EMPIANA

UX für AI-fähige Medizinprodukte - Hallucinating the Future

von Michael Engler

Disclaimer

Der Vortrag stellt einen Snap-Shot des mir bekannten Wissens heute dar.

Oder wie die FDA sagt:

Current Thinking

Was ist das Problem?

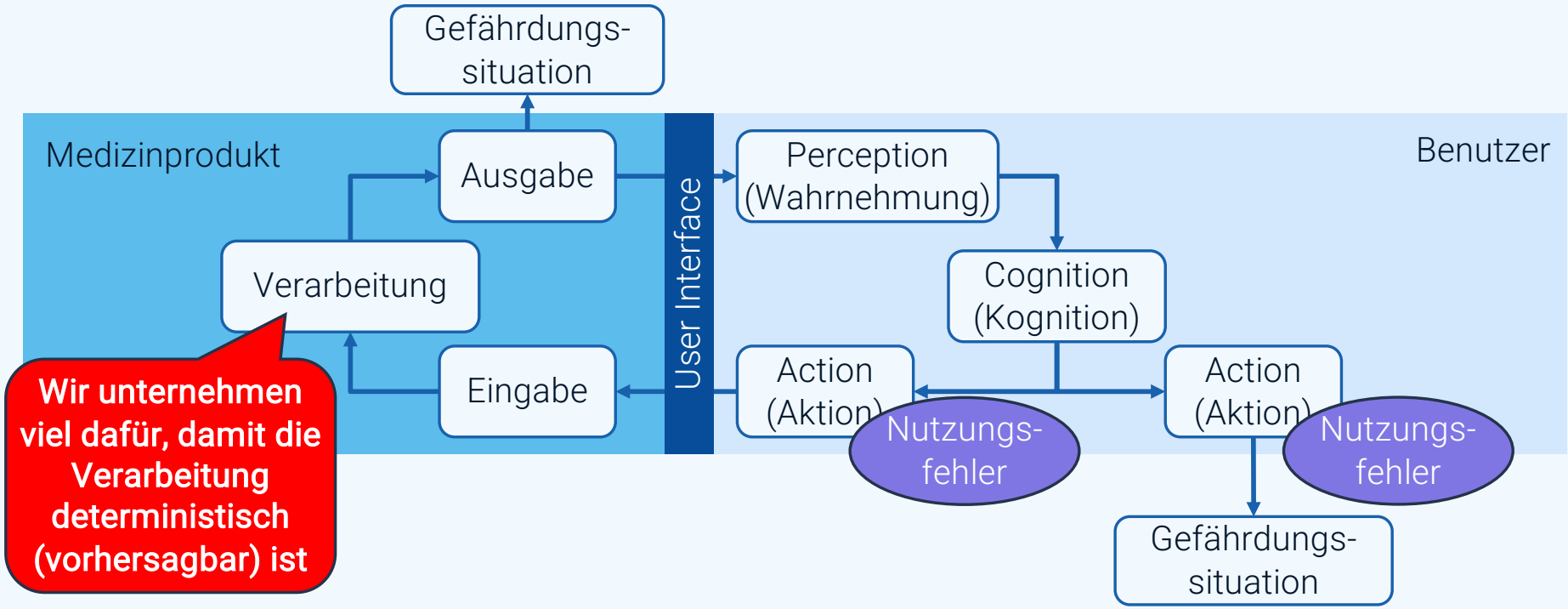
Arten von Interaktionen mit KI

1. Präsentation von Ergebnissen eines KI-Algorithmus
   - Beispiel: Zellzählung in der Pathologie, Highlights in bildgebende Diagnostik
2. Erarbeitung eines Ergebnisses in einem interaktiven Prozess zwischen Mensch und KI
   - Beispiel: Symptom-Checker, Clinical Decision Support Systems mit Large Language Model,
3. (Teil-)Autonome Steuerung von physikalische Prozessen
   - OP-Roboter mit KI-Unterstützung

Die Liste ist nicht erschöpfend, weitere KI-Interaktionen nehme ich gerne auf!

Was ist das Problem?

