

Herausforderungen des AI Acts aus technischer Sicht

Prof. Dr.-Ing. Holger Hermanns

Universität des Saarlandes

Transregionaler Sonderforschungsbereich 248



Über mich



- Professor für Informatik an der Universität des Saarlandes
- Schwerpunkt der Forschung: Verlässlichkeit cyber-physischer Systeme

Über mich



- Professor für Informatik an der Universität des Saarlandes
- Schwerpunkt der Forschung: Verlässlichkeit cyber-physischer Systeme
- Leiter des Sonderforschungsbereiches TRR 248

Grundlagen verständlicher Software-Systeme

Eine Initiative von Informatikern des Saarland Informatics Campus und der Technischen Universität Dresden

Verständliche Software-Systeme – Warum?



Das explosionsartig wachsende Potential

Software-basierter Innovationen

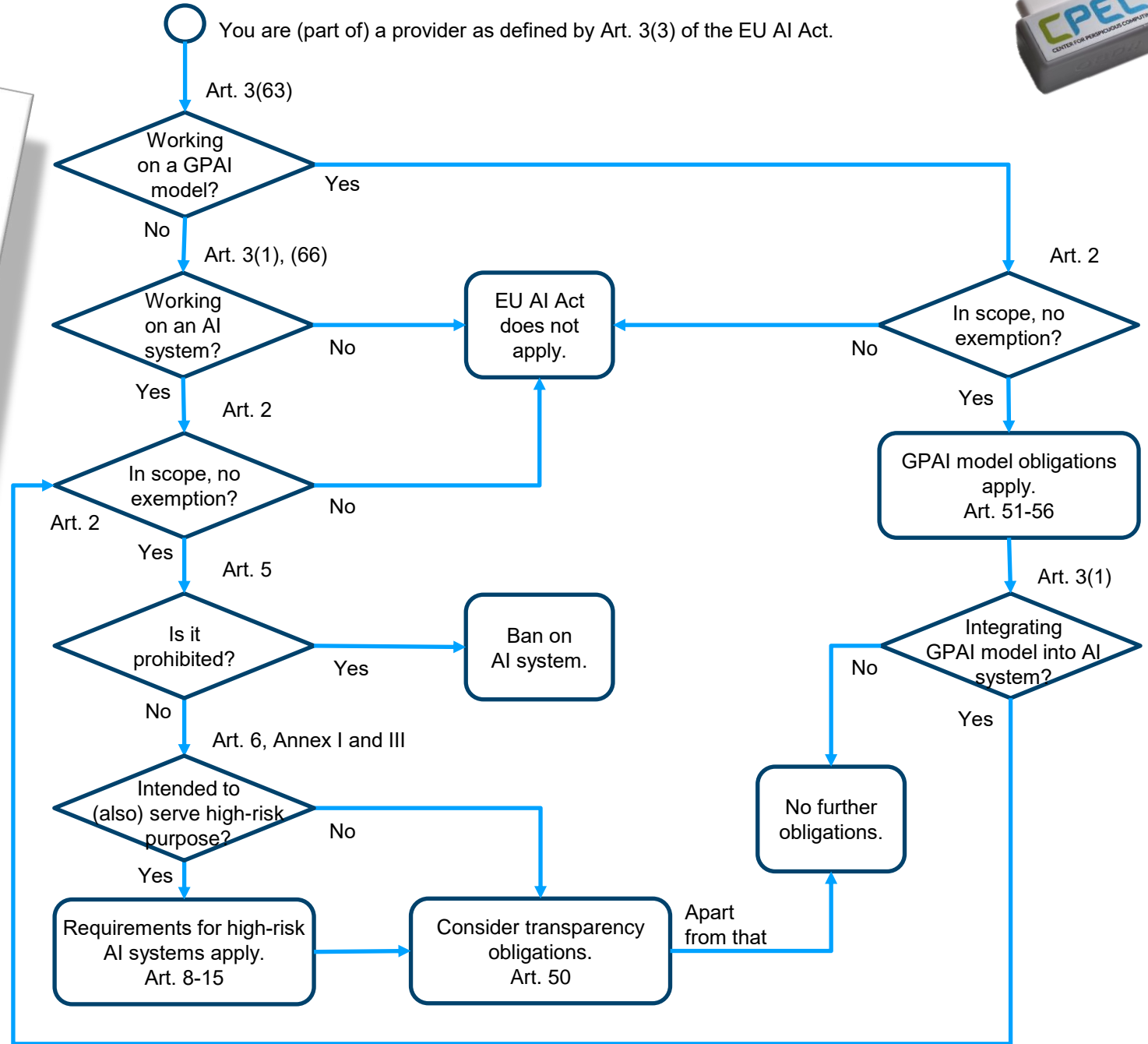
bedingt eine Implosion der Möglichkeiten und Fähigkeiten

diese zu verstehen und zu kontrollieren.





You are (part of) a provider as defined by Art. 3(3) of the EU AI Act.



AI Act for the Working Programmer*

Holger Hermanns¹, Anne Lauber-Rönsberg², Philip Meinel², Sarah Sterz¹, and Hanwei Zhang¹

¹ Saarland University, Saarland Informatics Campus, Saarbrücken, Germany
{hermanns, sterz, zhang}@depend.uni-saarland.de
² TU Dresden University of Technology, Institute of International Law, Intellectual Property and Technology Law, Dresden, Germany
{anne.lauber-roensberg, philip.meinel}@tu-dresden.de

Abstract. The European AI Act is a new, legally binding document that will enforce certain requirements on the development and use of AI technology potentially affecting people in Europe. It can be expected that the stipulations of the Act, in turn, are going to affect the work of many software engineers, software testers, data engineers, and other professionals across the IT sector in Europe and beyond. The 113 articles, 180 recitals, and 13 annexes that make up the Act cover more than 450 pages. This paper aims at providing an aid for navigating the Act from the perspective of some professional in the software domain, termed "the working programmer", who feels the need to know about the stipulations of the Act.

Introduction

Extensive deliberations, the European Union has taken the final step for adopting the AI Act [10]. The AI Act aims to ensure the development and deployment of trustworthy AI by relying on a risk-based approach – the higher the risks to society, the stricter the legal requirements.¹ However, the details of the regulated areas of AI often seem blurred. The idea of this paper is to provide the "working programmer"² with some initial help in navigating the complexities of the AI Act. In doing so, we make three main contributions:

1. We provide an overview of the regulated AI technologies and how to distinguish them. This is essential for the working programmer to determine which obligations under the AI Act might apply to their work.

2. We identify the relevant obligations to help the programmer understand which parts of the Act may be relevant. This is supported by a flowchart that helps to find the relevant obligations in simple questions and to narrow down the complexities of the Act.



KI-System

Begriffsbestimmungen

Für die Zwecke dieser Verordnung bezeichnet der Ausdruck

1. „KI-System“ ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können;

ein maschinengestütztes System,
das ... aus erhaltenen Eingaben ... ableitet,
wie Ausgaben ... erstellt werden, ...
die ... Umgebungen beeinflussen können.

KI-System



(12) Der Begriff „KI-System“ in dieser Verordnung sollte klar definiert und eng mit der Tätigkeit internationaler Organisationen abgestimmt werden, die sich mit KI befassen, um Rechtssicherheit, mehr internationale Konvergenz und hohe Akzeptanz sicherzustellen und gleichzeitig Flexibilität zu bieten, um den raschen technologischen Entwicklungen in diesem Bereich Rechnung zu tragen. Darüber hinaus sollte die Begriffsbestimmung auf den wesentlichen Merkmalen der KI beruhen, die sie von einfacheren herkömmlichen Softwaresystemen und Programmierungsansätzen abgrenzen, und sollte sich nicht auf Systeme beziehen, die auf ausschließlich von natürlichen Personen definierten Regeln für das automatische Ausführen von Operationen beruhen. Ein wesentliches Merkmal von KI-

... sollte sich nicht auf Systeme beziehen,
die auf ausschließlich von natürlichen Personen
definierten Regeln für das automatische
Ausführen von Operationen beruhen.

