 97

[…] the definition should be based on key characteristics of AI systems that distinguish it from simpler traditional software systems or programming approaches and should not cover systems that are based on the rules defined solely by natural persons to automatically execute operations. A key characteristic of AI systems is their capability to infer. […]

The techniques that enable inference while building an AI system include machine learning approaches […], and logic- and knowledge-based approaches […]. The capacity of an AI system to infer transcends basic data processing by enabling learning, reasoning or modelling. [...]

AI systems

Article 1

Article 2

Article 3

Article 4

⋮

(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means […]

(3) **provider** means […]

AI system capability to infer

traditional system fully programmed

AI systems

Article 1

Article 2

Article 3

Article 4

⋮

(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means […]

(3) **provider** means […]

AI system

capability to infer

if-then-else logic (e.g., max)

fixed, human-defined rules, behave exactly as programmed

traditional system

fully programmed

AI systems

Article 1

Article 2

Article 3

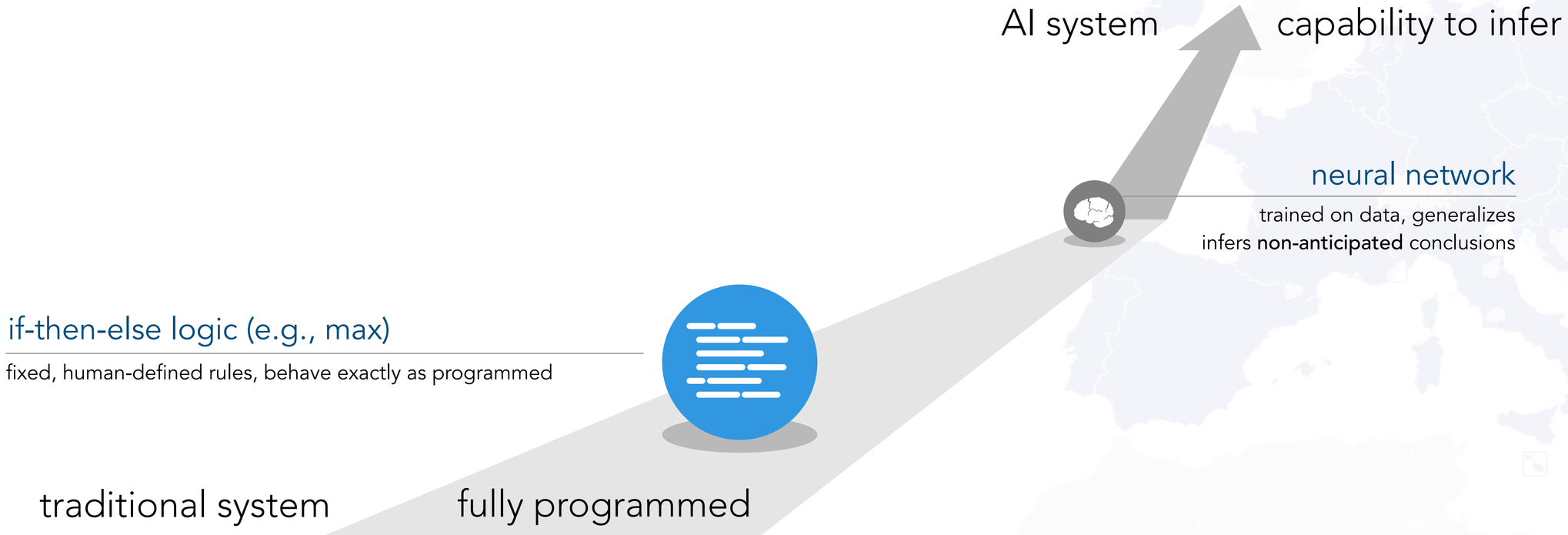
Article 4

⋮

(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means […]

(3) **provider** means […]



AI systems

Article 1

Article 2

Article 3

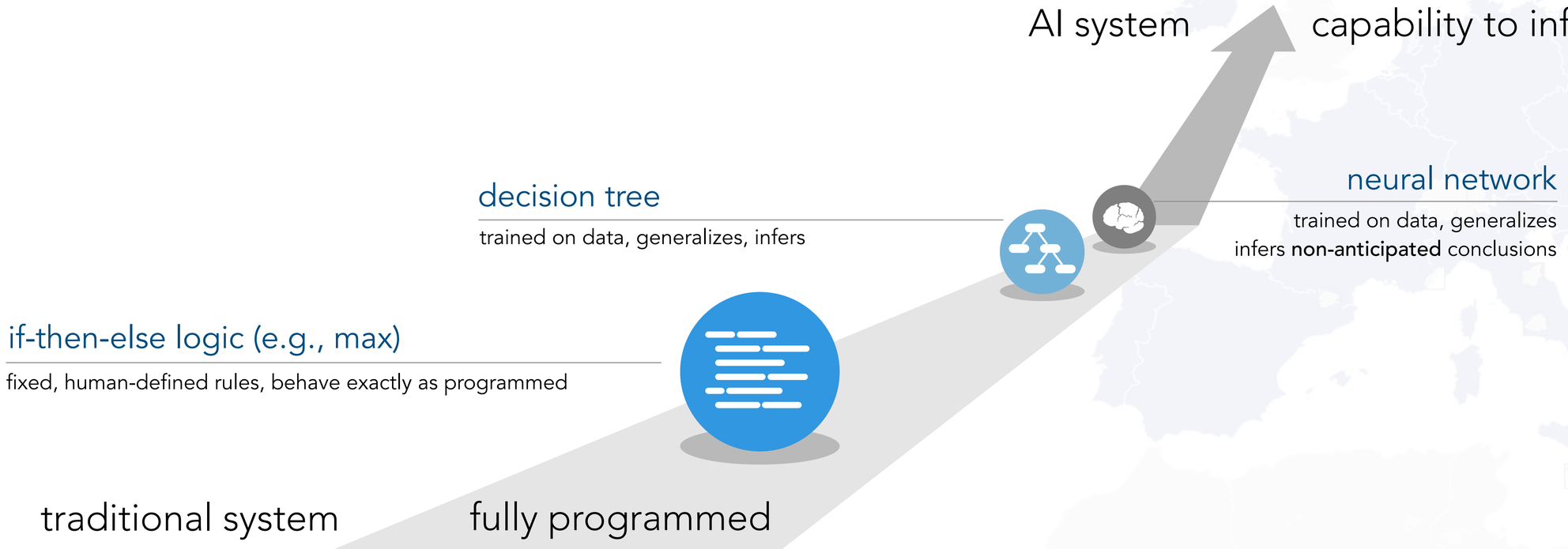
Article 4

⋮

(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means […]

(3) **provider** means […]



AI systems

Article 1

Article 2

Article 3

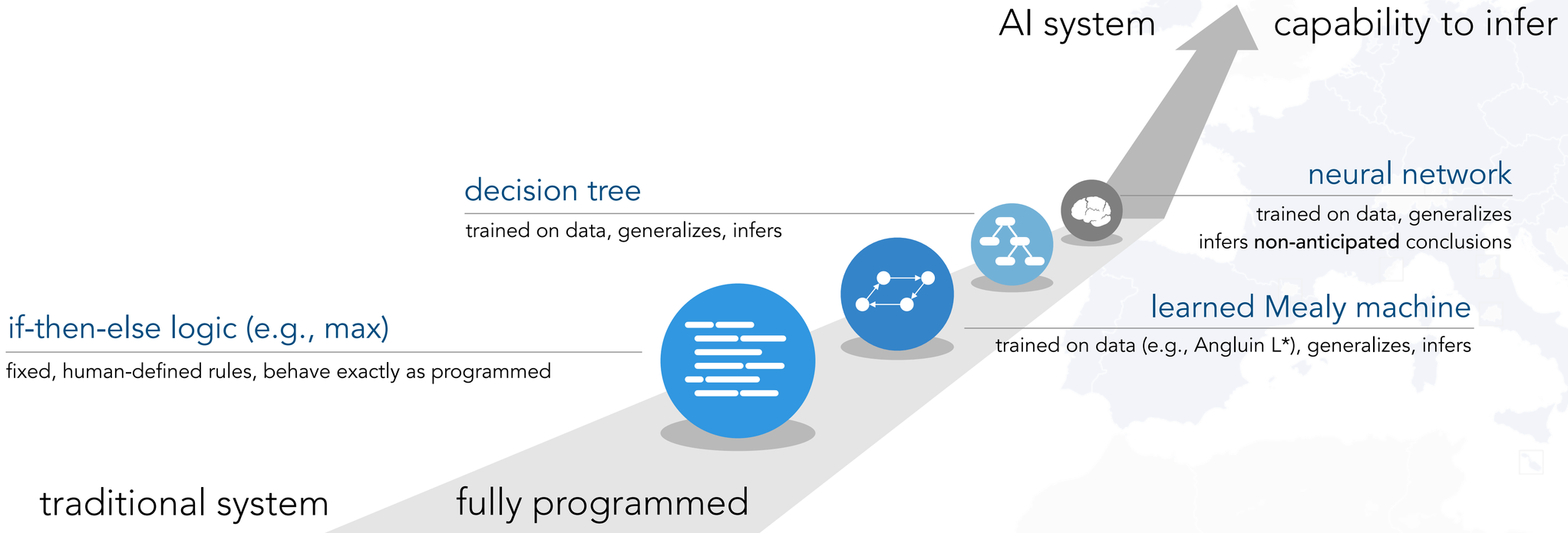
Article 4

⋮

(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means […]