Perception-Cognition-Action-Modell bei AI/ML-fähigen Medizinprodukten

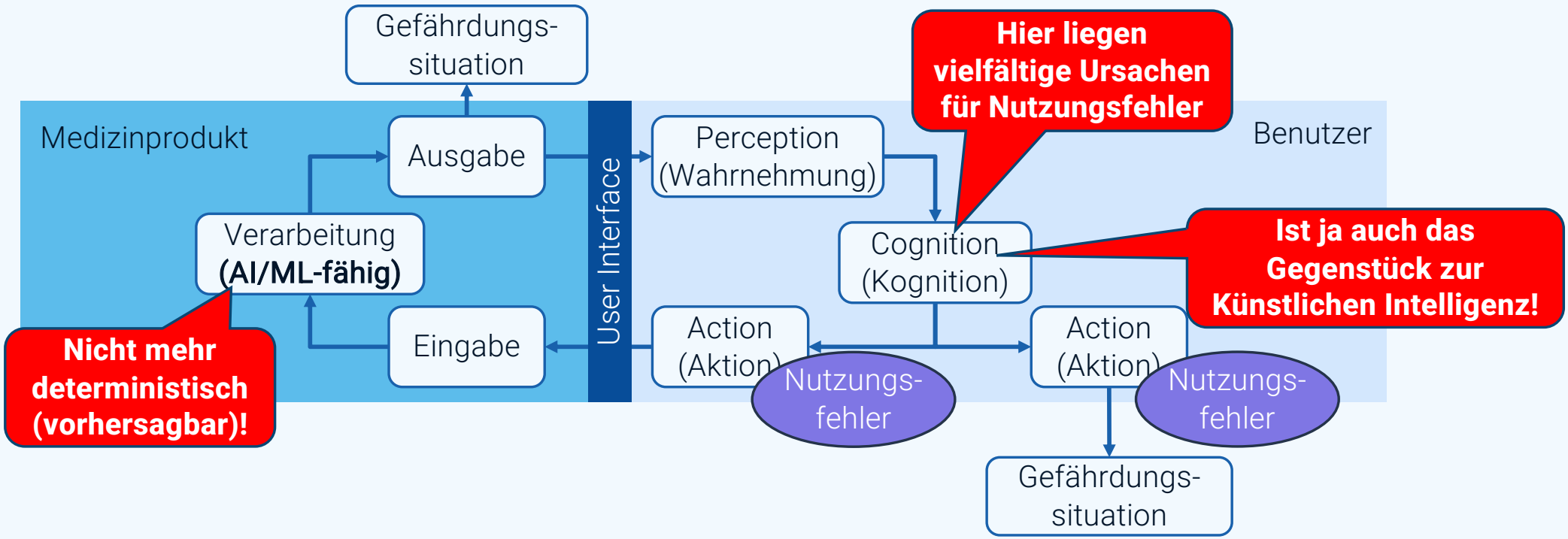


Quelle: IEC 6266-1 Medical devices – Part 1: Application of usability engineering to medical devices (eigene Übersetzung)

EMPIANA

Was ist das Problem?

Perception-Cognition-Action-Modell bei AI/ML-fähigen Medizinprodukten



Quelle: IEC 6266-1 Medical devices – Part 1: Application of usability engineering to medical devices (eigene Übersetzung)

EMPIANA

EMPIANA

1. Präsentation von Ergebnissen eines KI-Algorithmus

Wie vorgehen?

1. Präsentation von Ergebnissen eines KI-Algorithmus

Aus Usability-Sicht können wir Methoden weiterverwenden

- Methodik: **Qualitatives Usability Testing**
  - Ergebnis wird präsentiert
  - Benutzer wird bei der Interpretation des Ergebnisses beobachtet
  - Fehler werden erfasst und dann folgt ein Interview zur Bestimmung der Ursache
  - Anwendung der IEC 62366-1.

➔ Bekanntes Vorgehen, Methoden funktionieren weitgehend

➔ **Introspektion** ist wichtiger, da die Ursachen für Nutzungsfehler oft im Bereich der **Kognition** liegen.

Was ist das Problem?

Herausforderungen in der Bewertung von AI

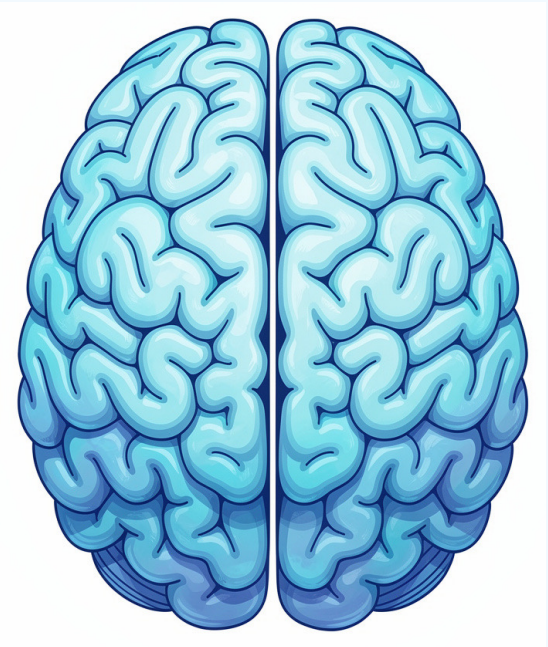
Jedoch zusätzlich:

- Werden die Benutzer einen **Bias** erkennen?
- Werden die Benutzer **Einschränkungen in der Leistungsfähigkeit** erkennen, wenn diese vorliegen?
- Wie werden die Benutzer die nötige **Überwachung** (human oversight) umsetzen können?

➔ Dafür müssen Benutzer entsprechende Mentale Modell aufbauen.

Was ist das Problem?

Mentales Modell



Ein mentales Modell ist eine interne Repräsentation der externen Realität: also, eine Art, die Realität im Geist abzubilden. Es wird angenommen, dass solche Modelle eine wesentliche Rolle bei der Kognition, dem Schlussfolgern und der Entscheidungsfindung spielen.

Source: https://en.wikipedia.org/wiki/Mental_model accessed 2024-04-24

- Das **Vorwissen über KI** und das **Wissen über das KI-Modell** spielen eine Rolle bei der Entwicklung eines robusten mentalen Modells.

