Human Oversight:
Der Faktor Mensch in der KI-Sicherheit

Eine Strategie zur Risikominderung oder ein neuer Angriffsvektor?

Dr. Matthias Heck

Federal Office
for Information Security

Warum Human Oversight?

Warum Human Oversight?

Artikel 14 KI VO: Human Oversight:

*„Hochrisiko-KI-Systeme werden so konzipiert und entwickelt, dass sie **während der Dauer ihrer Verwendung** […] **von natürlichen Personen wirksam beaufsichtigt** werden können.“*

*„Die menschliche Aufsicht dient der **Verhinderung oder Minimierung der Risiken für Gesundheit, Sicherheit oder Grundrechte**, die entstehen können, wenn ein Hochrisiko-KI-System […] verwendet wird […]“*

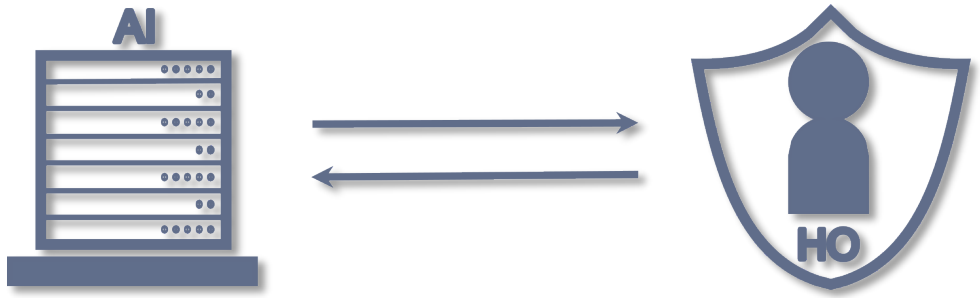
→ Akzeptanz von KI-Systemen

- Nutzende fordern menschliche Aufsicht im Sinne von Vertrauenswürdigkeit

→ Zunehmende Automatisierung von KI-Systemen

- Von Assistenzsystemen hin zu autonomen Systemen

In der operationalisierten Human Oversight werden KI-Systeme überwacht und Interventionen ermöglicht



Monitoring

- Überwachung relevanter Parameter
- Analyse der Outputs des HO IT-Systems

Interventionen

- Einzelentscheidungen des Systems manuell ändern
- System stoppen
- Modell ggf. anpassen/nachjustieren

Ziel: Risikominderung von Systemfehlern

Die vier Anforderungen an Human Oversight

Effektive Human Oversight

Epistemic Access

Der Akteur verfügt über ausreichendes Wissen über seine Entscheidungssituation.

Causal Power

Der Akteur hat die Fähigkeit, eine hinreichende kausale Verbindung zum relevanten Aspekt der Welt herzustellen.