(3) **provider** means […]



AI systems

Article 1

Article 2

Article 3

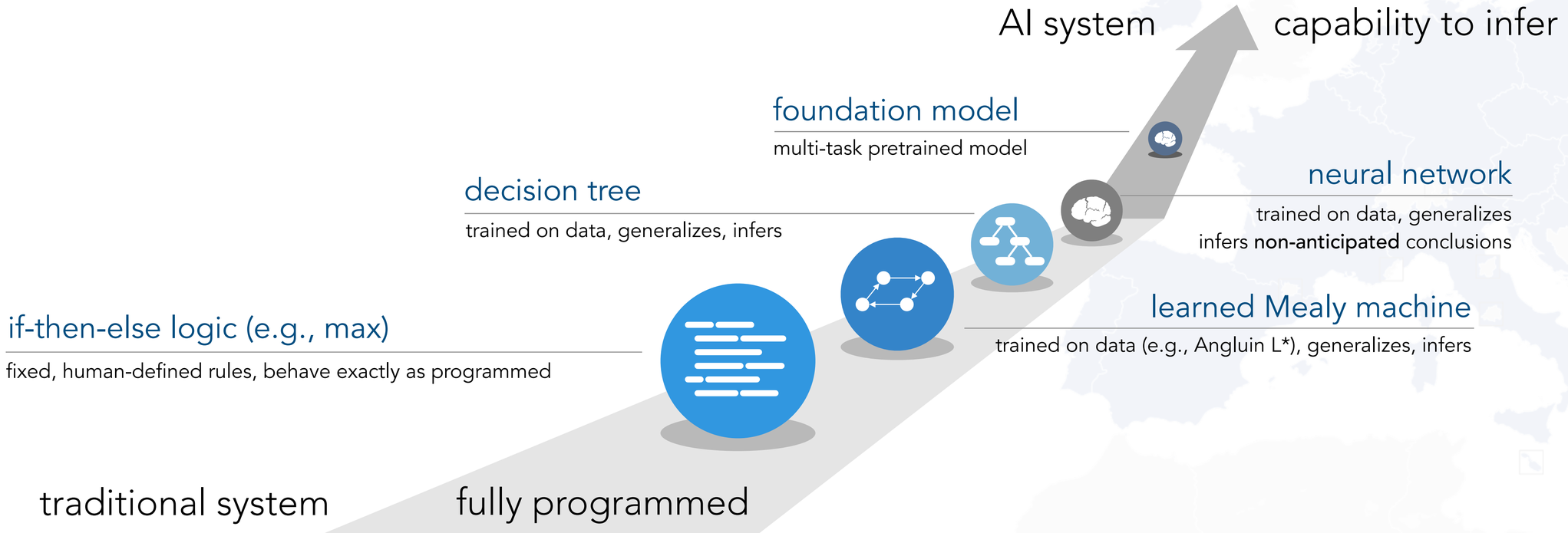
Article 4

⋮

(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means […]

(3) **provider** means […]



AI systems

Article 1

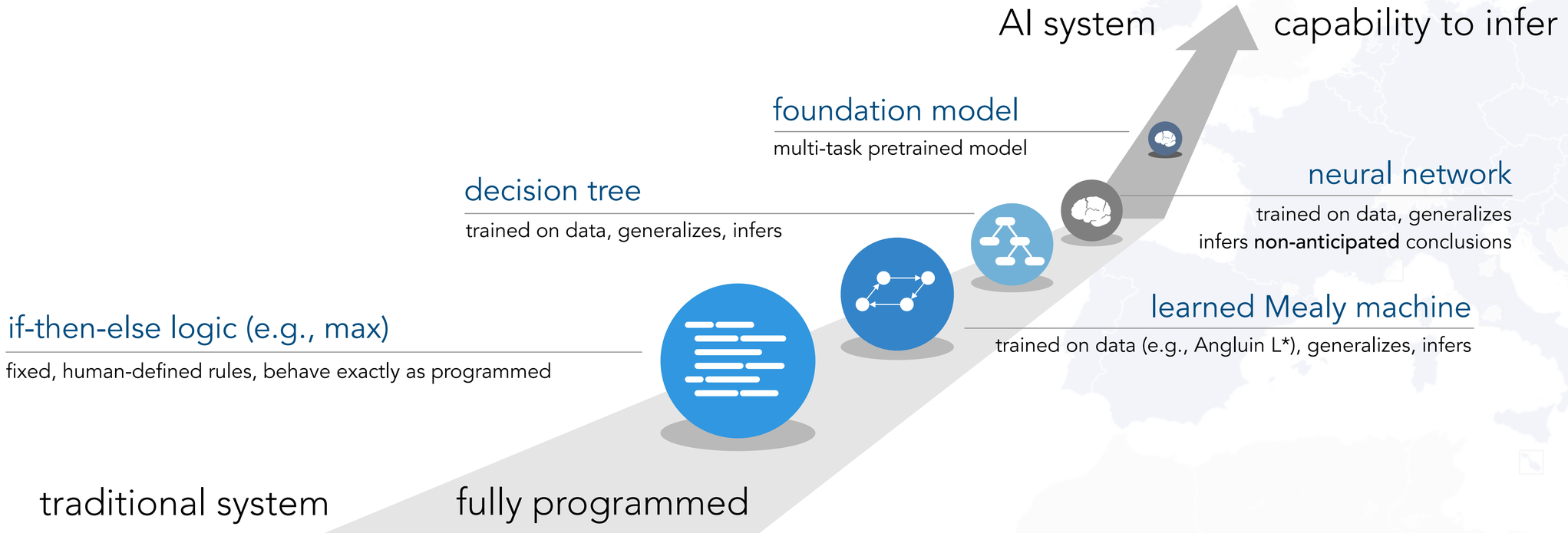
Article 2

Article 3

Article 4

⋮

(63) **general-purpose AI model** means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, […]



AI systems

Article 1

Article 2

Article 3

Article 4

⋮

(63) **general-purpose AI model** means an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, […]

- Definitions are deliberately closer to common sense than to definition a computer scientist would write.
- Arguably biased towards machine learning systems.

When does the AI Act **not** apply?

⚖️ *AI Act, Article 2 (Scope)*

- Military defence and national security
- Personal, non-professional use
- R&D and testing (non-commercial)
- Open-source/free AI components

AI system

⚖️ *Article 2 (3)*

⚖️ *Article 2 (10)*

⚖️ *Article 2 (6)*

⚖️ *Article 2 (12)*

When does the AI Act **not** apply?

⚖️ *AI Act, Article 2 (Scope)*

- Military defence and national security
- Personal, non-professional use
- R&D and testing (non-commercial)
- Open-source/free AI components

AI system

⚖️ *Article 2 (3)*

⚖️ *Article 2 (10)*

⚖️ *Article 2 (6)*

⚖️ *Article 2 (12)*

AI Act includes specific measures to support innovation (e.g., startups):

- Priority access to regulatory sandboxes to test high-risk AI systems.
- Some documentation obligations are proportionally reduced.

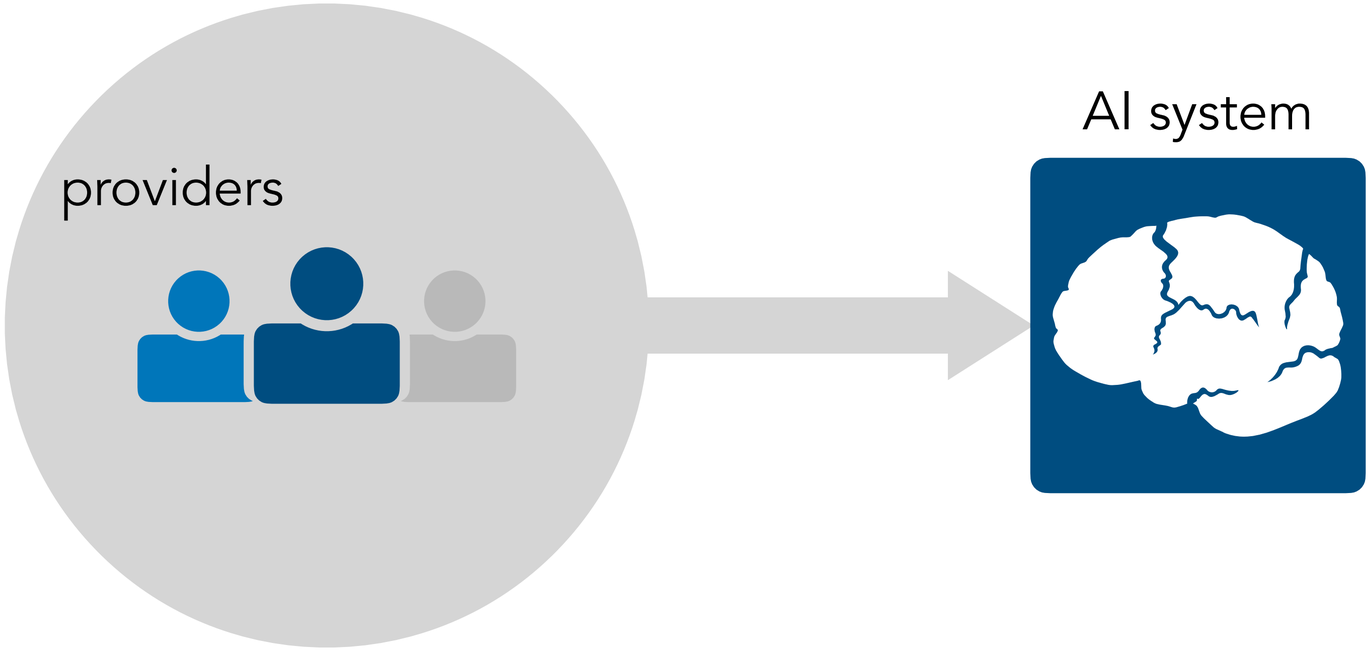
To whom does the AI Act apply?