Wie vorgehen?

Kommunikation über AI/ML-fähige Medizinprodukte

- FDA hat die Wichtigkeit von mentalen Modellen erkannt.
- Siehe FDA Draft Guidance:
  Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations

- Interessante Aspekte:
  - Kommunikation über AI/ML-fähige Medizinprodukte
  - Usability Evaluation

Wie vorgehen?

Kommunikation gemäß FDA Draft Guidance

- Transparenz
- Bias
- Leistungsmerkmale
- Einschränkungen
- Integration in Arbeitsabläufe
- Grad der Autonomie
- Menschliche Überwachung
- Wie Eingaben erfasst werden
- Erklärung was Ausgaben bedeuten
- Änderungen aufgrund neuer Modelle, etc.

Was ist das Problem?

Transparenz

- **Transparency** describes the degree to which appropriate information about a[n AI/ML-enabled medical device] [...] is clearly communicated to relevant audiences.

  Source: FDA Draft Guidance: Artificial Intelligence-Enabled DeviceSoftware Functions: Lifecycle Management and Marketing Submission Recommendations, 7 Jan. 2025

- beinhaltet
  - Zweckbestimmung (Welche medizinischen Indikationen werden bahandelt/diagnostiziert)
  - Entwicklung (e.g. Eigenschaften der Traingsdaten wie Patienteneigenschaften, Standorte für die Datenerhebung, genutzte Geräte zur Datenerhebung, usw.)
  - Leistung (statistische Parameter bspw. falsch negative/falsch positive Ergebnisse)
  - Logik, falls verfügbar (bswp. Typ des genutzten AI-Modells)

Herausforderung: Diese Informationen müssen während des Entwicklungsprozesses des AI/ML-Modells gesammelt werden.

Wie vorgehen?

Herausforderungen in der Bewertung von AI

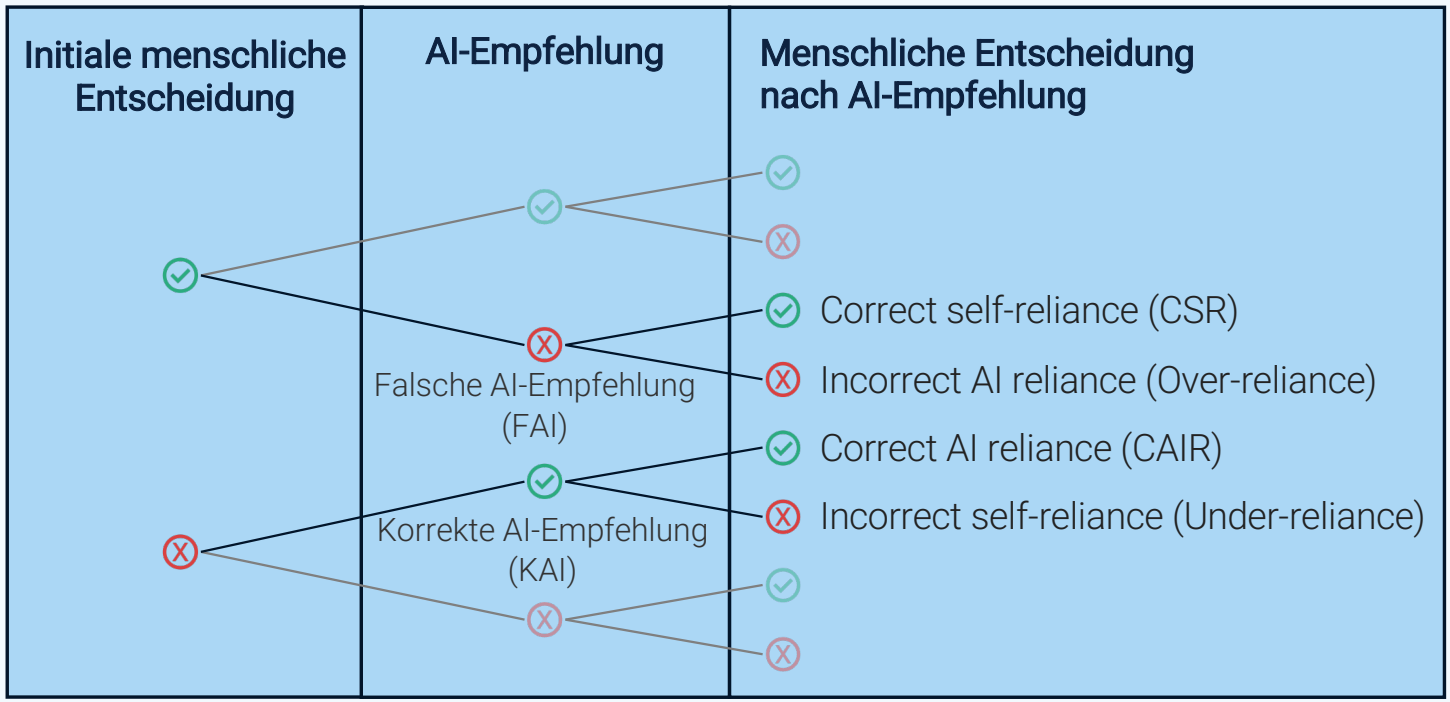
- Siehe FDA Draft Guidance:
  Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations
  - u.a. Appendix D: Usability Evaluation Considerations
  - User Interface Beschreibung
  - **Labeling** soll evaluiert werden.