KI-System



das automatische Ausführen von Operationen beruhen. Ein wesentliches Merkmal von KI-Systemen ist ihre Fähigkeit, abzuleiten. Diese Fähigkeit bezieht sich auf den Prozess der Erzeugung von Ausgaben, wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen, die physische und digitale Umgebungen beeinflussen können, sowie auf die Fähigkeit von KI-Systemen, Modelle oder Algorithmen oder beides aus Eingaben oder Daten abzuleiten. Zu den Techniken, die während der Gestaltung eines KI-Systems das Ableiten ermöglichen, gehören Ansätze für maschinelles Lernen, wobei aus Daten gelernt wird, wie bestimmte Ziele erreicht werden können, sowie logik- und wissensgestützte Konzepte, wobei aus kodierten Informationen oder symbolischen Darstellungen der zu lösenden Aufgabe abgeleitet wird. Die Fähigkeit eines KI-Systems, abzuleiten, geht über die

Zu den Techniken, ... die das Ableiten ermöglichen, gehören ...

Ansätze für maschinelles Lernen, ...

Ableiten sowie logik- und wissensgestützte Konzepte, ...

KI System? Hochrisiko?

~~KI~~

- Eine (typischerweise hochkomplexe) Programmierumgebung, die zum Erstellen des Codes für ein Airbag-Steuergerät verwendet wird.

Hochrisiko

KI

- Ein rein logikbasiertes System, das ableiten kann, wie zu entscheiden ist, ob der Airbag in einem Fahrzeug ausgelöst werden muss.

*wahrscheinlich
technisch unmöglich
(sagt der Experte)*

Hochrisiko

~~KI~~

- Ein rein logikbasiertes System, das ableiten kann, ob der Airbag in einem bestimmten Fahrzeug ausgelöst werden muss.

Hochrisiko

KI

- Ein System, bei dem maschinelles Lernen aus den Merkmalen vergangener Unfälle verwendet wurde, um zu entscheiden, ob der Airbag in einem Fahrzeug ausgelöst werden muss.

Hochrisiko



Begriffsbestimmungen

Für die Zwecke dieser Verordnung bezeichnet der Ausdruck

1. „KI-System“ ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können;

ein maschinengestütztes System,

das ... aus erhaltenen Eingaben ... **ableitet,**

wie Ausgaben ... erstellt werden, ...

die ... Umgebungen beeinflussen können.

KI-System

Hochrisiko KI (for the working programmer)

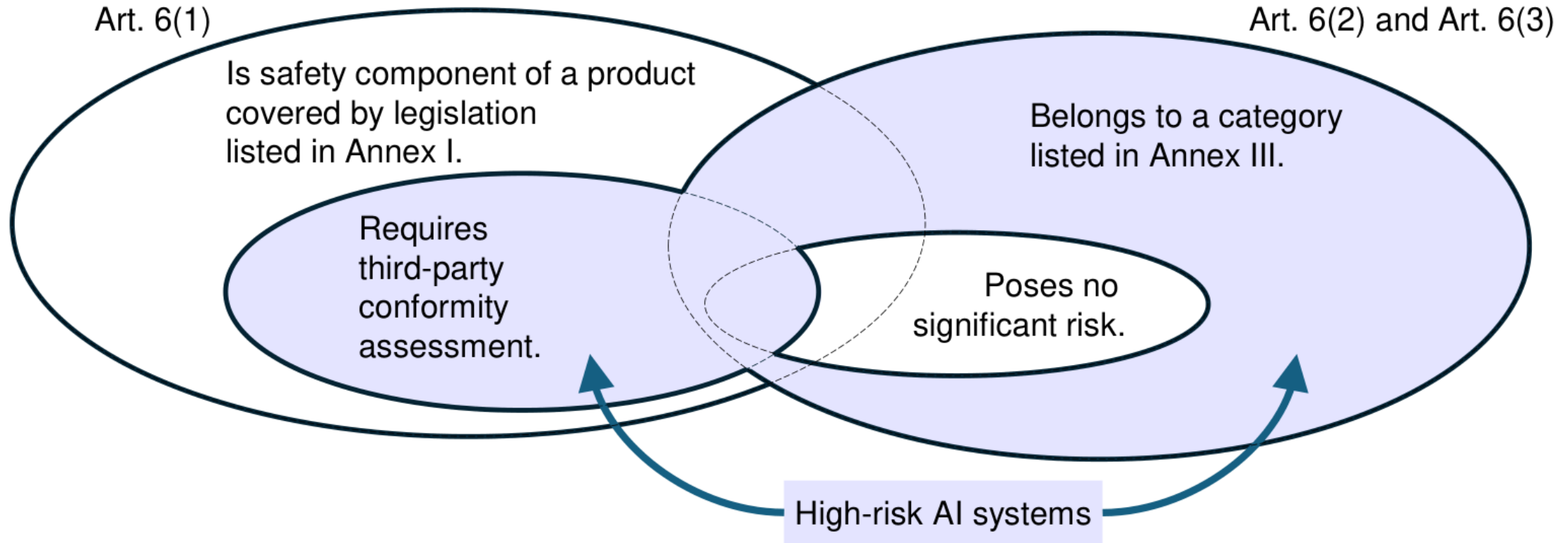


Fig. 2. Two avenues for an AI system to be classified as high-risk, represented as a Venn diagram.

Hochrisiko KI for the Working Programmer

Art 9: Risk management



Art 10: Data and data governance

Art 11: Technical documentation

Art 12: Record keeping

Art 13: Transparency and provision of information to users



Art 14: Human oversight

Art 15: Accuracy, robustness and cybersecurity

Menschliche Aufsicht: Artikel 14

For the purpose of implementing paragraphs 1, 2 and 3, the high-risk AI system shall be provided to the deployer in such a way that natural persons to whom human oversight is assigned are enabled, as appropriate and proportionate:

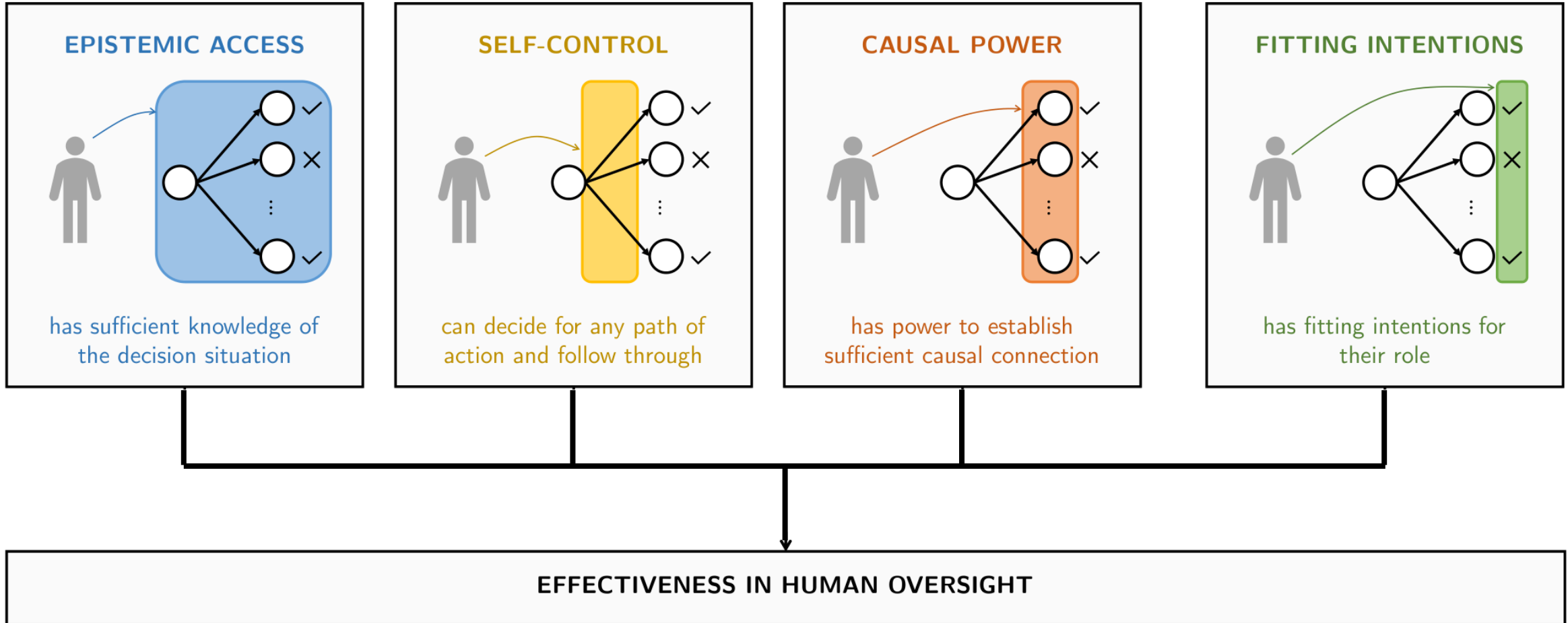
- (a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance;
- (b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
- (c) to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;
- (d) to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system;
- (e) to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state.