⚖️ *AI Act, Article 2 (Scope)*

AI system

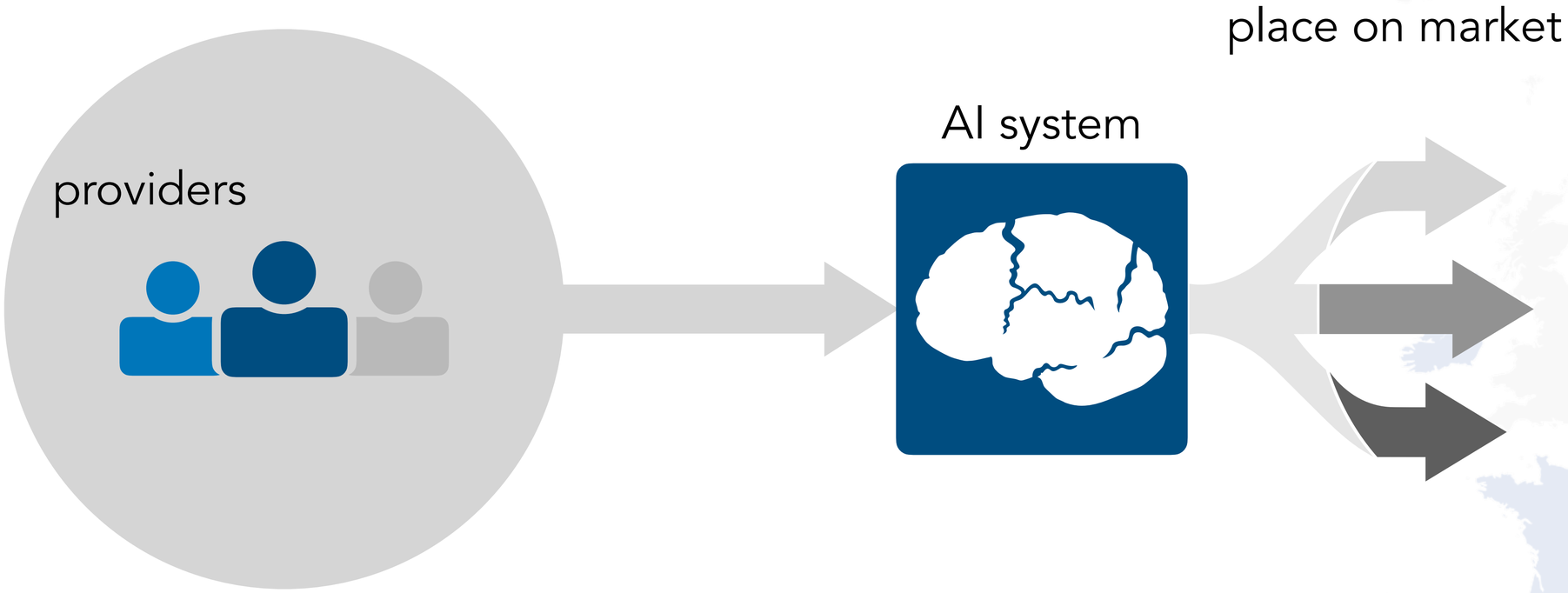
To whom does the AI Act apply?

⚖️ *AI Act, Article 2 (Scope)*



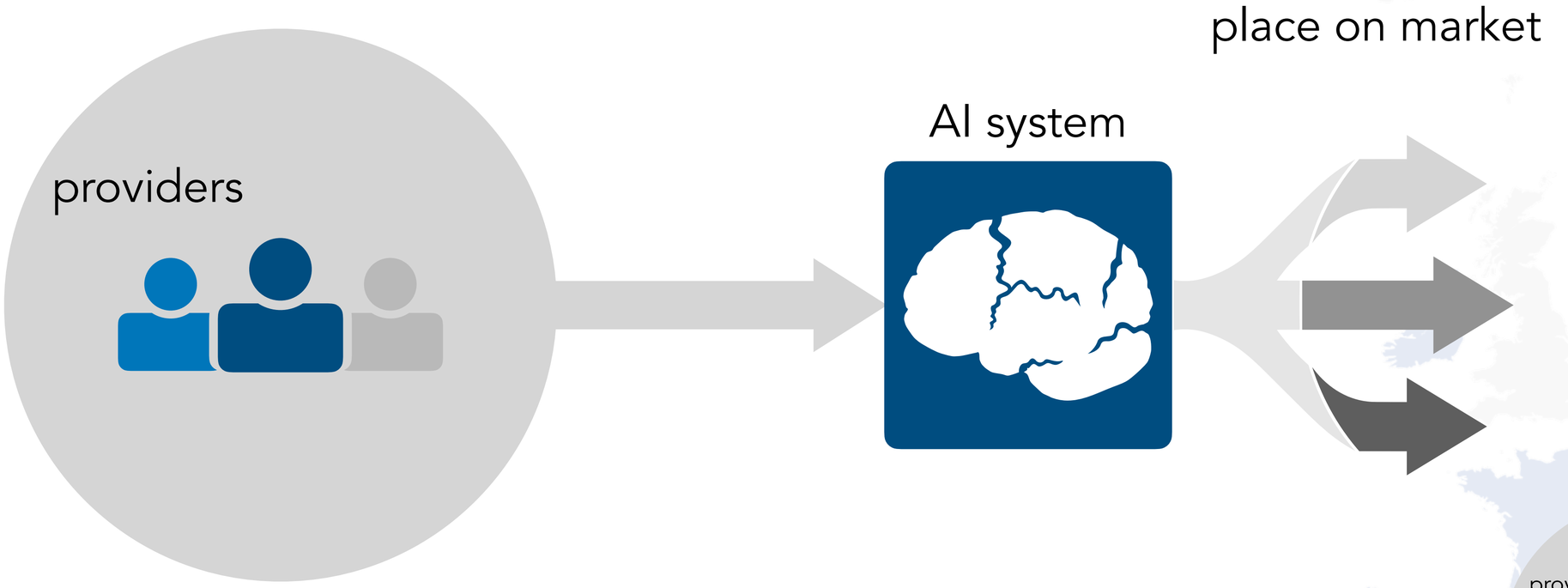
To whom does the AI Act apply?

⚖️ *AI Act, Article 2 (Scope)*



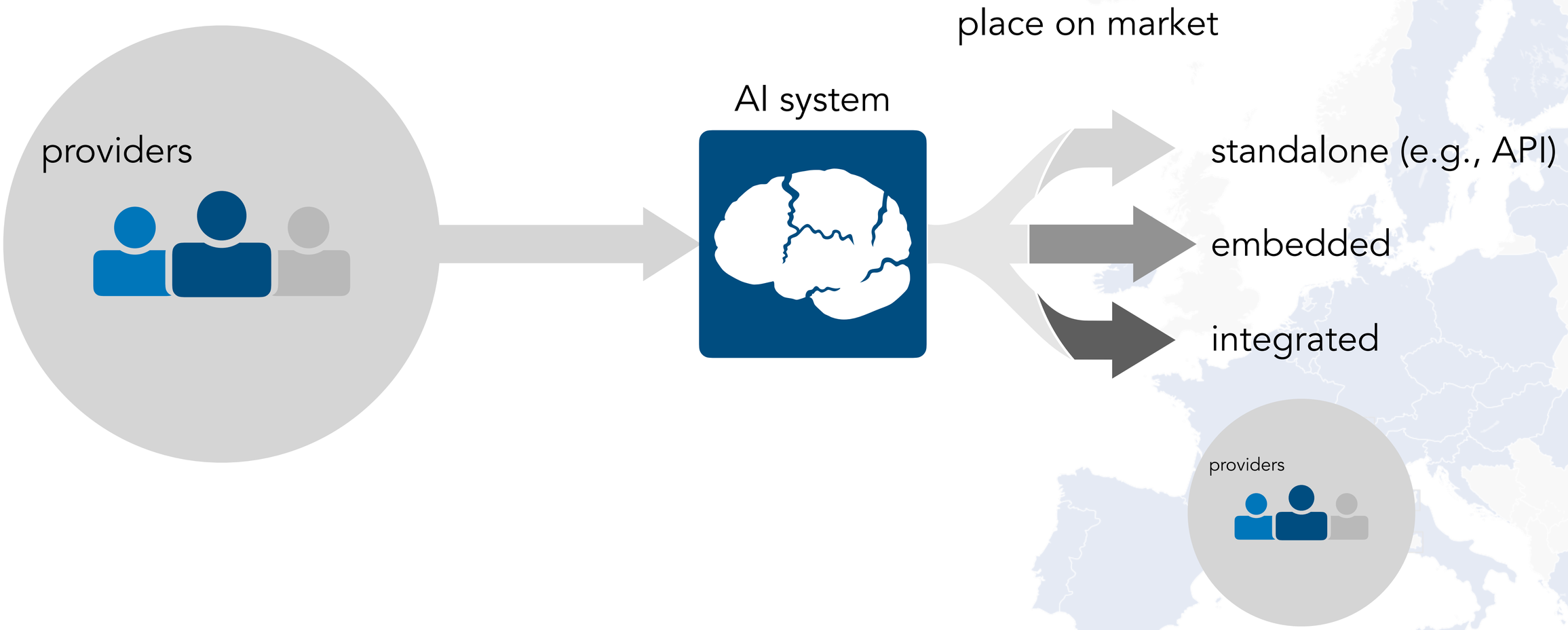
To whom does the AI Act apply?