Die FDA möchte Usability Evaluationen von **Informationen zur Transparenz** auch unterhalb der Schwelle von Critical Tasks im Human Factors Validation Testing.

EMPIANA

Wie vorgehen?

Approriateness of Reliance – Konzept zur Evaluation?



Maße für Appropriateness of Reliance:

Relative Self-Reliance = CSR/FAI

Relative AI Reliance = CAIR/KAI

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23). Association for Computing Machinery, New York, NY, USA, 410–422. https://doi.org/10.1145/3581641.3584066

EMPIANA

Wie vorgehen?

Approriateness of Reliance – Konzept zur Evaluation?

Kritik

- Zuerst muss immer eine menschliche Entscheidung ohne AI gefällt werden
  - Nicht immer sinnvoll/möglich
  - Verändert, wie stark etwas durchdacht wird
  - Human-AI-Team müsste ggf. direkt gegen ein anderes Human-AI-Team verglichen werden

Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23). Association for Computing Machinery, New York, NY, USA, 410–422. https://doi.org/10.1145/3581641.3584066

EMPIANA

2. Erarbeitung eines Ergebnisses in einem interaktiven Prozess zwischen Mensch und KI

Was ist das Problem?

2. Erarbeitung eines Ergebnisses in einem interaktiven Prozess zwischen Mensch und KI

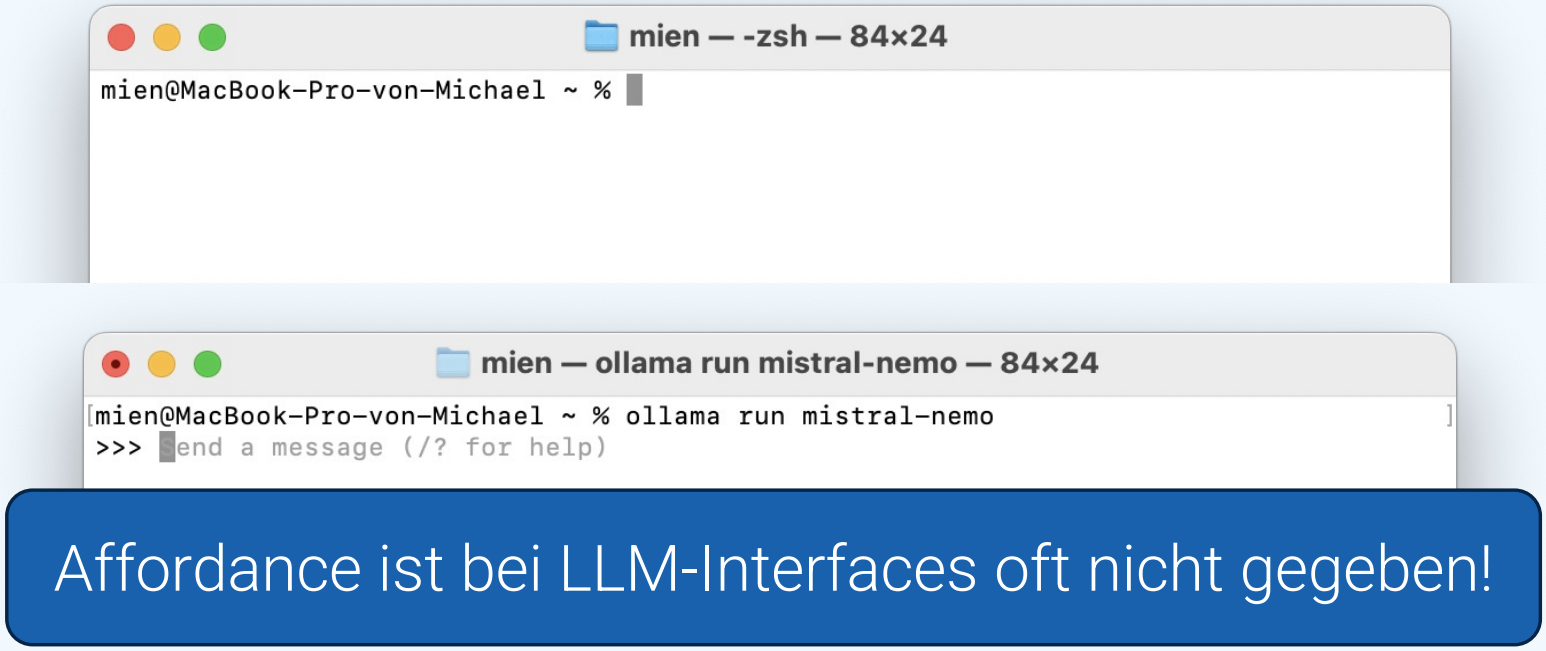
- Beispiele Symptom-Checker oder Decision Support Systeme mit Large Language Model (LLM): Infermedica, Ada Health, Prof Valmed etc.