Menschliche Aufsicht: Artikel 14

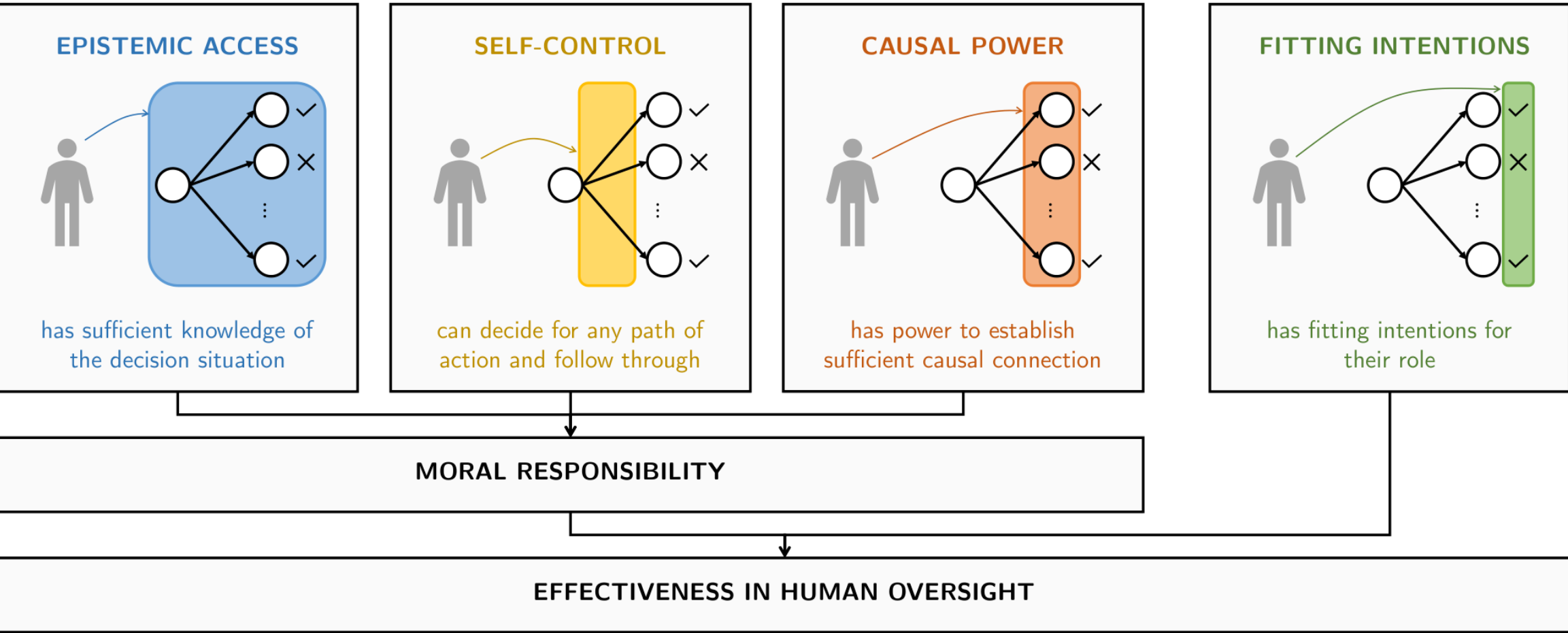
For the purpose of implementing paragraphs 1, 2 and 3, the high-risk AI system shall be provided to the deployer in such a way that natural persons to whom human oversight is assigned are enabled, as appropriate and proportionate:

- (a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance;
- (b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
- (c) to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;
- (d) to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system;
- (e) to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state.

Effektive menschliche Aufsicht



Effektive menschliche Aufsicht



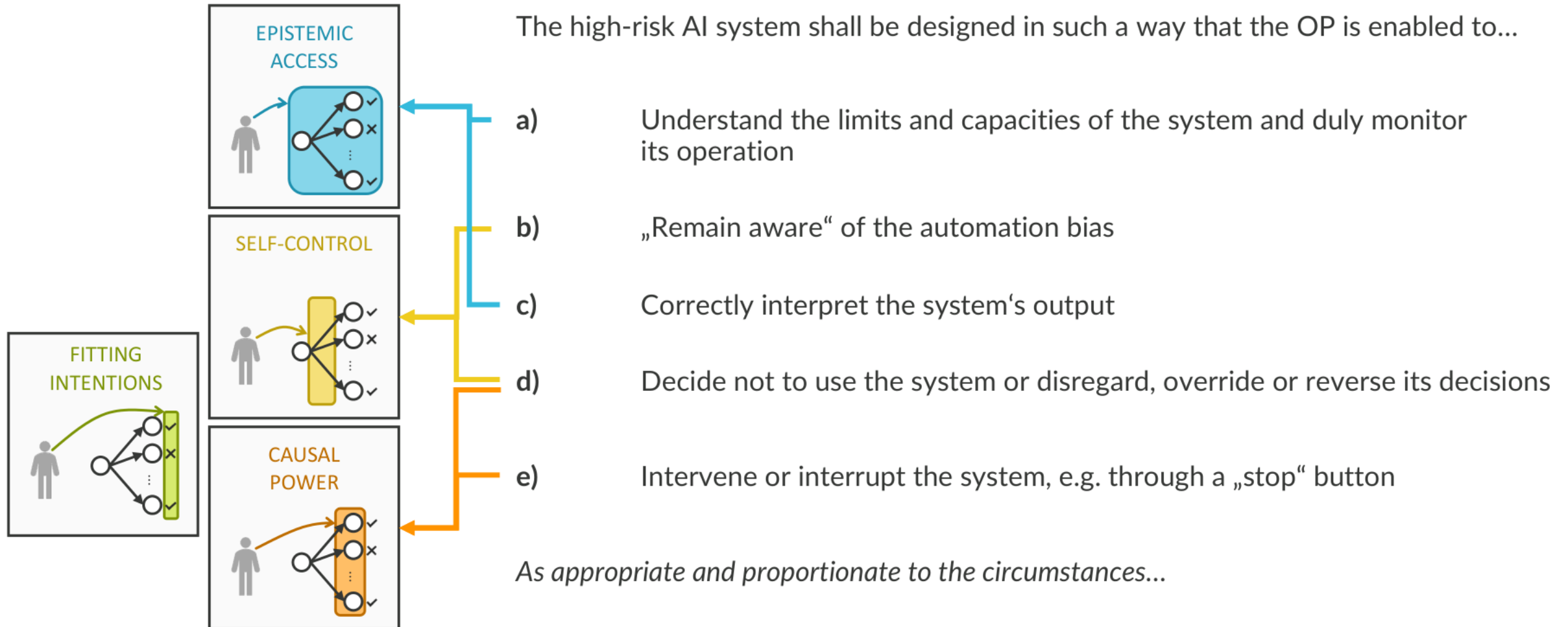
Menschliche Aufsicht: Artikel 14

The high-risk AI system shall be designed in such a way that the OP is enabled to...

- a) Understand the limits and capacities of the system and duly monitor its operation
- b) „Remain aware“ of the automation bias
- c) Correctly interpret the system's output
- d) Decide not to use the system or disregard, override or reverse its decisions
- e) Intervene or interrupt the system, e.g. through a „stop“ button

As appropriate and proportionate to the circumstances...

Menschliche Aufsicht: Artikel 14





Erleichternde und hemmende Faktoren für die Effektivität

	intervention options	system adaptability	system understandability	interpretability of in- and outputs	preselection of outputs to review	overseer training	domain expertise	conscientiousness	exhaustion	motivation	automation bias	adequate job design	role conflicts	independent thinking	accountability	time pressure
	technical design				individual factors						environment					
causal power	•	•				•	•									○
epistemic access		•	•	•	•	•	•	○	•	○	•		•	•		○
self-control						•	•	○	•	○	•			•		
fitting intentions						•	•	○	•	○	•	○		•/○		

Erleichternde und hemmende Faktoren für die Effektivität

	technical design						individual factors					environment				
	intervention options	system adaptability	system understandability	interpretability of in- and outputs	preselection of outputs to review	overseer training	domain expertise	conscientiousness	exhaustion	motivation	automation bias	adequate job design	role conflicts	independent thinking	accountability	time pressure
causal power	•	•				•	•	○	•	○	•					○
epistemic access		•	•	•	•	•	•	○	•	○	•		•	•		○
self-control						•	•	○	•	○	•			•		
fitting intentions						•	•	○	•	○	•	○		•/○		

Über Eingaben und Ausgaben



$P : \text{In} \rightarrow \text{Out}$

Was könnte das sein?