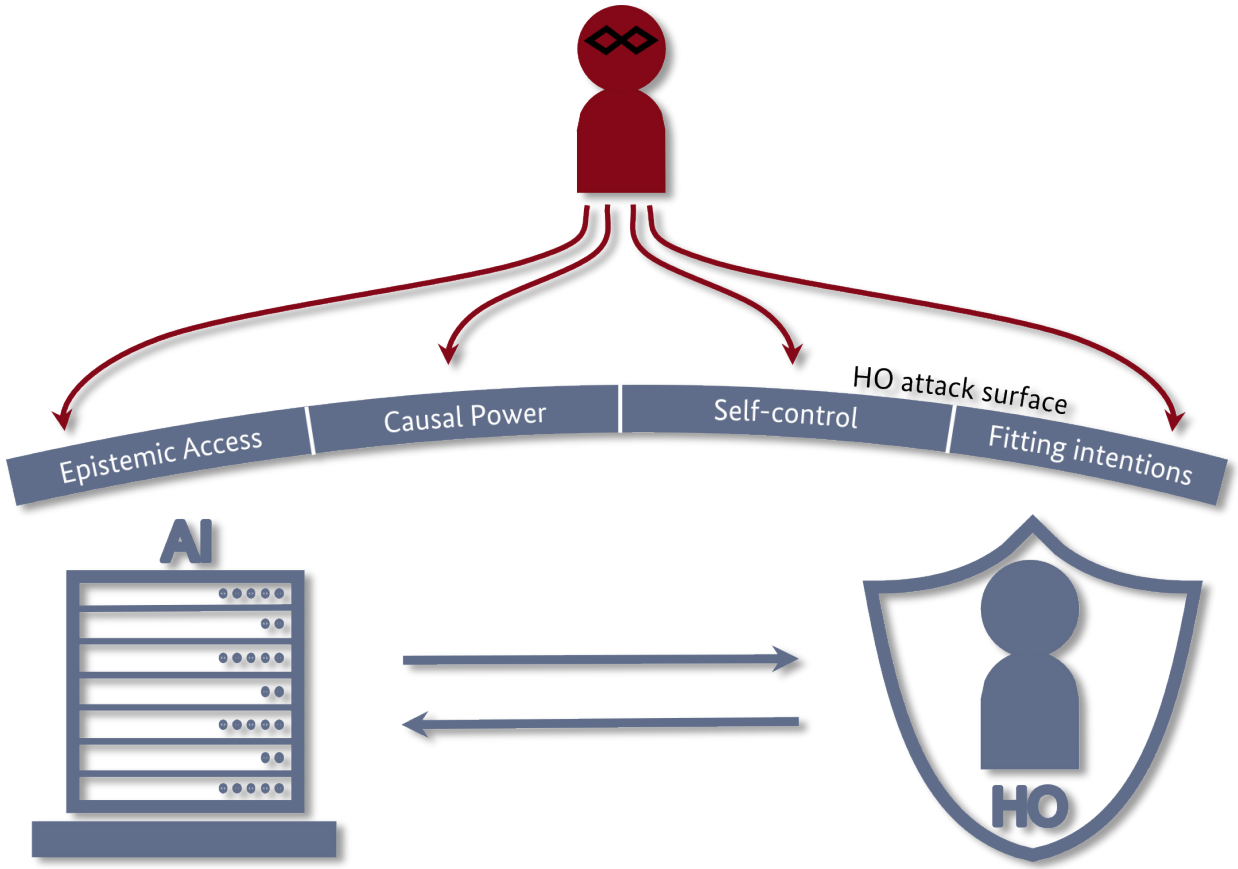
Self-control

Der Akteur kann sich in seiner Entscheidungssituation für jeden Handlungsweg entscheiden und diesen, wenn er dies tut, auch umsetzen

Fitting intentions

Der Akteur hat Absichten, die seiner Rolle (oder einem anderen relevanten Maßstab) angemessen sind.

