



Um ethische Reflexion erweiterte Model Cards für Autonome Systeme

Ein Leitfaden zur ethischen Gestaltung autonomer Systeme im
Rahmen des Projekts EGAR: Ethische und gesellschaftliche
Aspekte Autonomer Robotik

Aline Franzke (IZEW, Eberhard Karls Universität Tübingen)
Sebastian Gießler (IZEW, Eberhard Karls Universität Tübingen)
20. Januar 2026

Dieser Leitfaden entstand im Rahmen des Forschungsprojekts: "Ethische und gesellschaftliche Aspekte autonomer Robotik (EGAR)"

Unter der Leitung von Dr. Mone Spindler und Dr. Wulf Loh

Finanziert von der Baden-Württemberg Stiftung

1. Ethisch angepasste Model Cards für die Robotik

Model Cards sollen die Eigenschaften, Kapazitäten und intendierte Nutzung von Machine-Learning Systemen dokumentieren, und eine Transparenz über die beabsichtigte Verwendung und technische Grenzen bieten. Ziel dieses Dokuments ist es, zum einen Model Cards als strukturierte Dokumentation vorzustellen, und im folgenden Fragenkatalog um ethische und robotikspezifische Aspekte zu erweitern. Wir haben uns für Model Cards entschieden, da dieser Ansatz aus einer technischen Perspektive entwickelt wurde und wir uns von daher eine höhere Akzeptanz und eine bessere Integration in vorhandene Arbeitsprozesse und wissenschaftliche Praktiken versprechen. Dieser Ansatz ist explizit auch für Projekte anwendbar, die keine Machine Learning Ansätze verwenden.

Ein Model Card Ansatz für die Robotik wurde bisher nicht ausgearbeitet. Wir erweitern im Fragenkatalog Model Cards um Aspekte der Robotikentwicklung, die auf der Softwareebene nicht repräsentiert sind, etwa Interaktion mit Umwelt und Personen, und Roboterdesign und Körperlichkeit (*embodiment*). Wir verwenden diesen Ansatz, als ein Vorbild (Role Model) für andere computergestützte Technikwissenschaften, da Model Cards wesentliche Faktoren der Softwareentwicklung im Machine Learning Bereich berücksichtigen.

Was sind Model Cards? Der Model Card Ansatz kommt ursprünglich aus der Softwareentwicklung (*software engineering/SWE*) und wird vor allem in diesem Kontext eingesetzt. Der Fokus liegt dabei auf klassischen Problemfeldern im Machine Learning, etwa Details über das Modell, Datenherkunft, Verwendung von Trainingsdaten, intendierter Einsatzzweck eines Systems, vorgenommene Gewichtungen, und Limitierungen des Modells (Mitchell et al. 2019). Model Cards und dieser Fragenkatalog sind freiwillige Dokumentationen und keine rechtlich bindenden Nachweispflichten. Sie sind ein sogenanntes *Soft Law*, also eine freiwillige Übereinkunft oder Absprache, im Gegensatz zu *Hard Law*, einer zwingenden Rechtsnorm. *Soft Law* und *Hard Law* sind Begriffe der Corporate Governance. Beispiele für Hard Law sind etwa die EU-Maschinenverordnung (2023/1230/EU), dem deutschen Medizinproduktgesetz (MPDG-2020) oder der EU-AI-Act (2024/1689/EU). Weitere Beispiele für *Soft Law* sind Ethikrichtlinien und freiwillige Selbstverpflichtungen von Institutionen oder Unternehmen.

Ein Nachteil des Model Card Ansatzes ist, dass die technische Ebene zwar strukturiert bearbeitet wird, die ethische Ebene jedoch nicht. „Ethical Considerations“ werden zwar explizit mit Beispielen erwähnt, es wird aber keine strukturierte Anleitung, wie genau ethische Fragen zu bearbeiten sind, angeboten. (Mitchell et al. 2019: 6). Fragen zu ethischen Implikationen werden häufig an Ethikkommissionen, Institutional Review Boards (IRB) oder Research Ethics Boards (REB) verwiesen. Ein Problem dabei ist, dass diese Institutionen oftmals nicht für ethische Aspekte der Informatik oder der Ingenieurwissenschaften zuständig sind. Insbesondere in kleinen Organisationen, wie

Unternehmen oder Hochschulen, fehlt es oft gänzlich an einer entsprechenden Anlaufstelle. Dadurch sind Entwickler*innen häufig auf ihre individuelle Selbsteinschätzung angewiesen, was zu Inkonsistenten oder unzureichenden ethischen Bewertungen führen kann.