⚖️ *AI Act, Article 2 (Scope)*



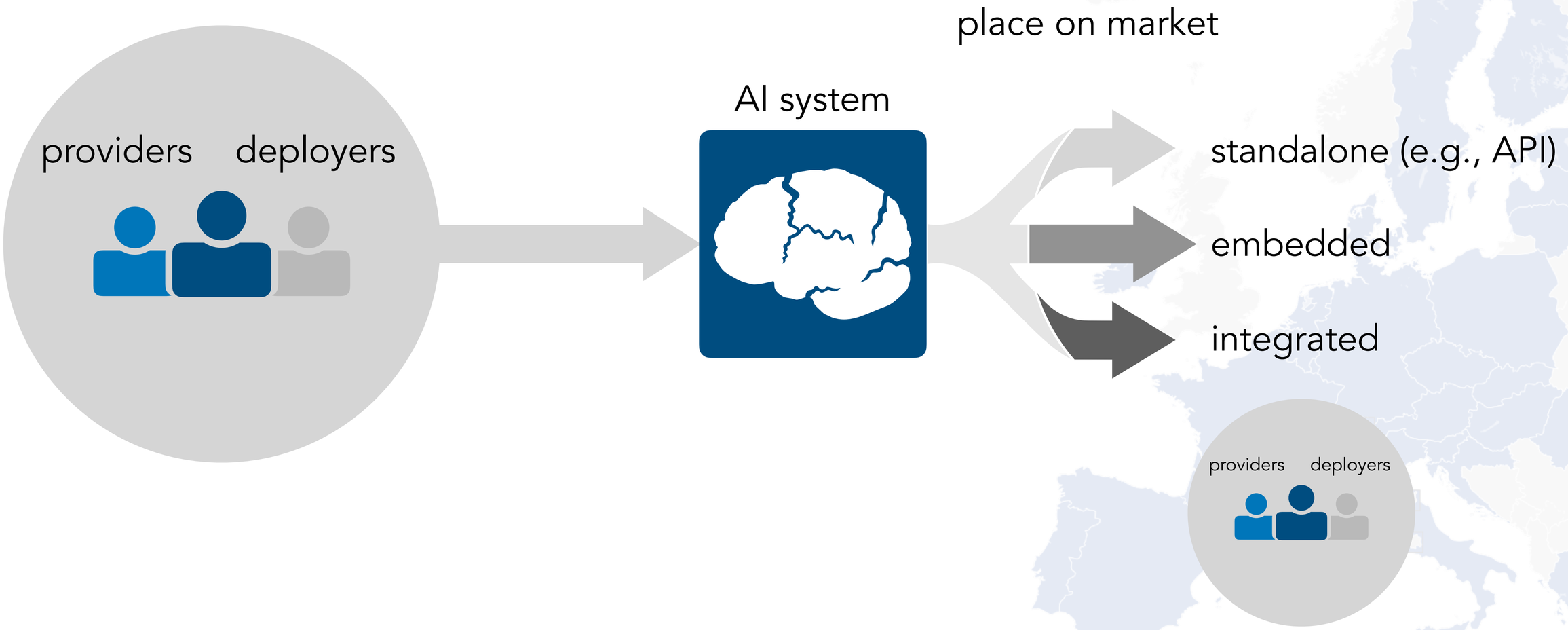
To whom does the AI Act apply?

⚖️ *AI Act, Article 2 (Scope)*



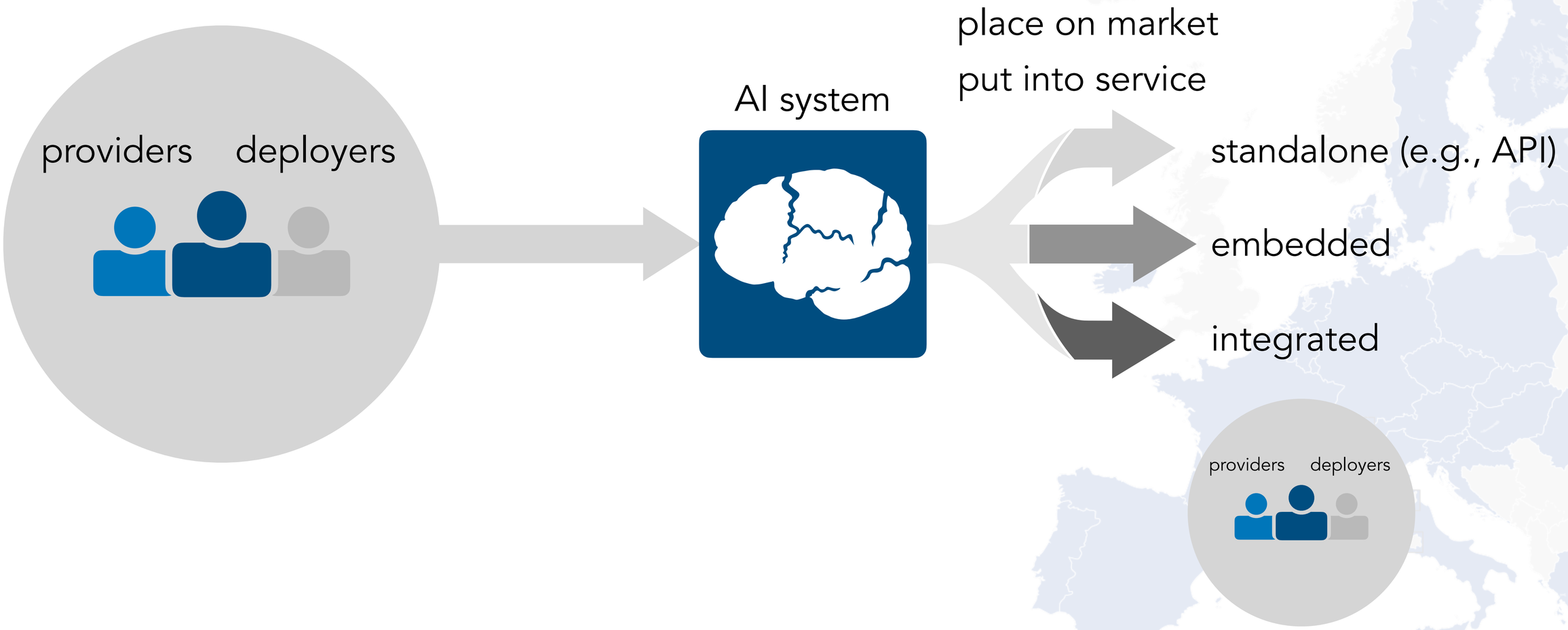
To whom does the AI Act apply?

⚖️ *AI Act, Article 2 (Scope)*



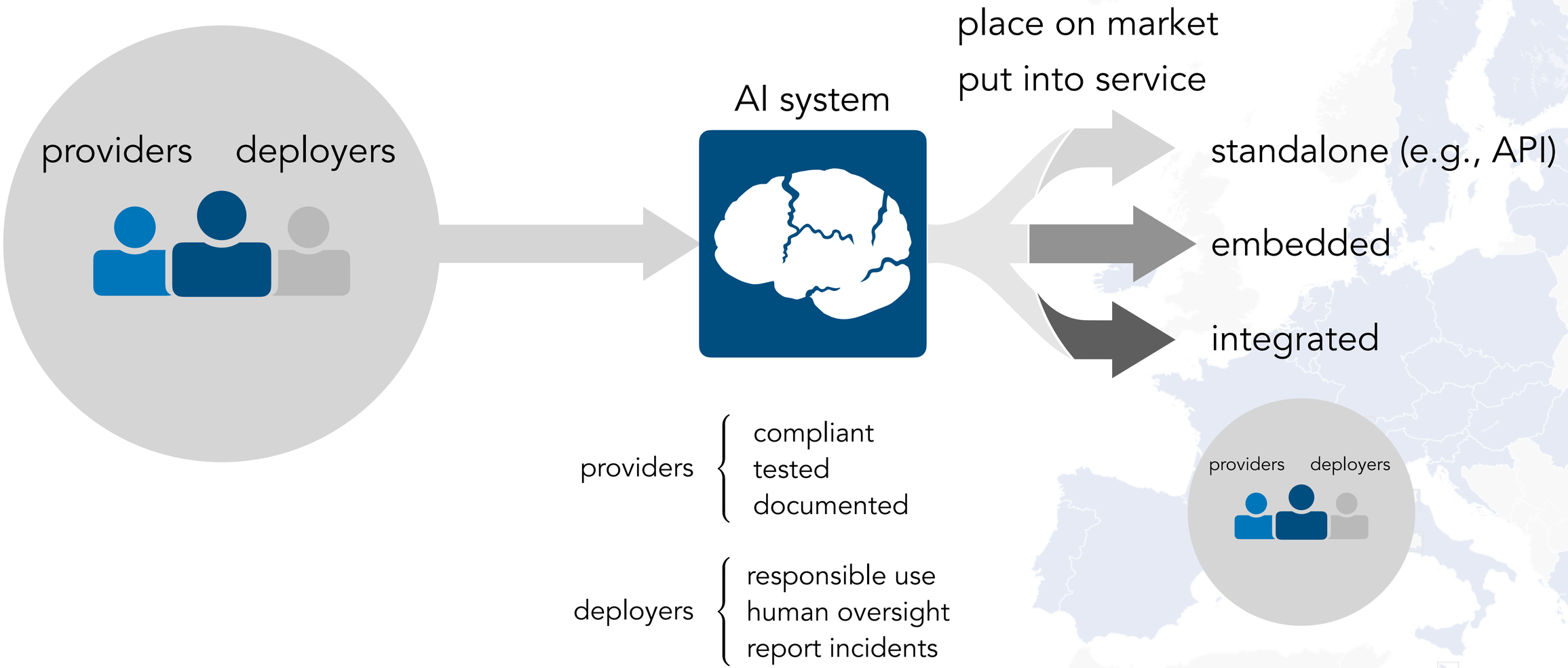
To whom does the AI Act apply?