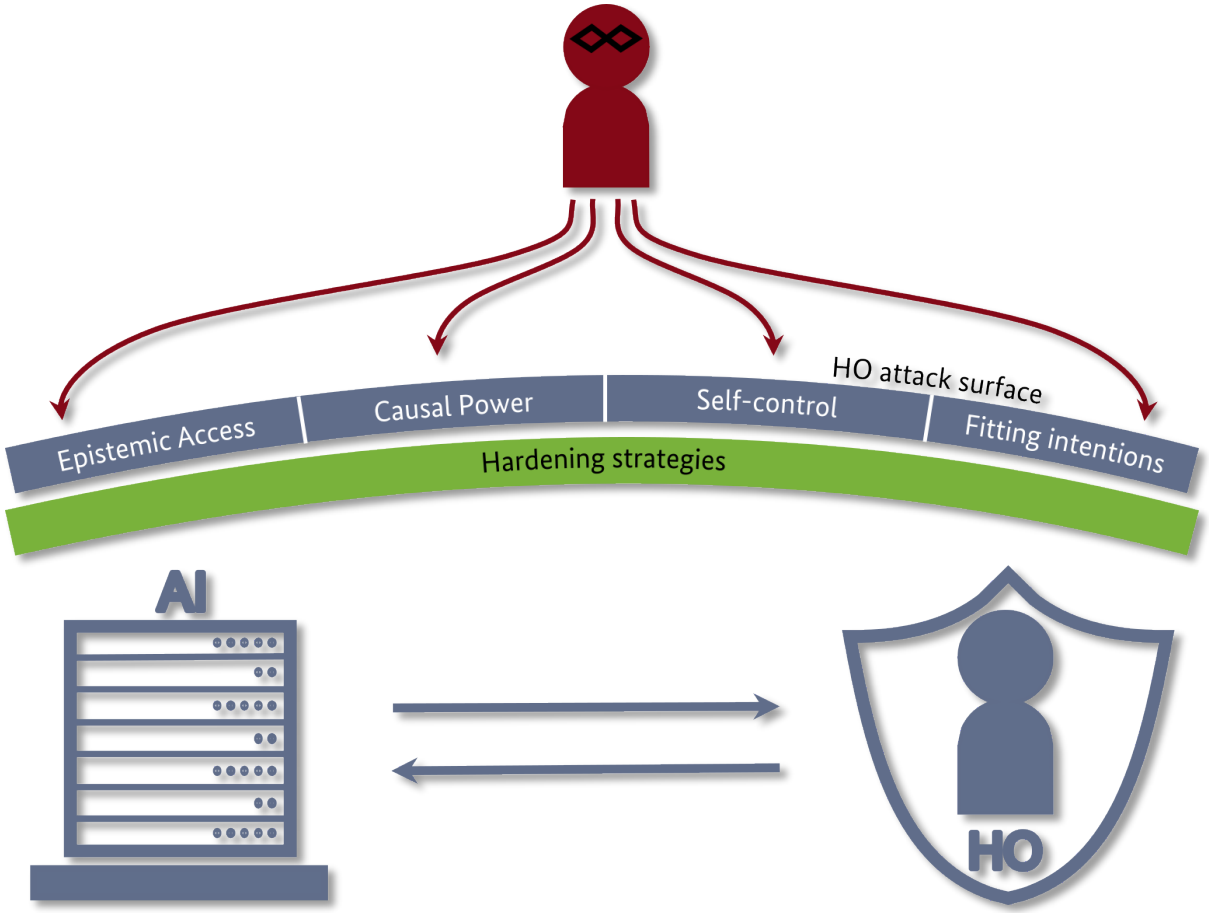
Human Oversight als Teil der Sicherheits-, Schutz- und Verantwortungsarchitektur von KI ermöglicht neue Angriffsvektoren