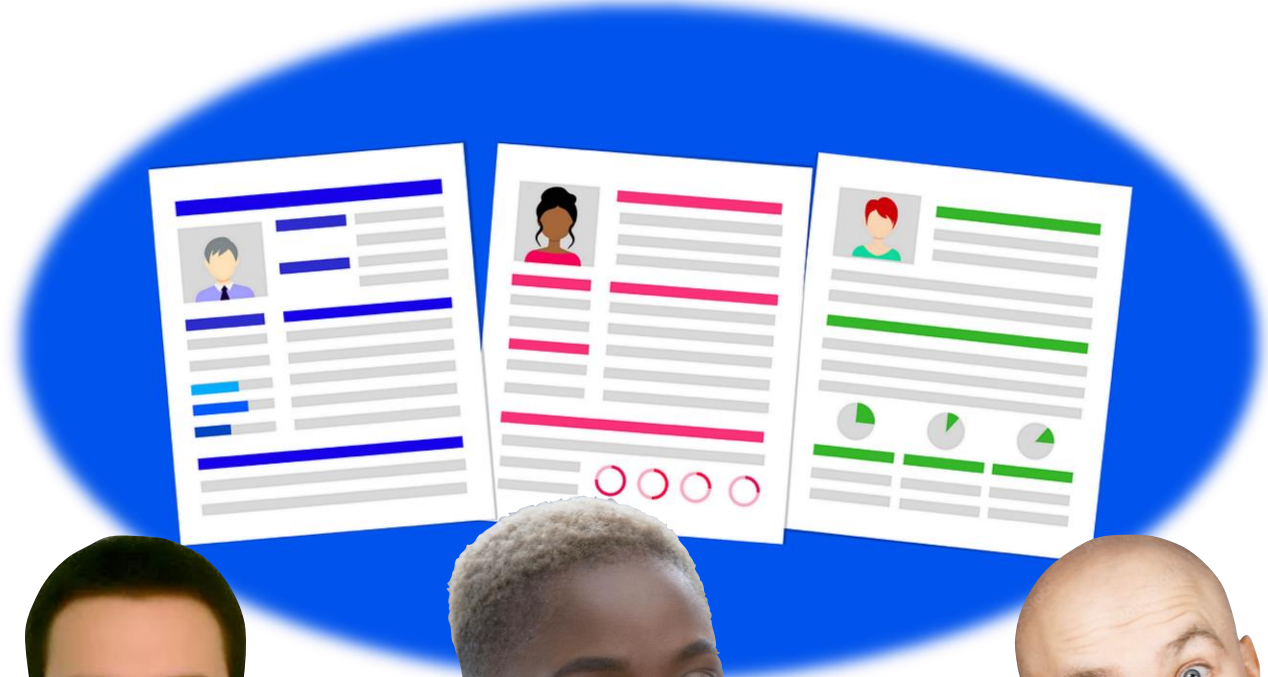
Über Eingaben und Ausgaben


$$P : \text{In} \rightarrow \text{Out}$$

Was könnte das sein?

- Eine mathematische Funktion, die Eingaben auf Ausgaben abbildet.
 - eventuell beschrieben durch einen Algorithmus,
 - eventuell beschrieben durch ein Programm,
 - eventuell beschrieben durch ein KI-Modell.
- Im letzteren Fall ist die Funktion für gewöhnlich gelernt.