Model Cards for Model Reporting

FAT* '19, January 29–31, 2019, Atlanta, GA, USA

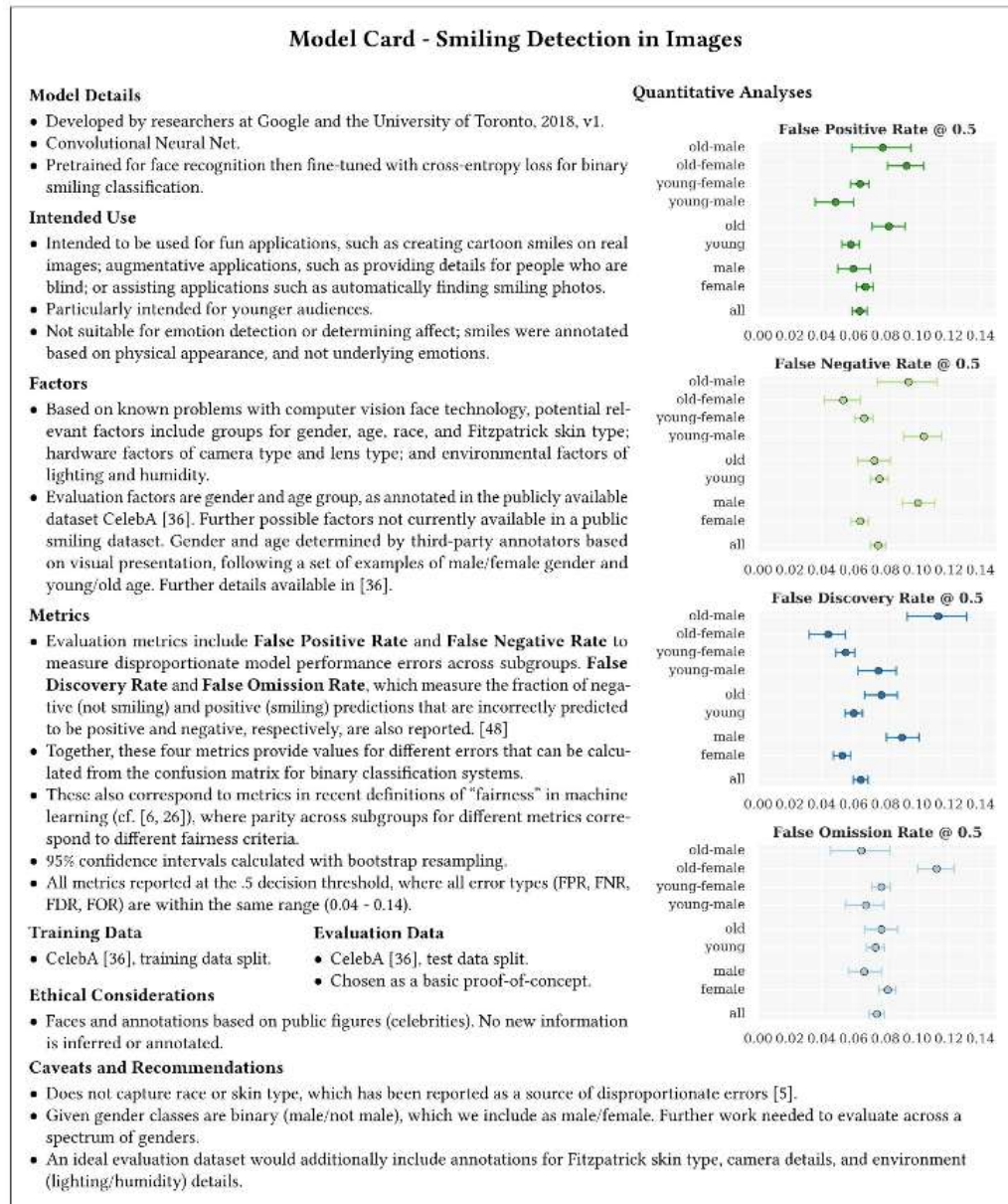


Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., et al. (2019). Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). <https://doi.org/10.1145/3287560.3287596>

Diese Systematik hat für die Projekte mehrere Vorteile:

*Robotikentwickler*innen*

Andere Entwickler*innen können anhand der Model Card for Robotics können nachvollziehen, wie ein technisches, soziales oder ethisches Problem adressiert wurde und für die eigenen Problemlösungen als Referenz verwenden.

*Policy*makers*

Politische Entscheidungsträger*innen können anhand der Model for Robotics ethische und soziale Risiken abschätzen, und auf deren Grundlage entscheiden, ob Projekte gefördert werden sollten, Regulation nötig ist und ob soziotechnische Lösungsversprechen realistisch sind.

Organisationen

Anhand von Model Cards for Robotics können Organisationen, z.B. Unternehmen, entscheiden, ob sie ein bestimmtes System anschaffen wollen, wie sie es rechtssicher einsetzen können, und wie dieses System optimal in die unternehmenseigenen Strukturen eingebunden werden kann.

Betroffene Personen

Betroffene Personen können anhand der Dokumentation leichter nachvollziehen, welche Auswirkungen das Autonome System auf Sie und ihre Lebenswelt haben könnte, und können fahrlässigen oder missbräuchlichen Einsatz leichter erkennen und kritisieren.