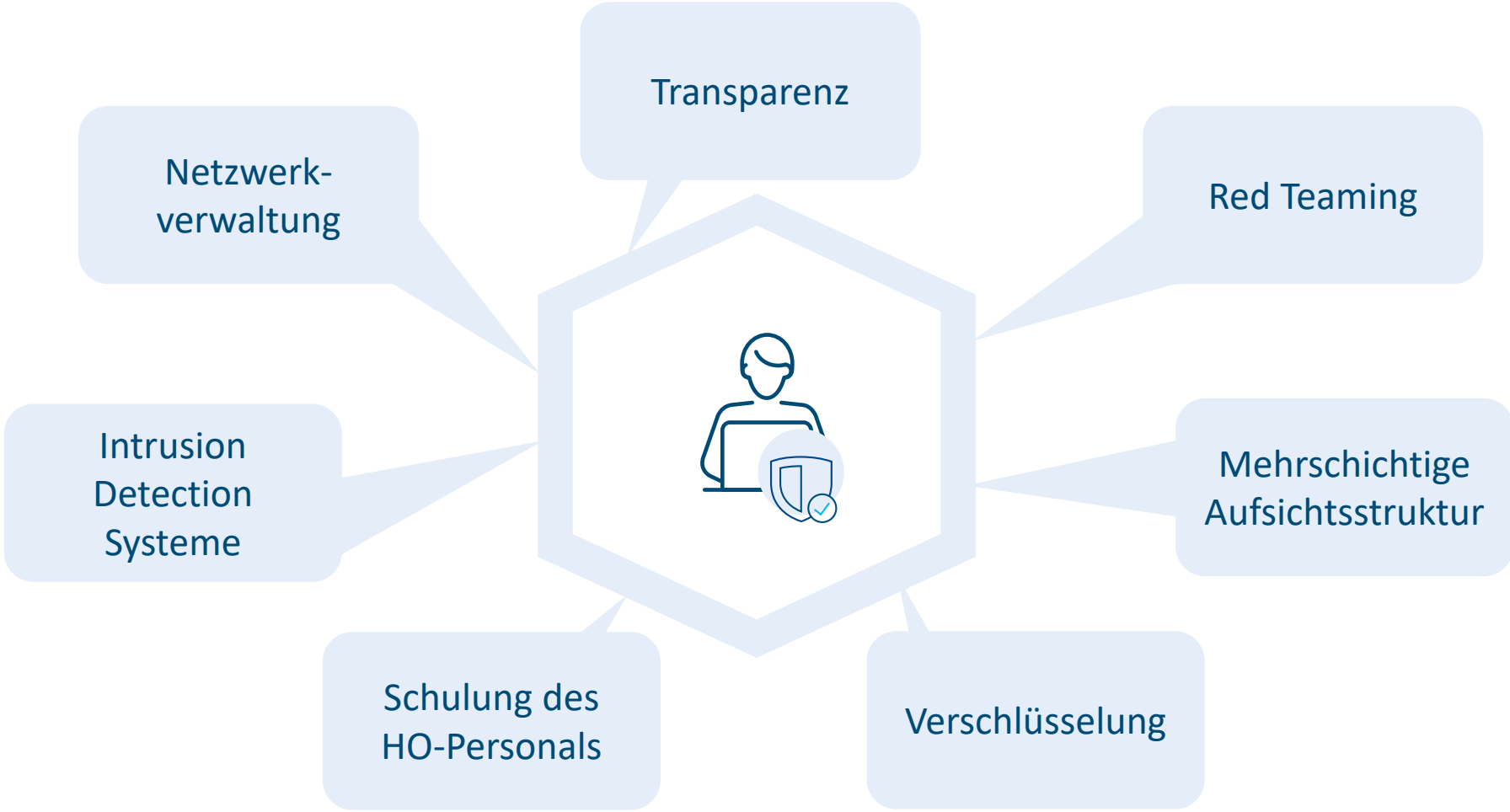
Analyse der Bedrohungslage von Human Oversight nach STRIDE

S	Spoofing	Angriffe, die versuchen Anmeldeinformationen unter Vortäuschung einer falschen Identität auszuleiten.
T	Tampering	Angriffe, die versuchen Informationen oder Systemfunktionen böswillig zu verändern.
R	Repudiation	Angriffe, die versuchen nicht erlaubte Operationen auf einem System auszuführen.
I	Information disclosure	Angriffe, die versuchen unbefugten Zugriff auf nicht-öffentliche Informationen zu erlangen.
D	Denial of Service	Angriffe, die versuchen Systeme durch Überlastung unzugänglich machen.
E	Elevation of privilege	Angriffe, die versuchen eine Verbesserung der Privilegien in System zu erreichen.

Gegenmaßnahmen und Härtungsstrategien erfordern security by design in der Umsetzung



Human Oversight: Eine Strategie zur Risikominderung, wenn sie effektiv und sicher implementiert wird!



Vielen Dank für Ihre Aufmerksamkeit

Dr. Matthias Heck
Referatsleiter

matthias.heck@bsi.bund.de

Bundesamt für Sicherheit in der Informationstechnik (BSI)
Godesberger Allee 87
53175 Bonn
www.bsi.bund.de

Illustrationen von Storyset.com

Bundesamt
für Sicherheit in der
Informationstechnik