Example – Individual Fairness



Eugene

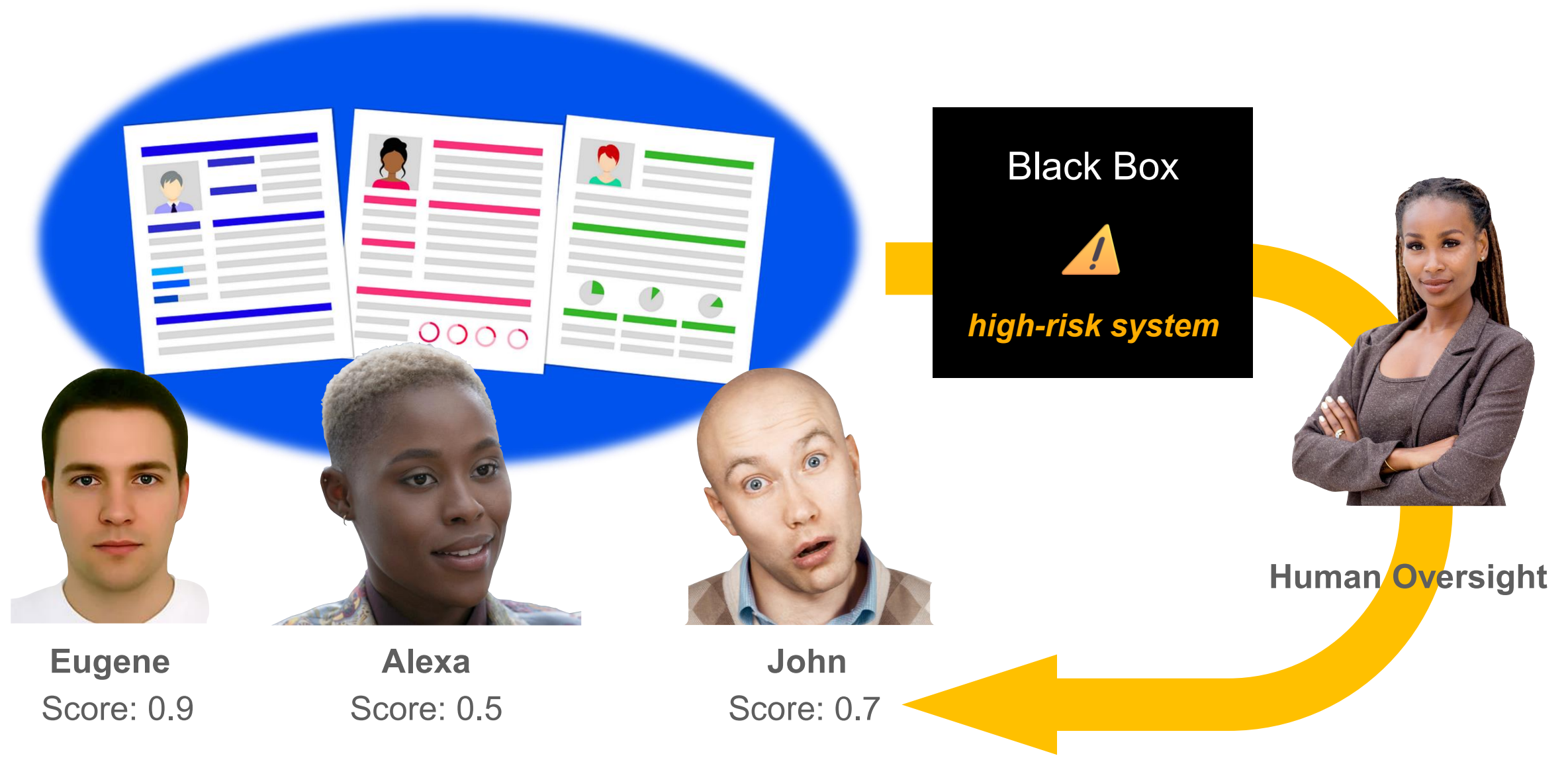


Alexa



John

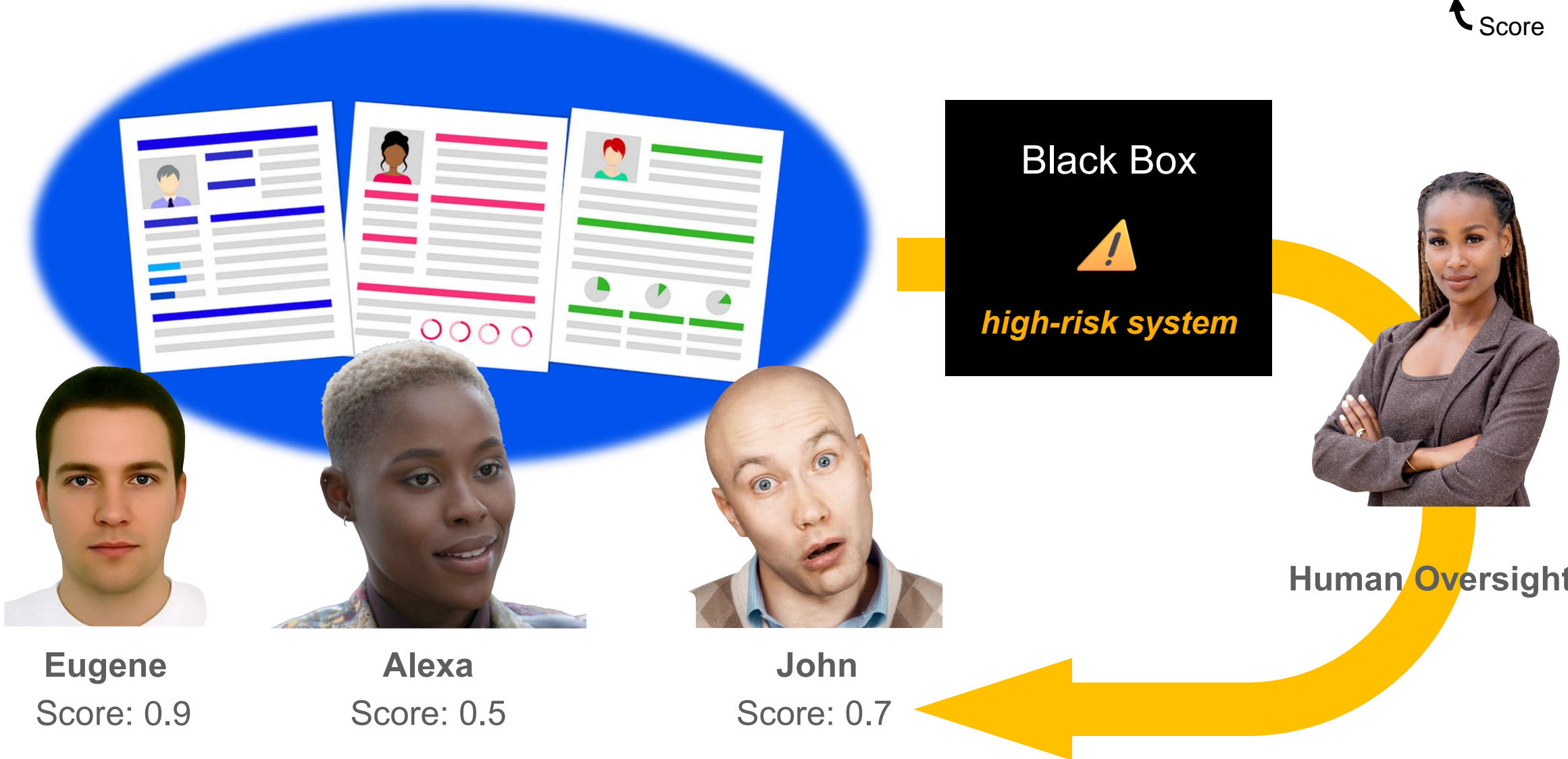
Example – Individual Fairness



Example – Individual Fairness

$$P : \text{In} \rightarrow \text{Out}$$

↙ Data about a human
↘ Score



Eugene
Score: 0.9



Alexa
Score: 0.5



John
Score: 0.7

Black Box

high-risk system



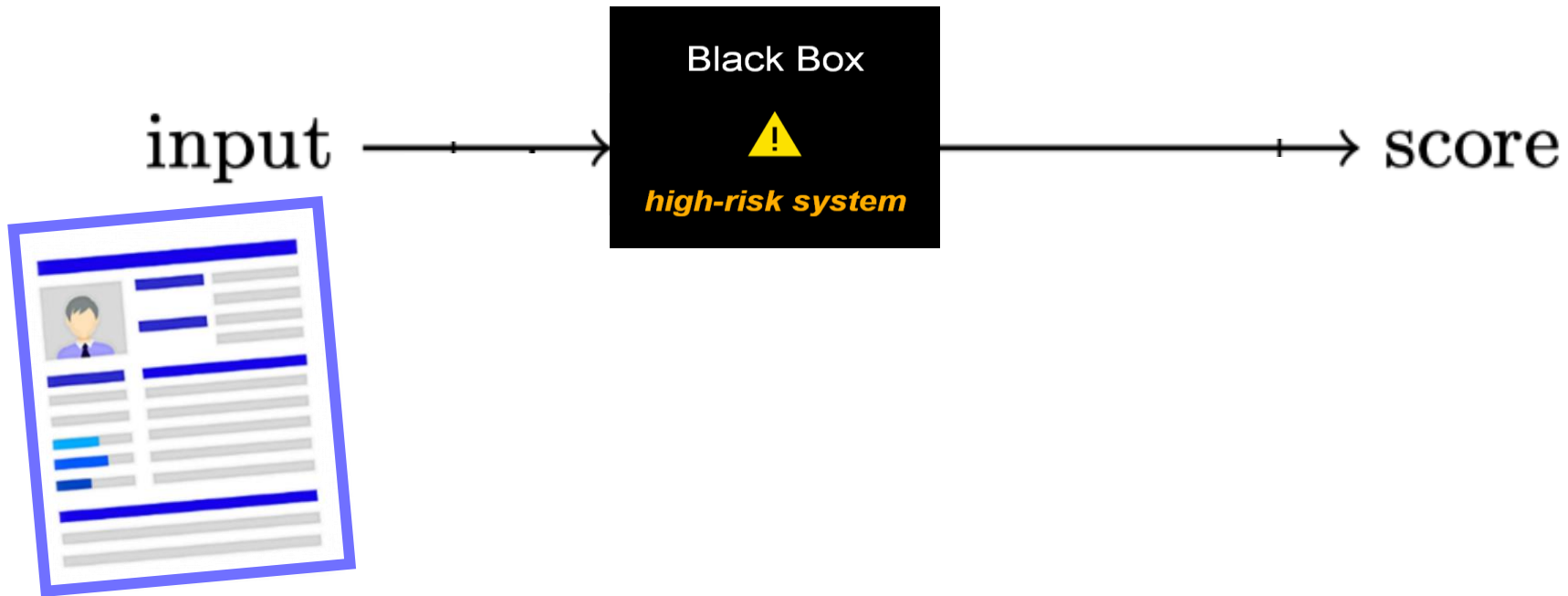
Human Oversight

Fairness Aware AI System

$$P : \text{In} \rightarrow \text{Out}$$

↙ Data about a human
↘ Score

For all $i_1 \in \mathcal{I}, i_2 \in \text{In}, d_{\text{Out}}(P(i_1), P(i_2)) \leq f(d_{\text{In}}(i, i'))$