- Herausforderung: Klassische qualitative Usability Tests funktionieren nur bedingt:
  - KI verhält sich nicht deterministisch.
  - Daher sind KI-Interaktionen immer anders.
  - Somit lassen sich vordefinierte Szenarien nicht mit der gleichen Reaktion testen.
  - Tests mit wenigen Nutzern (bspw. 15 pro Benutzergruppe) sind somit wenig reliabel. (Eine Wiederholung wird zu anderen Ergebnissen führen.)

Was ist das Problem?

Affordance bei LLM-Interfaces

– Affordance ist “ein Aspekt eines Objekts, der deutlich macht, wie das Objekt benutzt werden kann.”
(Quelle: CPUX-F Curriculum Certified Professional for Usability and User Experience Foundation Level Version 4.01 DE, 9. Januar 2023)



Affordance ist bei LLM-Interfaces oft nicht gegeben!

EMPIANA

Wie vorgehen?

Usability Tests

- Qualitative Usability Tests sind weiterhin wichtig, um zu verstehen, warum es Usability-Probleme gibt!
- Cleveres Testdesign ist nötig!
- Aber: Die Aussagekraft aufgrund des nicht-deterministischen Verhaltens des Systems ist begrenzt.
  - Wir müss(t)en mehr Interaktionen betrachten.
  - Wir müss(t)en statistisch relevante Aussagen treffen über Nutzungsfehler.
  - Aber: Dafür ist die Methode zu aufwendig.

Wie vorgehen?

Usability Tests – Beispiel eines Testdesigns: Symptom-Checker

- Usability Test eines Symptom-Checkers.
- Teilnehmende: Mussten in letzter Zeit einen Arzt wegen eines gesundheitlichen Problems aufgesucht haben.
- Testaufbau:
  - Symptome von Teilnehmenden beschreiben und dokumentieren lassen
  - Diagnose der Ärztin/des Artzes beschreiben und dokumentieren lassen
  - Teilnehmende Symptom-Check durchlaufen lassen
  - Eingabe in Symptom-Checker mit Symtomen vergleichen
  - Ergebnisse des Symptom-Checkers mit Diagnose vergleichen
  - Unterschiede evaluieren

Wie vorgehen?

Usability Tests – Beispiel eines Testdesigns: Symptom-Checker

- Vorteile
  - **Generiert Validierungs- oder Trainingsdaten → UXler können das!**
  - Objektive Messung der Nutzungs-Effektivität, durch Diagnose des Arztes
  - Nutzungsfehler können beobachtet und analyisert werden

- Nachteile
  - Wir brauchen eine deutlich größere Anzahl an Teilnehmenden als üblich, um eine Reliabilität zu erzeugen.
  - Beschreibung der Symptome vs. Eingabe der Symptome weist Unterschiede in alle Richtungen auf.
  - Testsetup funktioniert nur für die Fälle bei denen auch ein Arzt erforderlich war. Wenn Teilnehmende (richtigerweise) nicht zum Arzt gegangen sind, dann sehen wir diese Fälle nicht
  - ...

Wie vorgehen?

Unbeobachtete Usability Tests

- Teilnehmende führen Aufgaben am KI-System aus, werden dabei aber nicht von einem Moderator und Beobachter begleitet.
- Möglich bei ähnlichem Setup wie bei Symptom-Checkern
  - Einleitende Fragen
  - Aufgabe
  - Schließende Fragen
- Datenauswertung durch Logging, Fragebögen und Screen-Recordings.

Wie vorgehen?

Klinische Prüfungen in Kombination mit Usability Evaluation

- Gut denkbar bei Diagnostik
- Bspw. bei der einem Clinical Decision Support System
  - Beobachtete Diagnostik mit KI/ML-fähigem Medizinprodukt durch Teilnehmende.
  - Parallel dazu eine Diagonstik durch einen erfahrenen Spezialisten.
  - Vergleich der Ergebnisse und bei Abweichung eine Ursachen-Analyse.
  - Dies muss einen Anteil mit Moderator und Beobachter haben, kann aber auch teilweise unbeobachtet stattfinden.

Wie vorgehen?

Post-market survaillance

Klassisches Vorgehen

– Complaints
– Social Media
– etc.

Nur mit Software möglich:

– Logs / flight recorder
– Benutzer werden während der Nutzung angehalten Feedback zu geben

➔ **Das sind aber Features, die aktiv entwickelt werden müssen!**

Was ist das Problem?

Dynamisch generierte User Interfaces

– Nachgestelltes Beispiel mit generiertem Interface

EMPIANA

Wie vorgehen?

Automatische Tests auf Style Guides und automatisierte Usability Tests

- Agenten können die Einhaltung von Style-Guides prüfen, wenn der Style-Guide in maschinenlesbare From vorliegt.
- Bei Fehlern wird das UI erneut gerendert.
- Zudem Log-Eintrag zur Verbesserung der Generierung.

- Automatisierte Usability Tests beginnen eine Option zu sein.

EMPIANA

3. (Teil-)Autonome Steuerung von physikalische Prozessen

Was ist das Problem?