⚖️ *AI Act, Article 2 (Scope)*



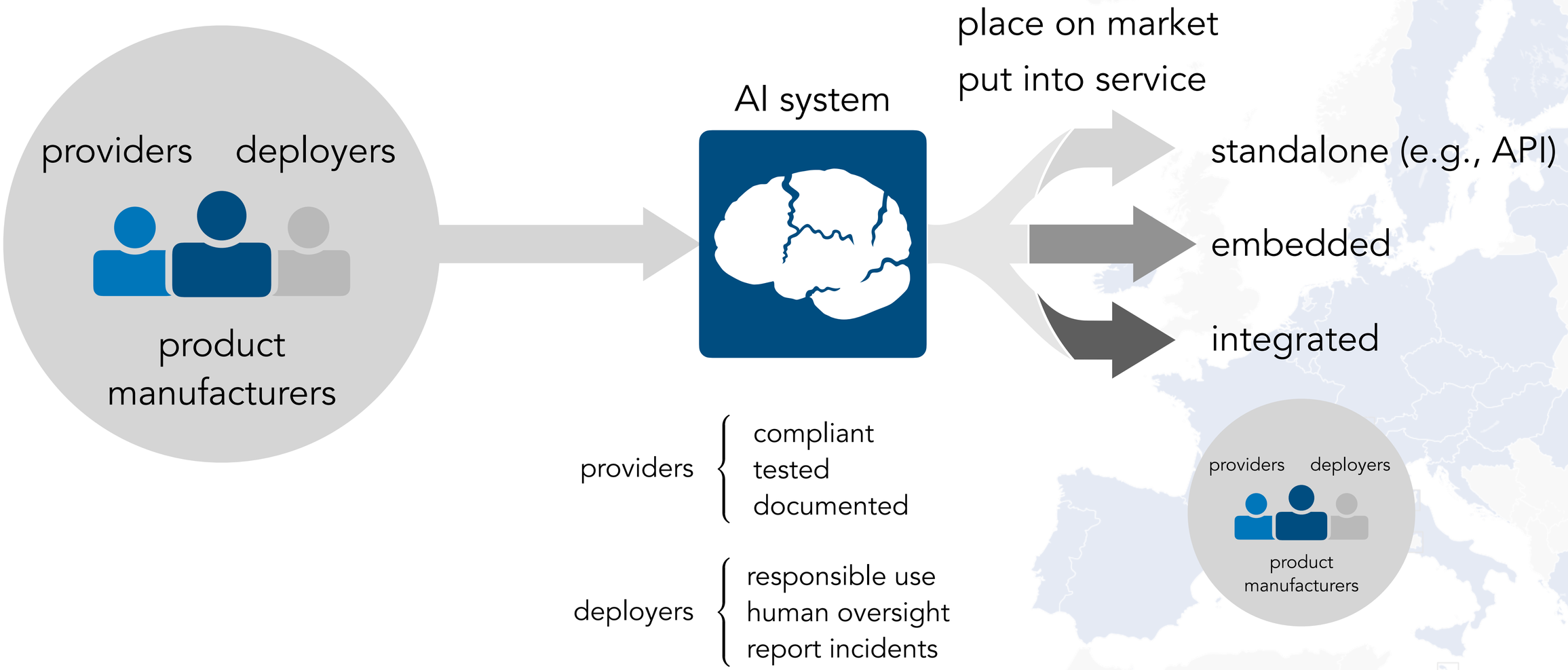
To whom does the AI Act apply?

⚖️ *AI Act, Article 2 (Scope)*



To whom does the AI Act apply?

⚖️ *AI Act, Article 2 (Scope)*



How does the AI Act apply?

Article 1

Article 2

Article 3

Article 4

⋮

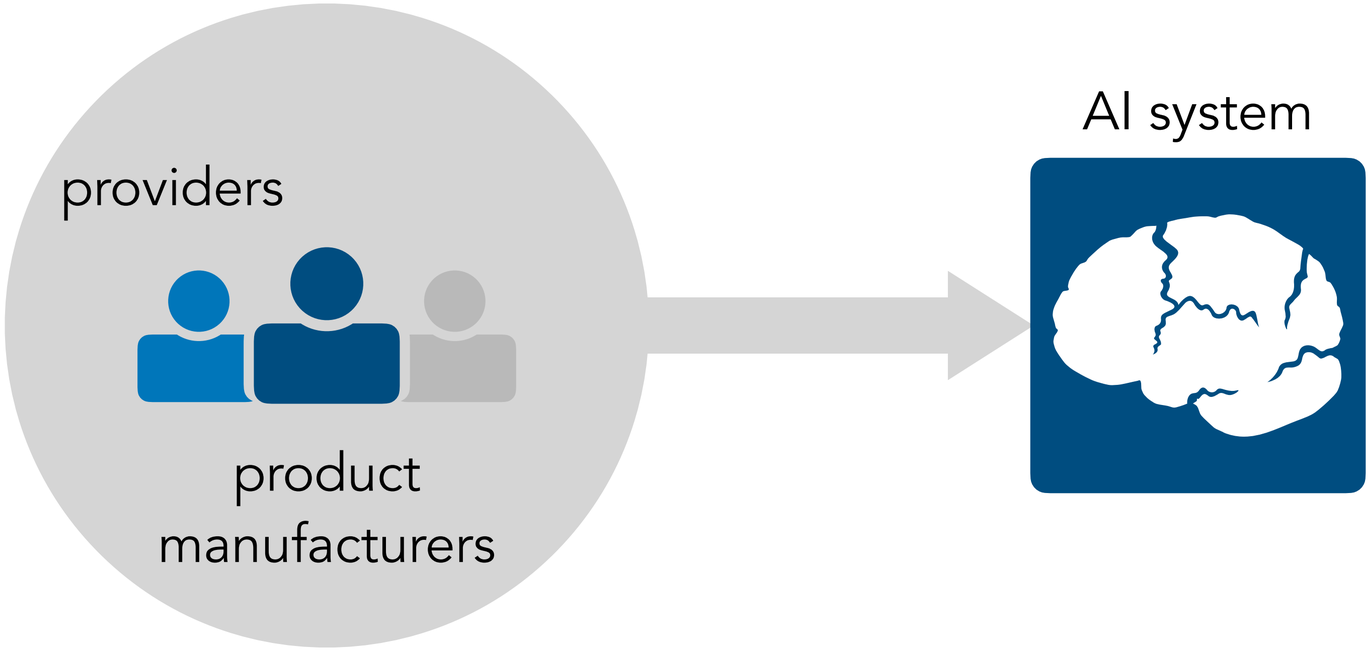
(1) **AI system** means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that […] infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments; […]

(2) **risk** means means the combination of the probability of an occurrence of harm and the severity of that harm;

(3) **provider** means means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge;

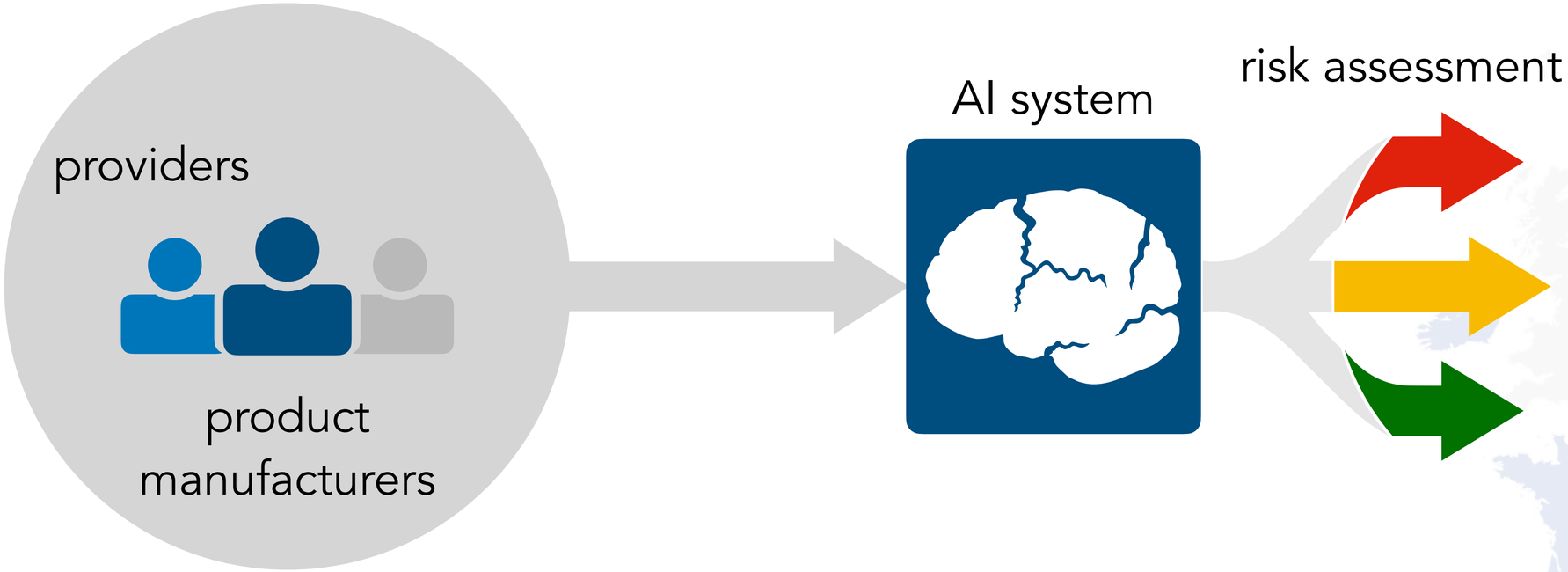
How does the AI Act apply?