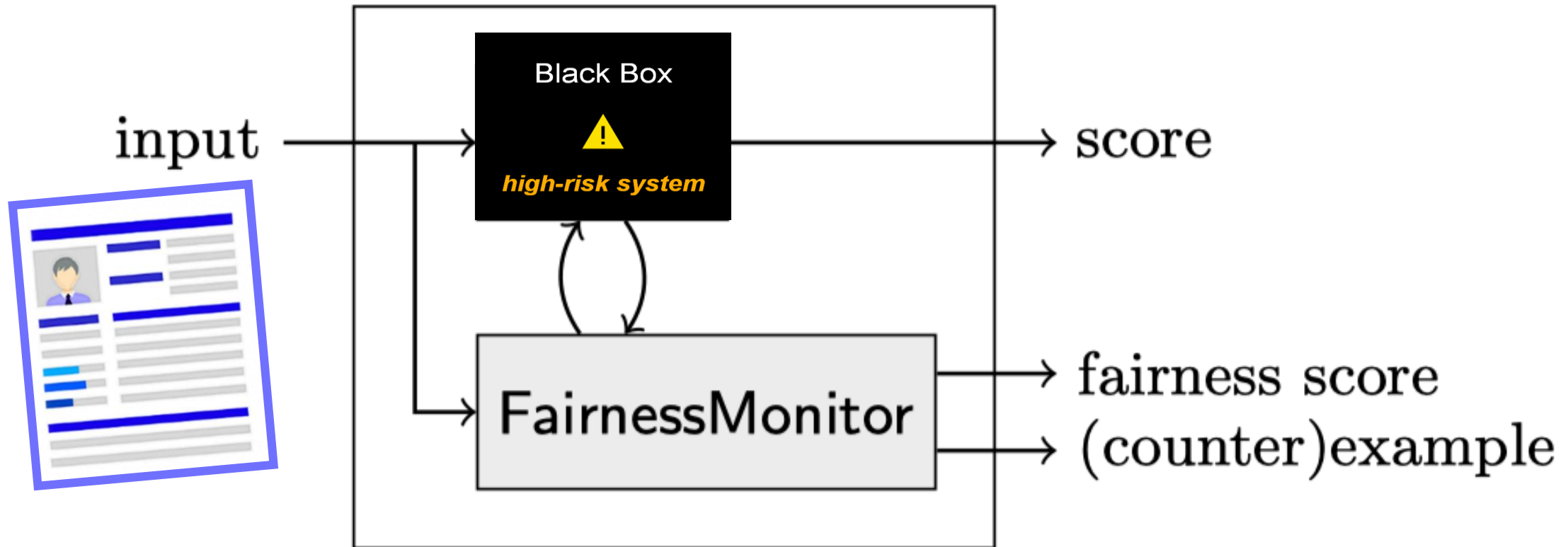


Fairness Aware AI System

$$P : \text{In} \rightarrow \text{Out}$$

↙ Data about a human
↘ Score

For all $i_1 \in \mathcal{I}, i_2 \in \text{In}, d_{\text{Out}}(P(i_1), P(i_2)) \leq f(d_{\text{In}}(i, i'))$



Fairness Monitoring

For all $i_1 \in \mathcal{I}, i_2 \in \text{In}, d_{\text{Out}}(P(i_1), P(i_2)) \leq f(d_{\text{In}}(i, i'))$

Algorithm 2.1 Monte-Carlo falsification

Input: w : Initial trace, \mathcal{R} : Robustness function, PS: Proposal Scheme

Output: $w \in M$

```

1: while  $\mathcal{R}(w) > 0$  do
2:    $w' \leftarrow \text{PS}(w)$ 
3:    $\alpha \leftarrow \exp(-\beta(\mathcal{R}(w') - \mathcal{R}(w)))$ 
4:    $r \leftarrow \text{UniformRandomReal}(0, 1)$ 
5:   if  $r \leq \alpha$  then
6:      $w \leftarrow w'$ 
7:   end if
8: end while

```

Fairness score – Robustness estimate

$$F(i_a, i_s) := f(d_{\text{In}}(i_a, i_s)) - d_{\text{Out}}(P(i_a), P(i_s))$$

$$F(\mathcal{I}, i_s) := \min\{F(i_a, i_s) \mid i_a \in \mathcal{I}\}$$

$$\mathcal{R}_{\mathcal{I}}(i_s) := F(\mathcal{I}, i_s)$$

Fairness Monitoring

Algorithm 2 FairnessMonitor,

with ξ -min $S = (\xi, i_1, i_2)$ only if $(\xi, i_1, i_2) \in S$ and for all $(\xi', i'_1, i'_2) \in S, \xi' \geq \xi$

Falsification Parameters: PS: Proposal scheme, β : Temperature parameter

Input: System $P : \text{In} \rightarrow \text{Out}$, Fairness contract $\mathcal{F} = \langle d_{\text{In}}, d_{\text{Out}}, f \rangle$, and set of actual inputs \mathcal{I}

Output: A minimal fairness score triple from $\mathbb{R} \times \mathcal{I} \times \text{In}$.

- 1: $i_s \leftarrow$ any input $i_a \in \mathcal{I}$
- 2: $(\xi, i_{\min}, i_s) \leftarrow \xi\text{-min}\{F(i_a, i_s), i_a, i_s \mid i_a \in \mathcal{I}\}$
- 3: $(\xi_{\min}, i_1, i_2) \leftarrow (\xi, i_{\min}, i_s)$
- 4: **while not** timeout **do**
- 5: $i'_s \leftarrow \text{PS}(i_s, P(i_s))$
- 6: $(\xi', i'_{\min}, i'_s) \leftarrow \xi\text{-min}\{F(i_a, i'_s), i_a, i'_s \mid i_a \in \mathcal{I}\}$
- 7: $(\xi_{\min}, i_1, i_2) \leftarrow \xi\text{-min}\{(\xi_{\min}, i_1, i_2), (\xi', i'_{\min}, i'_s)\}$
- 8: $\alpha \leftarrow \exp(-\beta(\xi' - \xi))$
- 9: $r \leftarrow \text{UniformRandomReal}(0, 1)$
- 10: **if** $r \leq \alpha$ **then**
- 11: $i_s \leftarrow i'_s$
- 12: $\xi \leftarrow \xi'$
- 13: **end if**
- 14: **end while**
- 15: **return** (ξ_{\min}, i_1, i_2)

Algorithm 2.1 Monte-Carlo falsification

Input: w : Initial trace, \mathcal{R} : Robustness function, PS: Proposal Scheme

Output: $w \in M$

- 1: **while** $\mathcal{R}(w) > 0$ **do**
- 2: $w' \leftarrow \text{PS}(w)$
- 3: $\alpha \leftarrow \exp(-\beta(\mathcal{R}(w') - \mathcal{R}(w)))$
- 4: $r \leftarrow \text{UniformRandomReal}(0, 1)$
- 5: **if** $r \leq \alpha$ **then**
- 6: $w \leftarrow w'$
- 7: **end if**
- 8: **end while**

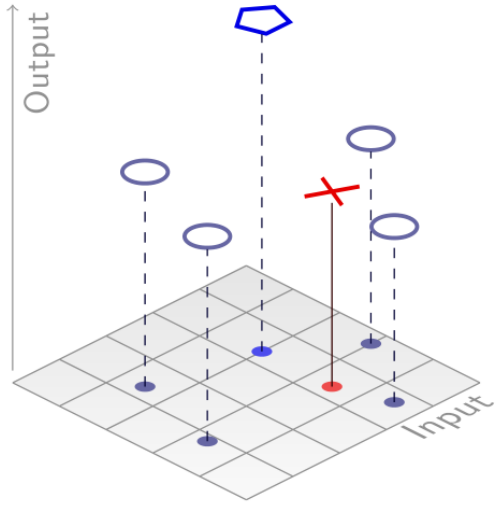
Fairness score – Robustness estimate

$$F(i_a, i_s) := f(d_{\text{In}}(i_a, i_s)) - d_{\text{Out}}(P(i_a), P(i_s))$$

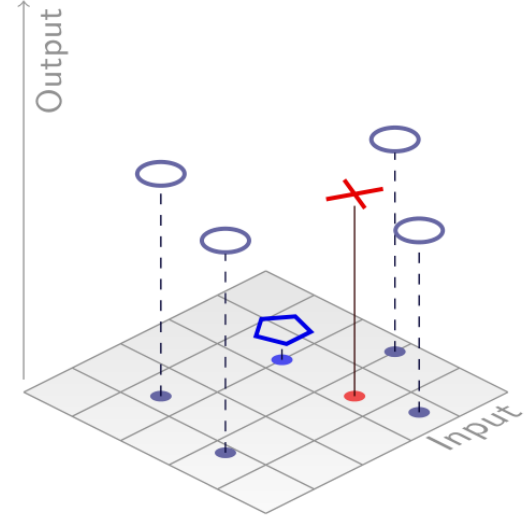
$$F(\mathcal{I}, i_s) := \min\{F(i_a, i_s) \mid i_a \in \mathcal{I}\}$$

$$\mathcal{R}_{\mathcal{I}}(i_s) := F(\mathcal{I}, i_s)$$

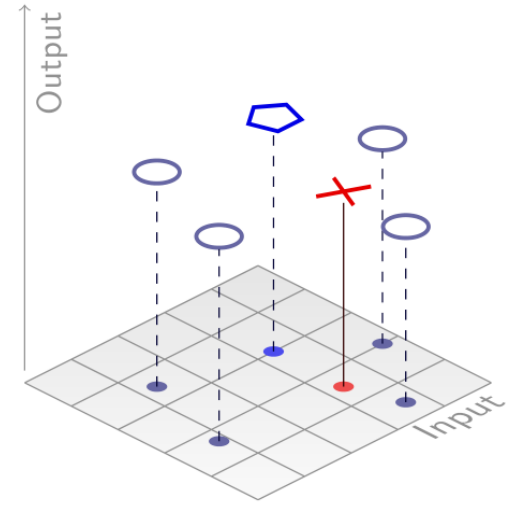
Cases of Unfairness



Individual scores worse than synthetic counterpart.