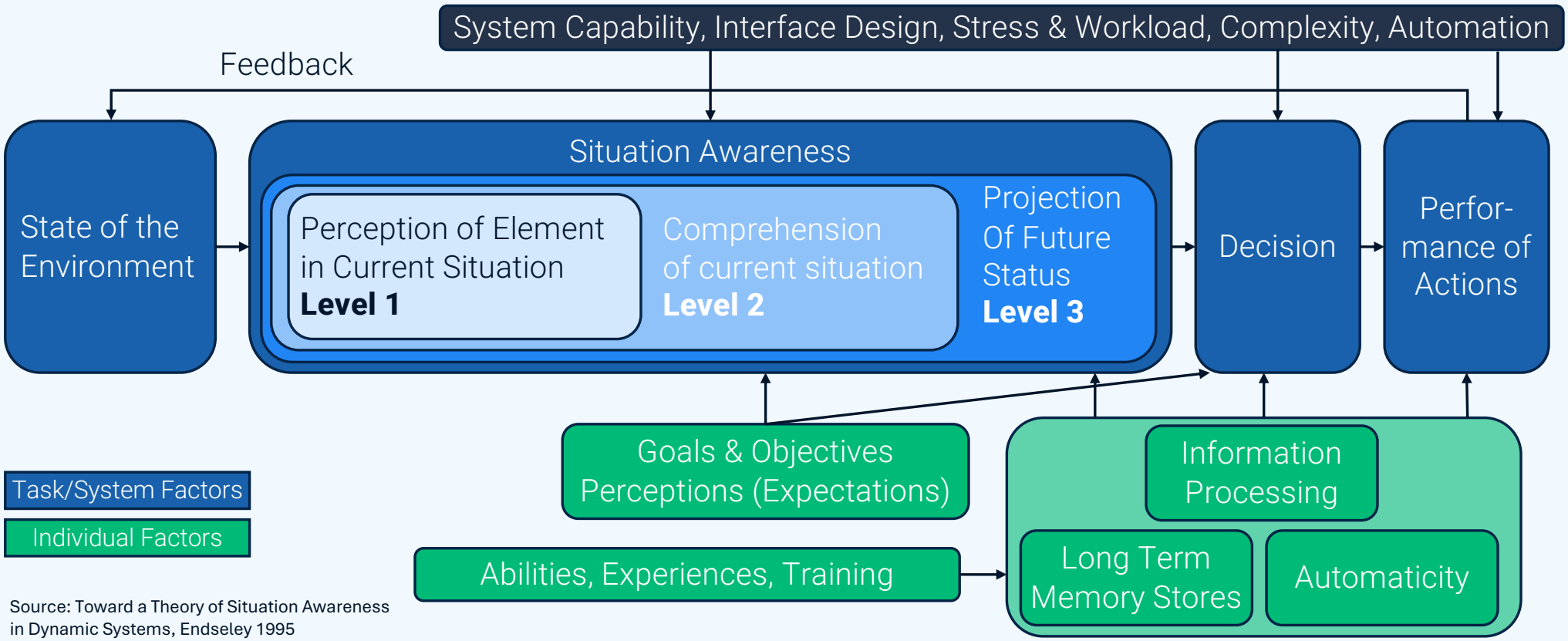
3. (Teil-)Autonome Steuerung von physikalische Prozessen

– Hat der Nutzer ein Situationsbewusstsein (Situational Awareness)?

– Dies ist auch ein Teil von Human-AI teaming.

Was ist das Problem?

Situation(al) Awareness



Source: Toward a Theory of Situation Awareness in Dynamic Systems, Endseley 1995

EMPIANA

Wie vorgehen?

3. (Teil-)Autonome Steuerung von physikalische Prozessen

Aufwendigs Vorgehen, aber für komplexe Systeme vermutlich sinnvoll:

- Informationsbedarf ermitteln.
  - Bspw. mit einer Goal-Directed Task Analysis (GDTA)
  - Dabei werden die Ziele und Aufgaben zerlegt und der **Informationsbedarf** zu jemandem Zeitpunkt ermittelt

- User Interface evaluieren mit Situation Awareness Global Assessment Technique (SAGAT)
  - Fragen zur Siutational Awareness aus dem Informationsbedarf entwickelt
  - Benuzter führen Test-Aufgaben aus. Die Aufgaben werden **unterbrochen** und **Fragen zur Situational Awareness** gestellt

EMPIANA

Normung: IEC/TS 62366-3

Medical devices – Part 3: Guidance on the application of usability engineering to medical devices using artificial intelligence and machine learning technology

Was normen wir?

IEC/TS 62366-1

- Technical Specification basiert auf
  - “Usability Engineering for Medical Devices using Artificial Intelligence and Machine Learning Technology” von Stangenberg, Martin (Project leader), Engler, Michael (Editor), Johner, Christian (Editor), Krepcke, Martin (Editor), Martinez Torres, Manuel Isaac (Editor)
  - Link: https://zenodo.org/records/14203190

- Interpretation der IEC 62366-1 in Bezug auf AI/ML-enabled medical devices
- Erste Überarbeitungsrunde beendet
- Erste öffentliche Version zur Kommentierung folgt

EMPIANA

Was normen wir?