⚖️ *AI Act, AI Act, Art. 5 (unacceptable risk), Art. 6 (high-risk classification)*



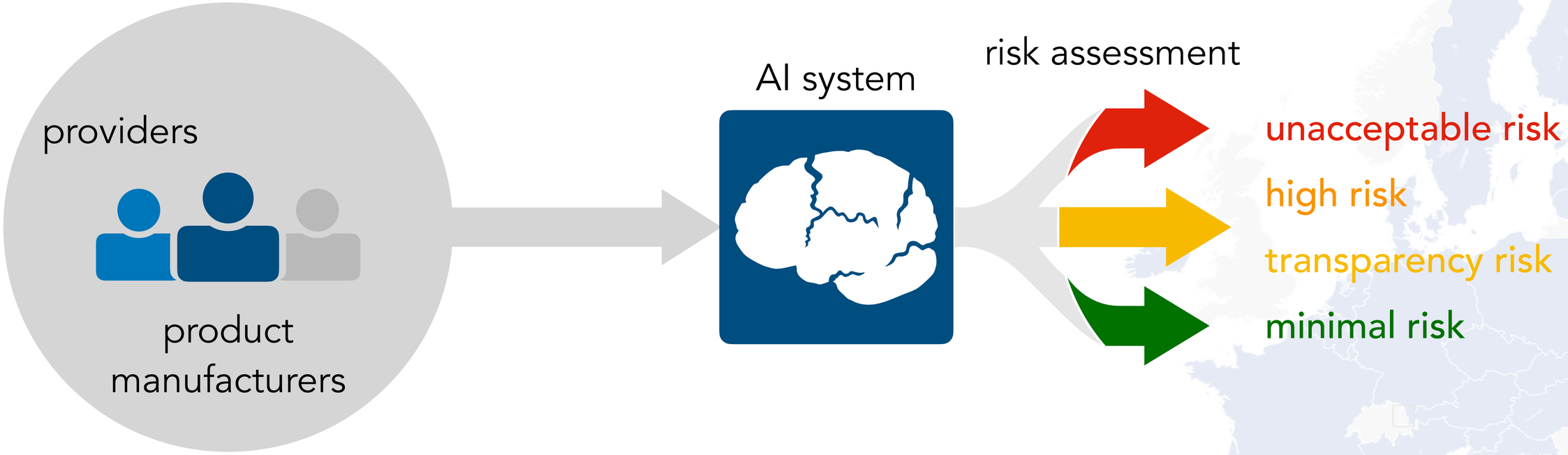
How does the AI Act apply?

⚖️ *AI Act, AI Act, Art. 5 (unacceptable risk), Art. 6 (high-risk classification)*



How does the AI Act apply?

⚖️ *AI Act, AI Act, Art. 5 (unacceptable risk), Art. 6 (high-risk classification)*

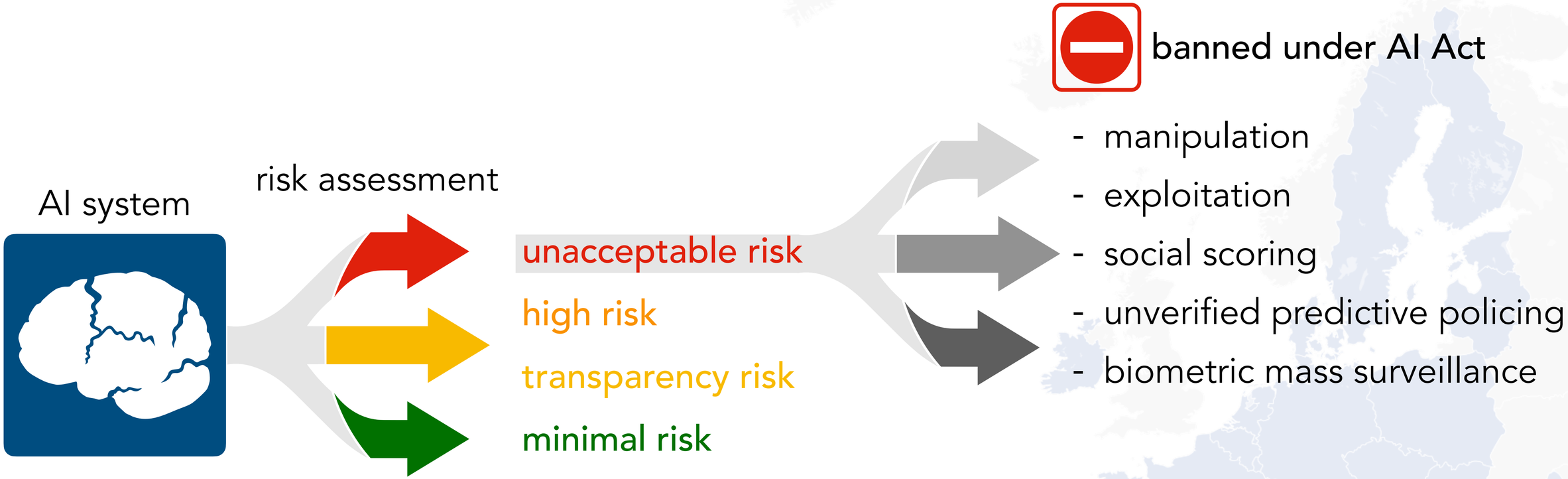


How does the AI Act apply?

⚖️ *AI Act, Article 5*

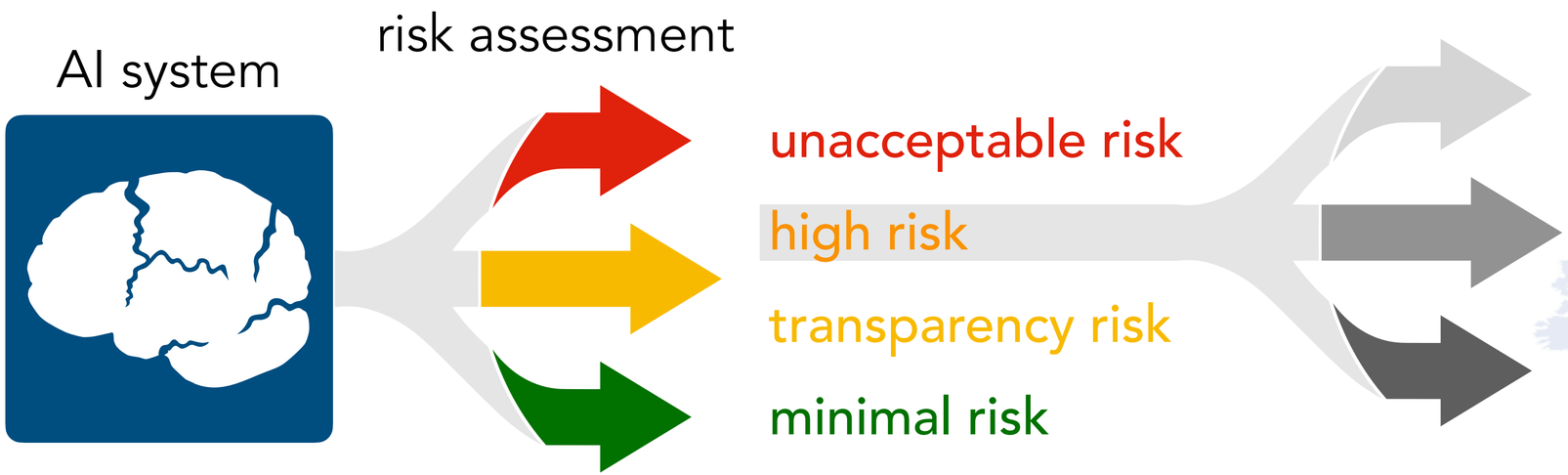
How does the AI Act apply?

⚖️ *AI Act, Article 5*



How does the AI Act apply?