Individual scores better than synthetic counterpart.



No unfairness detected.

In Practice



Eugene

Score: 0.9



Alexa

Rainbow University

Score: 0.5



John

Poor Grades

Trump University

Score: 0.7



The score should be greater than 0.5...



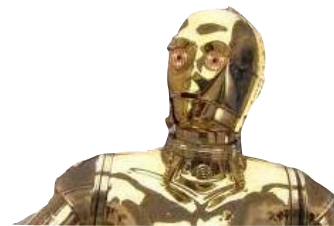
Very similar to Eugene

Score: 0.75



Snow University

Score: 0.6

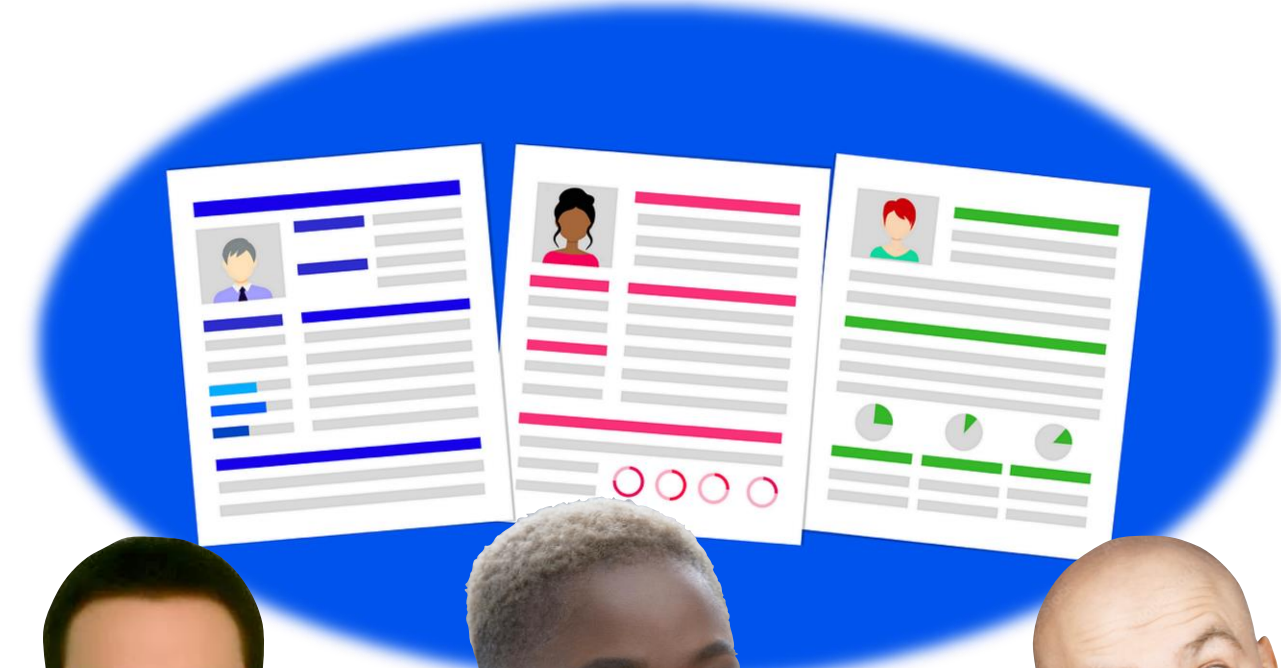


Same Poor Grades


Saarland University

Score: 0.4

Und jetzt mit GPAI!



GPAI-in-the-Box



high-risk system



Eugene
Score: 0.9



Alexa
Score: 0.5



John
Score: 0.7



Human Oversight





Daten und Daten-Governance

- (1) Hochrisiko-KI-Systeme, in denen Techniken eingesetzt werden, bei denen KI-Modelle mit Daten trainiert werden, müssen mit Trainings-, Validierungs- und Testdatensätzen entwickelt werden, die den in den Absätzen 2 bis 5 genannten Qualitätskriterien entsprechen, wenn solche Datensätze verwendet werden.
- (2) Für Trainings-, Validierungs- und Testdatensätze gelten Daten-Governance- und Datenverwaltungsverfahren, die für die Zweckbestimmung des Hochrisiko-KI-Systems geeignet sind. Diese Verfahren betreffen insbesondere
- (3) Die Trainings-, Validierungs- und Testdatensätze müssen im Hinblick auf die Zweckbestimmung relevant, hinreichend repräsentativ und so weit wie möglich fehlerfrei und vollständig sein. Sie müssen die geeigneten statistischen Merkmale, gegebenenfalls auch bezüglich der Personen oder Personengruppen, für die das Hochrisiko-KI-System bestimmungsgemäß verwendet werden soll, haben. Diese Merkmale der Datensätze können auf der Ebene einzelner Datensätze oder auf der Ebene einer Kombination davon erfüllt werden.

Was fordert Artikel 10?

- Daten-Governance- und Datenverwaltungsverfahren, die für die Zweckbestimmung des Hochrisiko-KI-Systems geeignet sind.

Benötigt werden unter anderen

- eine Bewertung der Verfügbarkeit, Menge und Eignung der benötigten Datensätze,
- eine Untersuchung bzgl möglicher Verzerrungen (Bias),
- Die Datensätze müssen im Hinblick auf die Zweckbestimmung relevant sein, sowie hinreichend repräsentativ und so weit wie möglich fehlerfrei und vollständig.
- Sie müssen die geeigneten statistischen Merkmale haben, gegebenenfalls auch bezüglich der Personen oder Personengruppen, für die das Hochrisiko-KI-System bestimmungsgemäß verwendet werden soll.

Dies braucht sehr tiefe Einblicke in die Daten.

Was liefern Artikel 51-56 sowie 25?

Annex XII enthält, was der GPAI-Anbieter bereitstellen muss, um Integration zu ermöglichen. Im wesentlichen:

- eine allgemeine Beschreibung des KI-Modells einschließlich der Information, worin es integriert werden kann;
- die anwendbaren Regelungen der akzeptablen Nutzung;
- gegebenenfalls wie das Modell mit Hardware oder Software interagiert, oder dafür verwendet werden kann;

Artikel 53 I (b) verlangt eine Dokumentation, die Anbieter von KI-Systemen in die Lage versetzen (soll), die Fähigkeiten und Grenzen des GPAI-Modells gut zu verstehen und ihren Pflichten gemäß dieser Verordnung nachzukommen.

Artikel 25 (4) verlangt eine vertragliche Vereinbarung zwischen dem Anbieter des GPAI-Modells und dem Anbieter der Hochrisiko-KI. Quelloffene GPAI-Modelle sind ausgenommen.

Was bedeutet das?



Die Informationspflichten haben keine Präzedenz über die Geschäftsgeheimnisse des Anbieters.

Die meisten Anbieter von GPAI-Modellen werden die bereitgestellten Informationen auf ein Minimum reduzieren, und damit die Überprüfung durch nachgelagerte Anbieter erschweren.

Es ist daher zu bezweifeln, dass die Informationsanforderungen für GPAI-Modellanbieter ausreichend sind, um die Pflichten nachgelagerter Anbieter erfüllbar zu machen.

Dies gilt insbesondere für die Verpflichtungen in Bezug auf die Trainingsdaten, die von zentraler Bedeutung zu sein scheinen.



Können da nicht die KI-Experten helfen?

Fundamental Nein.