Scope

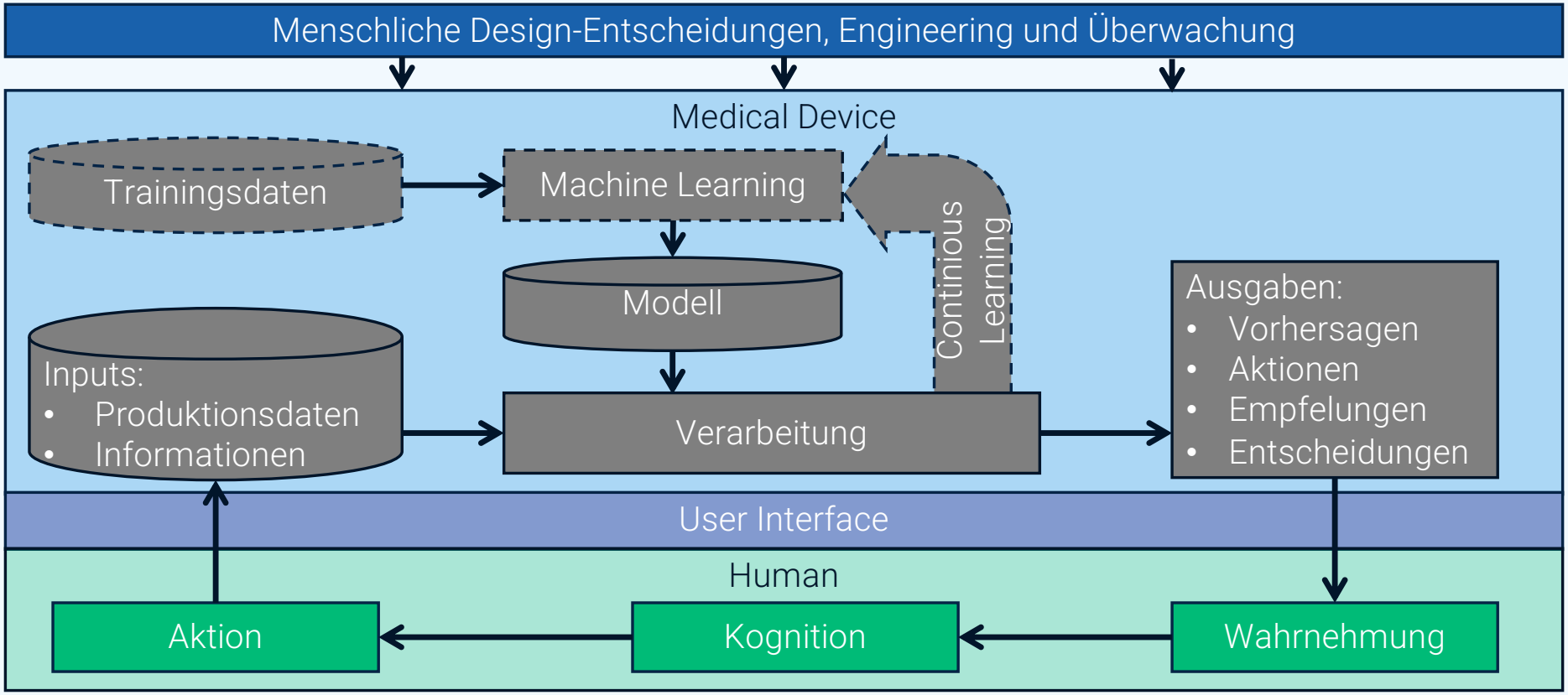
Nutzungsbezogene Risiken:

- Es geht nur um sichere Nutzung!
- Effektivität, Effizienz und Zufriedenheit in der Nutzung werden nicht betrachtet.
- Kann aber interessant sein für Marketing Claims!

- Initiales Paper geht auch auf Continious Learning ein, dass ist in der TS nicht mehr der Fall!

Was normen wir?