⚖️ *AI Act, Annex I (safety component) & Annex III*



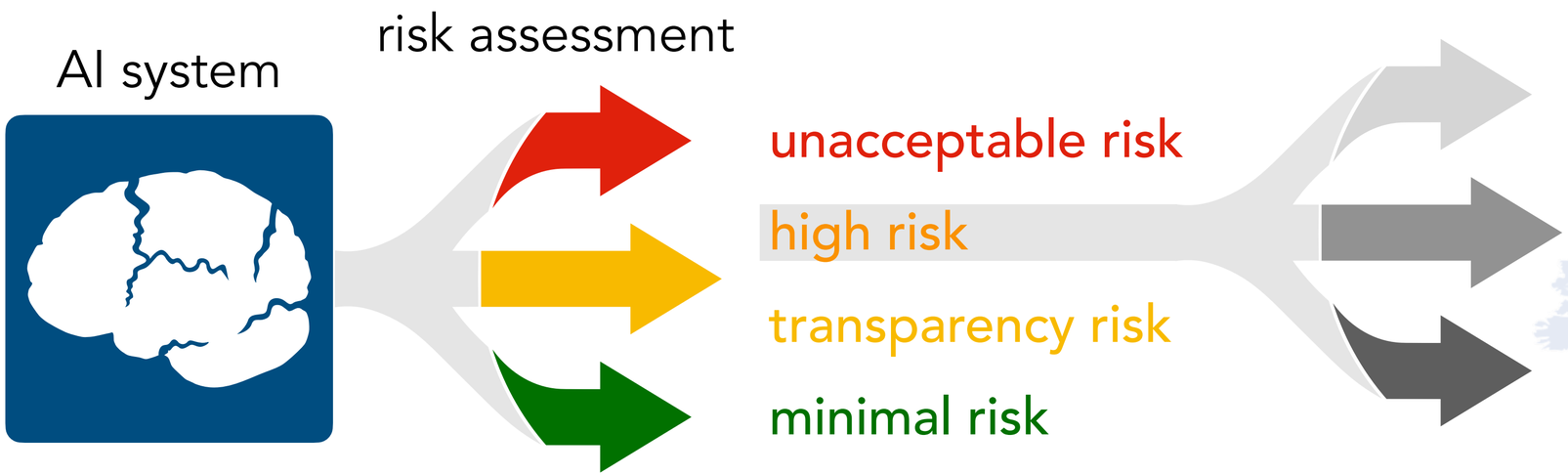
strict obligations apply

- health care
  (e.g., AI diagnosis, medical device control)
- education
  (e.g., automated grading, admission decisions)
- employment and HR
  (e.g., CV screening, hiring recommendations)
- law enforcement
  (e.g., biometric identification, crime analytics)
- critical infrastructure
  (e.g., energy grid control, water supply)
- justice
  (e.g., AI support for judges, voter influence)
- migration and border control
  (e.g., visa application assessment, lie detection)

(exceptions if task is "narrow")

How does the AI Act apply?

⚖️ *AI Act, Annex I (safety component) & Annex III*



GPAI models may trigger **systemic risk** obligations:
powerful & widespread use poses risk for society

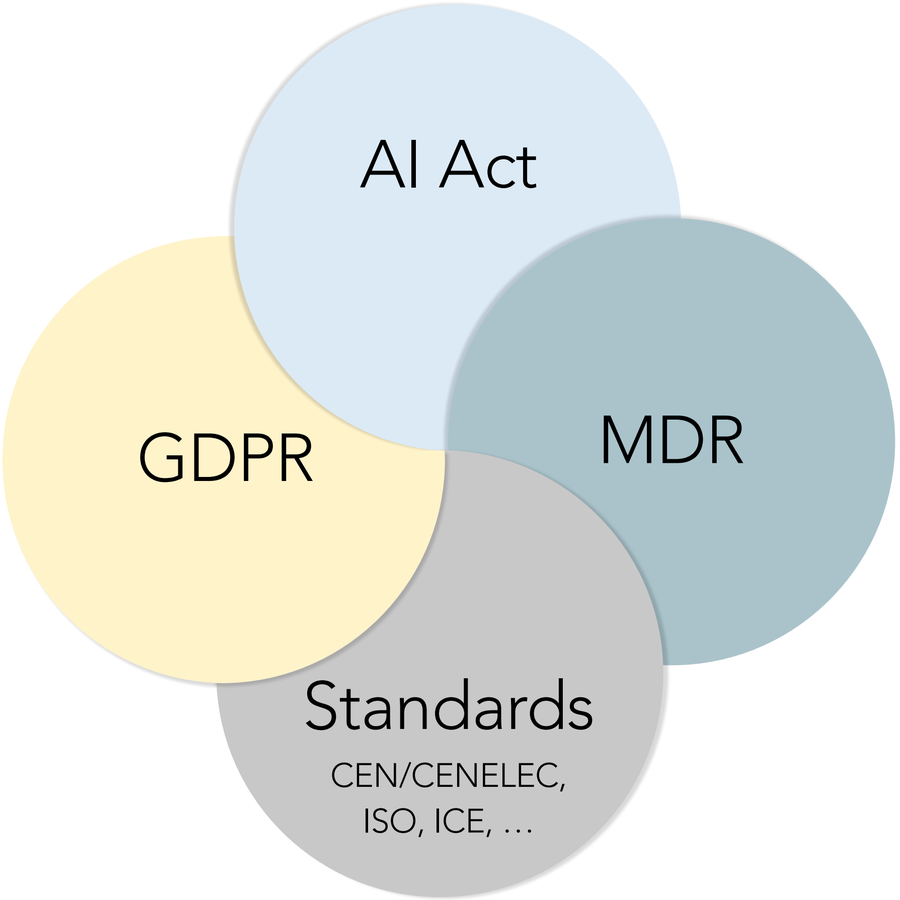
strict obligations apply

- health care
  (e.g., AI diagnosis, medical device control)
- education
  (e.g., automated grading, admission decisions)
- employment and HR
  (e.g., CV screening, hiring recommendations)
- law enforcement
  (e.g., biometric identification, crime analytics)
- critical infrastructure
  (e.g., energy grid control, water supply)
- justice
  (e.g., AI support for judges, voter influence)
- migration and border control
  (e.g., visa application assessment, lie detection)

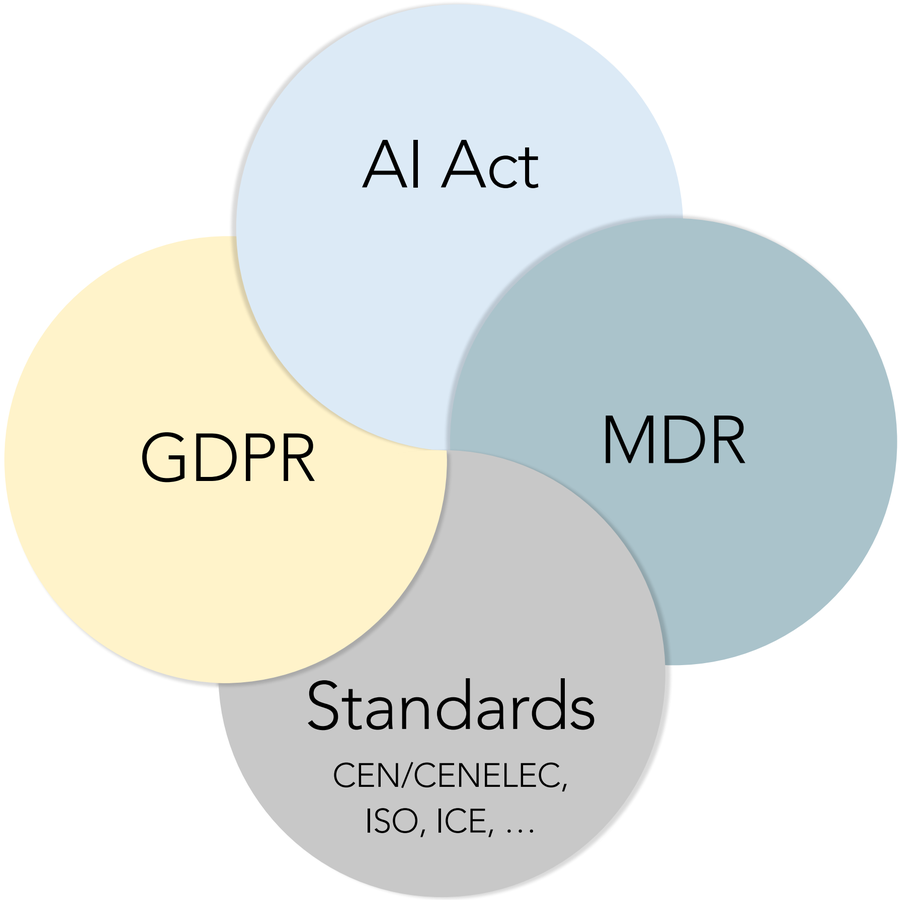
(exceptions if task is "narrow")