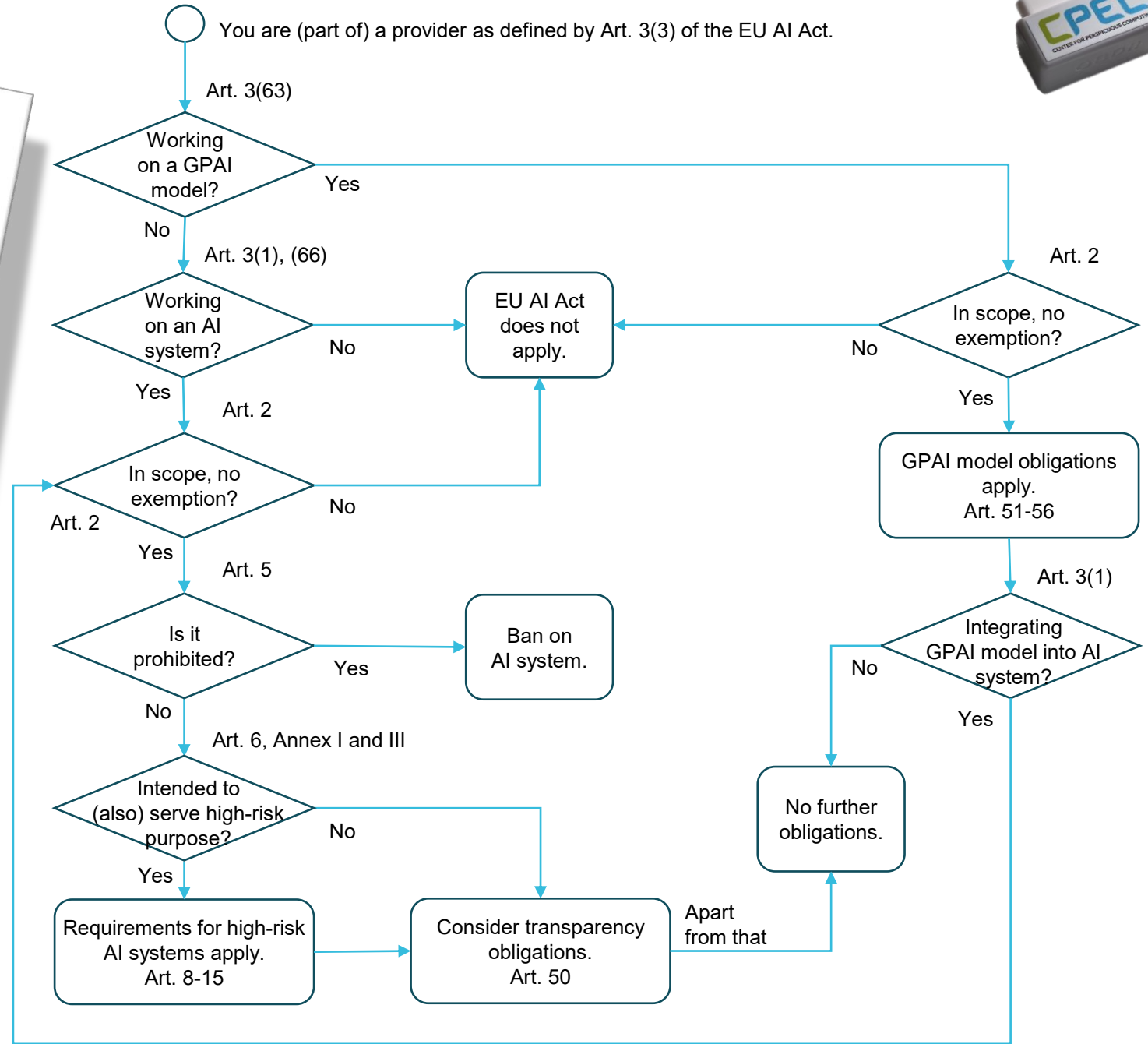
Es gibt viele Verfahren, um ein bestehendes GPAI-Modell

- zu verbessern,
- zu tunen,
- zu heilen.

Aber diese können die benötigten Garantien bezüglich der Trainingsdaten nicht heilen.



You are (part of) a provider as defined by Art. 3(3) of the EU AI Act.



AI Act for the Working Programmer*

Holger Hermanns¹, Anne Lauber-Rönsberg², Philip Meinel², Sarah Sterz¹, and Hanwei Zhang¹

¹ Saarland University, Saarland Informatics Campus, Saarbrücken, Germany
{hermanns, sterz, zhang}@depend.uni-saarland.de
² TU Dresden University of Technology, Institute of International Law, Intellectual Property and Technology Law, Dresden, Germany
{anne.lauber-roensberg, philip.meinel}@tu-dresden.de

Abstract. The European AI Act is a new, legally binding document that will enforce certain requirements on the development and use of AI technology potentially affecting people in Europe. It can be expected that the stipulations of the Act, in turn, are going to affect the work of many software engineers, software testers, data engineers, and other professionals across the IT sector in Europe and beyond. The 113 articles, 180 recitals, and 13 annexes that make up the Act cover more than 450 pages. This paper aims at providing an aid for navigating the Act from the perspective of some professional in the software domain, termed "the working programmer", who feels the need to know about the stipulations of the Act.

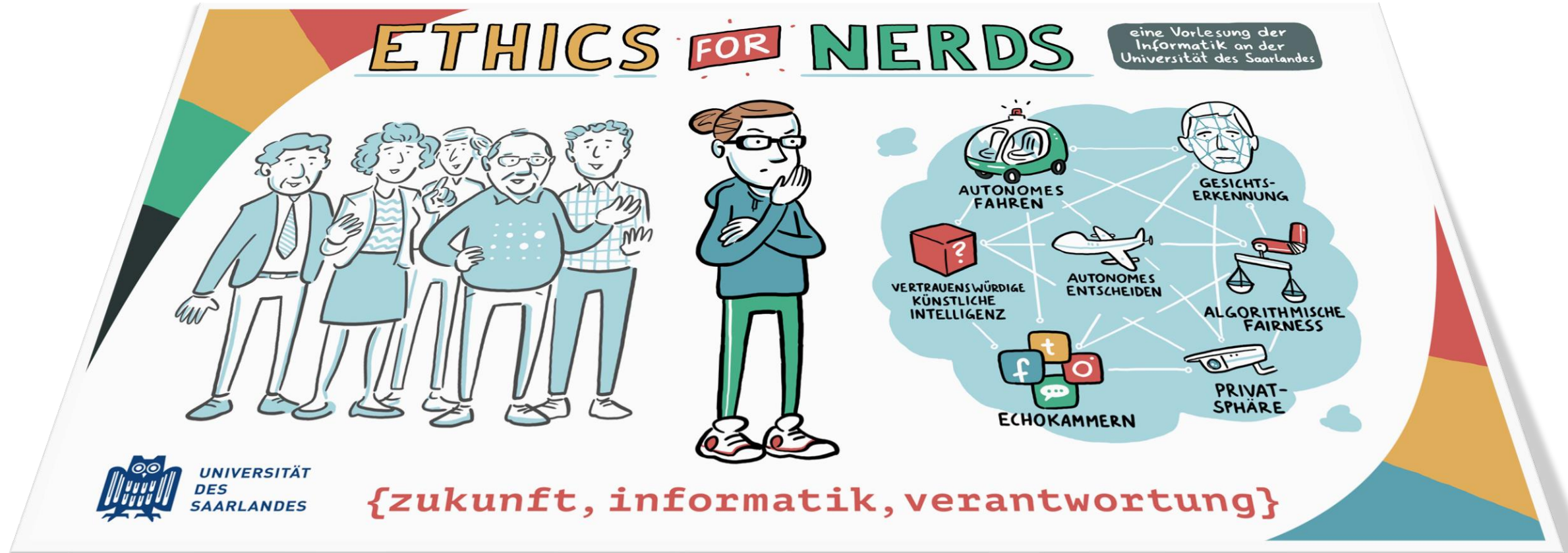
Introduction

Extensive deliberations, the European Union has taken the final step for adopting the AI Act [10]. The AI Act aims to ensure the development and deployment of trustworthy AI by relying on a risk-based approach – the higher the risks to society, the stricter the legal requirements.¹ However, the details of the regulated areas of AI often seem blurred. The idea of this paper is to provide the "working programmer"² with some initial help in navigating the complexities of the AI Act. In doing so, we make three main contributions:

1. We provide an overview of the regulated AI technologies and how to distinguish them. This is essential for the working programmer to determine which obligations under the AI Act might apply to their work.

2. We identify the relevant obligations to help the programmer understand which parts of the Act may be relevant. This is supported by a flowchart that helps to find the answers to simple questions and to narrow down the complexities of the Act.

Hochschulperle 2019



Eine Initiative von

CPEC CENTER FOR
PERSPICUOUS
COMPUTING

und

ES Explainable
Intelligent
Systems