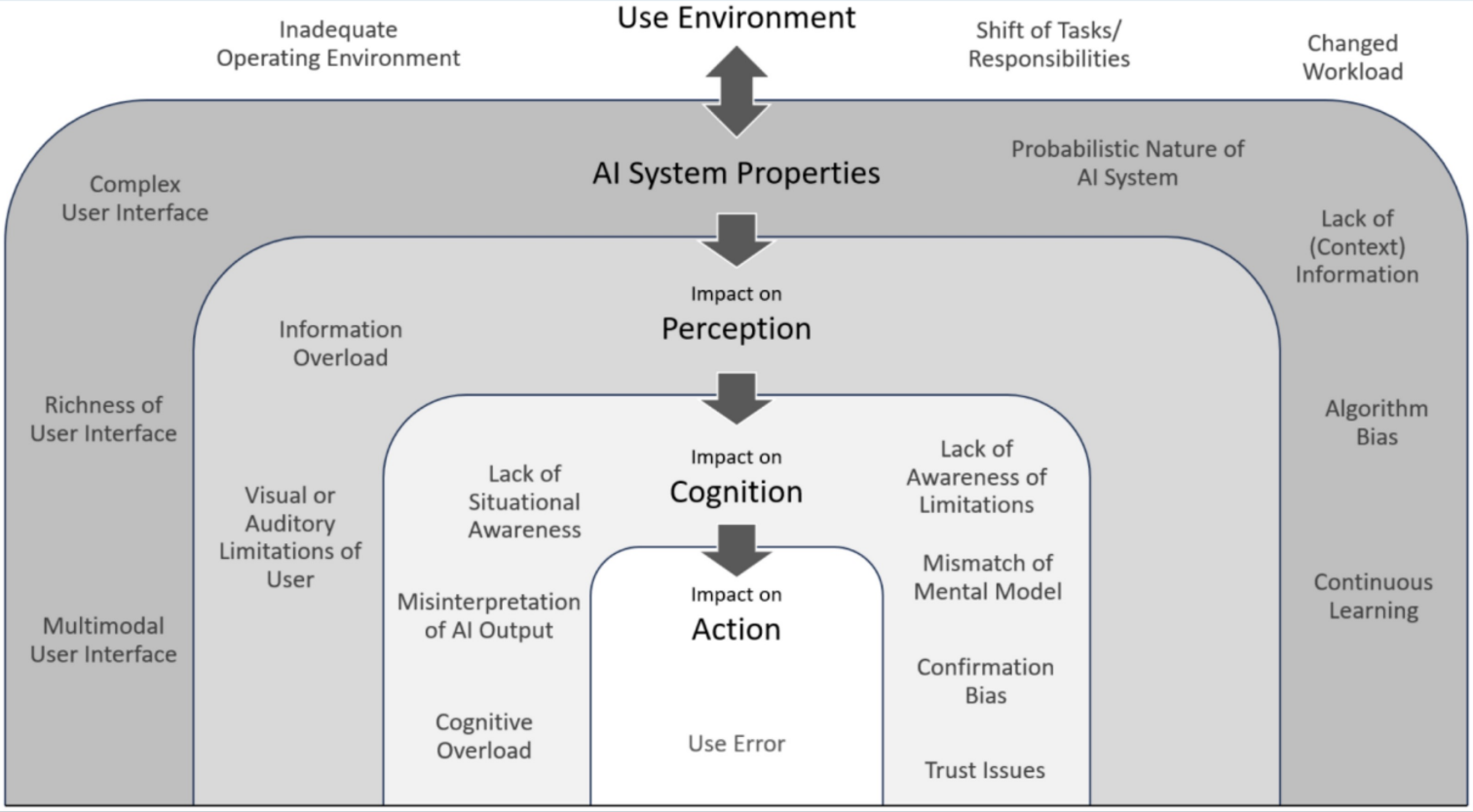
Adaptiertes PCA-Modell



EMPIANA

Was normen wir?

Faktoren, die zu Use Errors führen



Quelle: Usability Engineering for Medical Devices using Artificial Intelligence and Machine Learning Technology; Engler, Johner,Krepcke, Martinez-Torres, Stangenberg

EMPIANA

Was normen wir?

Inhalt

– Empfehlungen zu den einzelnen Kapiteln der IEC 62366-1, die (noch) wenig über die IEC 62366-1 hinausgehen.

Adressiert wird vor allem der erste Punkt:

1. Präsentation von Ergebnissen eines KI-Algorithmus
2. Erarbeitung eines Ergebnisses in einem interaktiven Prozess zwischen Mensch und KI
3. (Teil-)Autonome Steuerung von physikalische Prozessen

Was normen wir?

Inhalt

- Das Paper geht auch auf Post-Market Survaillance ein.
- Es ist umstritten, ob die IEC/TS 62366-3 auch auf Post-Market Surveillance eingehen soll.
  - Dafür spricht: Für AI/ML-fähige Medizinprodukte muss hier aus Usability-Sicht einiges getan werden.
  - Dagegen spricht: IEC 62366-1 hat nur den Sopce der Entwicklung.

Was sollten wir normen?

Ideen

- Stärkerer Fokus auf **Transparenz-Informationen**, um anschlussfähig an die FDA Guidances zu sein.
- Anhang zum Thema, dass **Labeling-Tools** menschliche Fehlerquellen beinhalten und eine Usability-Betrachtung sinnvoll ist.
- Weitere **kognitive Verzerrungen** beschreiben, nicht nur Confirmation Bias
  - Automation bias, etc.
  - Mehr zu Human-AI-Teaming und wie der Nutzungsworkflow gestaltet sein muss.
- **Methoden für summative Evaluation** von LLM-basierten und dynamischen Interfaces einbringen: Als Teil der Klinischen Evaluation, unbeobachtet Usability-Tests, statistisch relevante Usability-Tests, etc.
- **Post-Market Surveillance** Aktivitäten aus Usability-Sicht ☺
- **Weitere Ideen sind gern gesehen!**

Welche Erfahrungen habt ihr?
Was können wir zusammen tun?

Michael Engler

EMPIANA GmbH
Rüttenscheider Straße 120
45131 Essen

michael.engler@empiana.com
+49 1522 9240498

https://www.linkedin.com/in/michael-engler-med/