Obligations & Compliance Mechanisms

AI Act in the regulation landscape



AI Act in the regulation landscape

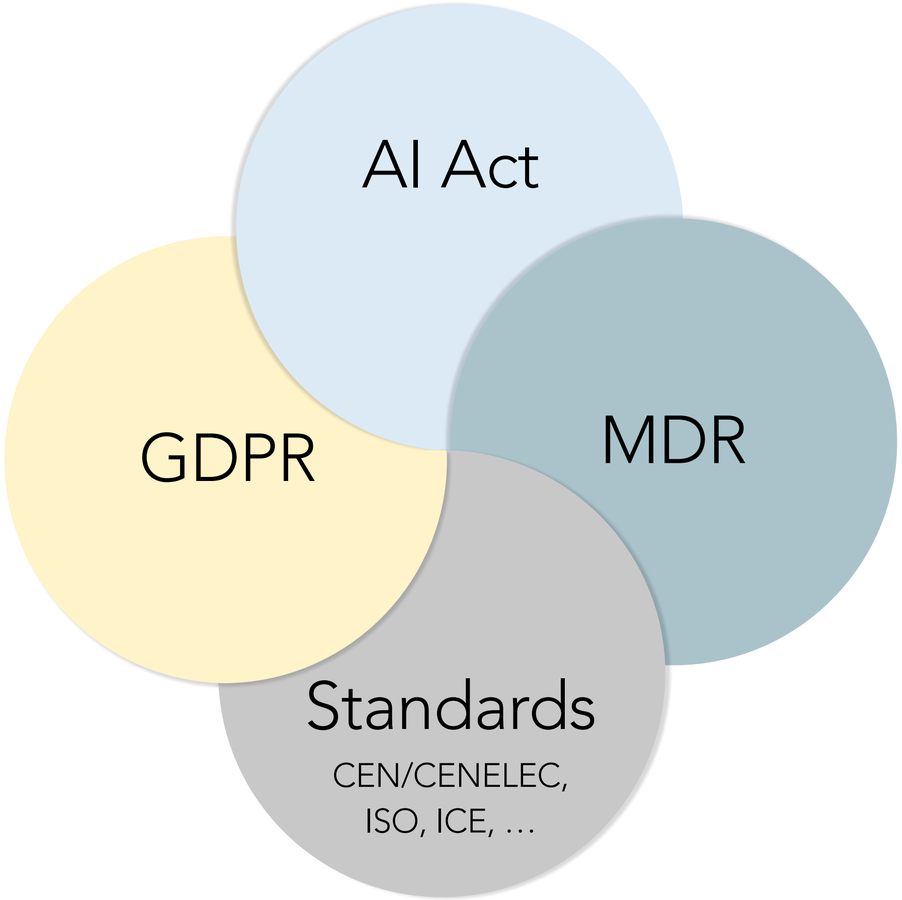


Right to be forgotten

The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay […]

⚖️ *GDPR, Article 17(1)*

AI Act in the regulation landscape



Right to be forgotten

The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay […]

⚖️ *GDPR, Article 17(1)*

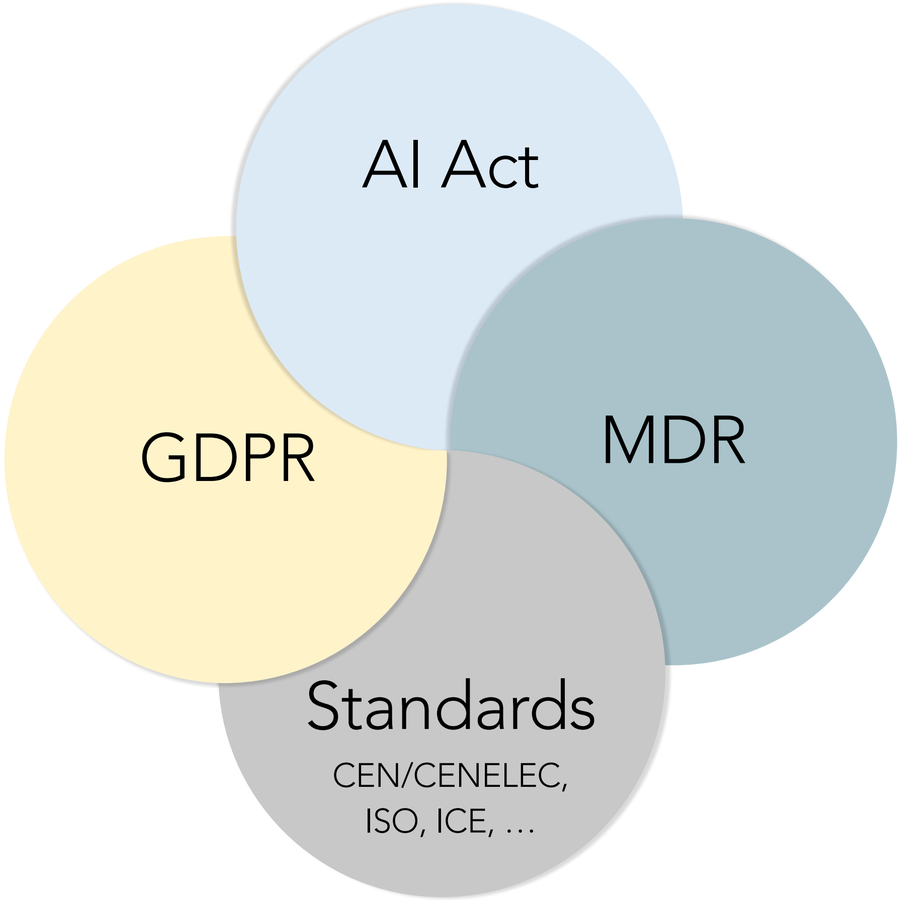
Challenge decisions

[…] the right not to be subject to a decision based solely on automated processing […]

[…] at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

⚖️ *GDPR, Article 22*

AI Act in the regulation landscape



Compliance with European harmonized standards provides a legal presumption of conformity with the regulation.

Right to be forgotten

The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay […]

⚖️ *GDPR, Article 17(1)*

Challenge decisions

[…] the right not to be subject to a decision based solely on automated processing […]

[…] at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

⚖️ *GDPR, Article 22*

Obligations for high-risk systems and GPAI models

Article	description	related attributes		
Article 5	Prohibited AI Practices	ethical	human-centric	
Article 9	Lifecycle Risk Management	safe	robust	trustworthy
Article 10	Data Quality and Fairness	fair	ethical	trustworthy
Article 11	Technical Documentation	transparent	verifiable	accountable
Article 13	Transparency and Explainability	explainable	interpretable	transparent
Article 14	Human Oversight	human-centric	trustworthy	accountable
Article 15	Accuracy and Robustness	safe	robust	trustworthy
Article 24	Conformity Assessment	verifiable	accountable	
Article 55	GPAI Obligations	trustworthy	ethical	human-centric
Article 72	Post-Market Monitoring	accountable	trustworthy	robust

Obligations for high-risk AI systems

Article 10 **Data and Data Governance**

⋮

(3) Training, validation and testing data sets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used. […]

fair | ethical | trustworthy

ISO/IEC TR 24027:2021 — Bias in AI Systems

- Detect bias in data (e.g., sampling)
- Mitigate bias in data (e.g., re-sampling, re-weighting)
- Bias detection metrics for models (e.g., statistical parity)

Obligations for high-risk AI systems

⋮

Article 13

⋮

Transparency and Provision of Information to Deployers

(1) High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. […]

explainable | interpretable | transparent

ISO/IEC TR 24028:2020 — Trustworthiness in AI

- Local surrogate models for per-decision explanations (e.g., LIME)
- Feature attribution techniques (e.g., SHAP-like methods)
- Explanation methods should be tailored to the user's role and technical background, ensuring that outputs are interpretable and actionable in context

More formal-methods opportunities

Human Oversight ⚖️ AI Act, Article 14

[Natural persons] shall be enabled, as appropriate and proportionate […] to intervene […]

Post-Market Monitoring ⚖️ AI Act, Article 72

Providers shall establish and document a post-market monitoring system […]

Process Mining

Build process models and verify human oversight.

[Pery, Rafiei, Simon, van der Aalst: *Trustworthy Artificial Intelligence and Process Mining: Challenges and Opportunities.* ICPM Workshops 2021: 395-407]

Runtime Monitoring

Monitor log files at runtime to detect deviations.

[Colombo, Pace, Seychell: *Runtime Verification and AI: Addressing Pragmatic Regulatory Challenges.* AISoLA 2024: 225-241]

Transparency ⚖️ AI Act, Article 13

[…] enable deployers to interpret a system's output and use it appropriately. […]

Automata Learning

Learn automata as surrogate models of RNNs.

[Bollig, Leucker, Neider: *A Survey of Model Learning Techniques for Recurrent Neural Networks.* Lecture Notes in Computer Science 13560, 2022: 81-97]

Logical Reasoning

Outsource logical reasoning in LLMs.

[Liu, Xu, Huang, Wang, Wang, Yang, Li: Logic-of-Thought: Injecting Logic into Contexts for Full Reasoning in Large Language Models. CoRR abs/2409.17539 (2024)]

Conclusion: The AI Act as a work in progress

- The AI Act requires high-risk AI systems to be fair, robust, and explainable.
- Logic and formal methods can help make these goals precise and verifiable.
- We are in a transition phase: The legal framework is in place, but the technical standards are still being developed.
- Great opportunity for formal-methods researchers to contribute.

References

Explainability

- Darwiche: *Logic for Explainable AI.* LICS 2023.
- Marques-Silva: *Logic-Based Explainability: Past, Present and Future.* ISoLA (4) 2024.
- Marques-Silva, Ignatiev: *Delivering Trustworthy AI through Formal XAI.* AAAI 2022.

Verification of neural networks

- Albarghouthi: *Introduction to Neural Network Verification.* 2021. http://verifieddeeplearning.com
- Bollig: *Verification of Neural Networks.* MPRI Lecture Notes, 2024. https://lmf.cnrs.fr/downloads/BenediktBollig/nn_verification.pdf
- Katz, Barrett, Dill, Julian, and Kochenderfer: *Reluplex: An efficient SMT solver for verifying deep neural networks.* CAV 2017.
- Zhang, Xu, Wang, Hsieh: *Formal Verification of Deep Neural Networks: Theory and Practice.* 2022. https://neural-network-verification.com

Formal methods and machine learning

- Urban, Miné. *A Review of Formal Methods applied to Machine Learning.* CoRR, abs/2104.02466, 2021.

AI Act

- https://artificialintelligenceact.eu/
- Leucker: *The AI Act and Some Implications for Developing AI-Based Systems.* The Combined Power of Research, Education, and Dissemination 2025.