Ethische angepasste Model Cards für Autonome Systeme

Die *Model Card for Robotics* ist idealerweise ein einseitiger 'Beipackzettel', welcher für die relevanten Stakeholder, Förderorganisationen und andere Forscher*innen vorgehalten wird. Die Verschränkung von Model Cards mit ethischer Reflexion kann als ein Werkzeug für ethische Mitigation verstanden werden. Diese bietet für Roboterentwicklungsprojekte einen bedeutenden Mehrwert während der Design- und Entwicklungsphase. Internationale Zeitschriften und Konferenzen setzen ebenso zunehmend Ethikevaluationen voraus.

Durch die Integration der ethischen Reflexion und der Dokumentation durch Model Cards adressieren wir ein weiteres, strukturelles, Problem: Im Vergleich zu anderen Ansätzen im Bereich der Technikfolgenabschätzung und ELSA, wird der ethische Impuls durch die Entwickler*innen selbst durchgeführt und in die Projektpraktiken verankert: ethische und soziale Aspekte werden so in den Forschungsprozess integriert, und ELSA-Forscher*innen werden nicht erst am Ende des Projekts hinzugezogen (Post-hoc) und bewerten einen fertig entwickelten Demonstrator.

Modellcards und EGAR - Fragenkatalog

Dieser Fragenkatalog soll diese Lücke schließen und damit beide Ebenen verbinden – eine strukturierte technische Dokumentation und darauf aufbauend, eine strukturierte ethische Analyse und Risikobeurteilung. Um die Zugangshürde zu senken, kann dieser Fragenkatalog ohne vorheriges Erstellen einer Model Card bearbeitet werden. Einige Themen und Fragen kommen daher sowohl beim Model Card Ansatz als auch in diesem Fragenkatalog vor.

Der optimale Zeitpunkt für das Bearbeiten des EGAR-Fragenkatalogs ist, wenn bereits ein Grundkonzept erarbeitet und die Anforderungen vorliegen. Das ist wichtig, um zu klären, *welche* Probleme *wie* mit einem Roboter gelöst werden sollen bevor die verwendeten Technologien, die Datensätze sowie der Einsatzzweck eines Projekts bereits feststehen. Dies entspricht in der Regel dem ersten Drittel bis zur Hälfte des Entwicklungsprozesses.

Im Idealfall gibt es eine definierte Person oder Rolle im Entwicklungsteam, die für die ethische Bewertung zuständig ist. Dadurch wird sichergestellt, dass eine Person diese Themen bearbeitet, darauf achtet, dass Richtlinien eingehalten werden und diese durchsetzt. Dennoch bleiben ethische und soziale Fragestellungen eine Querschnittsaufgabe und sollten daher im gesamten Entwicklungsteam integriert werden.

Der Fragenkatalog hat zwei Funktionen: Fragen nach technischen Eigenschaften, Leistungsfähigkeit und intendierten Einsatz autonomer Systeme werden überprüft und in ethisch-normative Argumente eingebettet. Beispielsweise führt die Frage, welche Datensätze verwendet wurden, zu Folgefragen, ob bei eigenen und fremden Datensätzen ethische und soziale Aspekte berücksichtigt wurden. Für die technische Ebene sind dabei die Model Cards nach Mitchell et al. 2019 ein Vorbild¹, die für die domänenspezifischen Bedürfnisse der Robotikentwicklung erweitert werden. Diese Überlegungen richten sich an Entwickler*innen, Förderinstitutionen, und anderen Forschungsgruppen, die technische Systeme auf ethische und soziale Probleme analysieren und basiert auf einer Kombination von reflexiven und formalen Ansätzen der Technikfolgenabschätzung.

Reflexive Ansätze

Reflexive Ansätze konzentrieren sich auf die kritische Untersuchung der Überzeugungen, Annahmen und Praktiken von Forscher*innen und wie diese den Forschungsprozess beeinflussen. Reflexivität erfordert das Hinterfragen der eigenen Position und bewussten und unbewussten Überzeugungen des Forschers und die Anerkennung ihrer Rolle bei der Gestaltung des Forschungsprozesses. Reflexive Ansätze können Forscher*innen herausfordern, sowohl ihre persönlichen Perspektiven als auch die Annahmen innerhalb ihrer Disziplin zu untersuchen.

¹ Vorbild im Sinne eines Role Models oder established protocol Ansatzes.

Diese Methoden basieren auf der Kommunikation und Praxis reflexiver Argumente in strukturierten Diskussionen. Diskussionen beruhen in der Regel auf einem dialogischen Konzept, das den Austausch zwischen ethischen Expert*innen (oder Werkzeugen) und Entwickler*innen fördert. Sie sind oft niedrigschwellig angelegt, um ethische Probleme sichtbar und intersubjektiv zugänglich zu machen. Gewählte Methoden sind in der Regel strukturierte oder teilstrukturierte Fragebögen, Workshops oder dialogische Zielgruppenanalysen und stützen sich auf interdisziplinäre Perspektiven, um ethische Probleme zu bearbeiten.

Reflexive Ansätze sind oft mit Herausforderungen konfrontiert, da sie nur begrenzt operationalisierbar sind und von der Bereitschaft der Teilnehmenden zum Dialog abhängen. Sie sind außerdem zeitaufwändig und erfordern interdisziplinäre Arbeit, die eine gemeinsame Zielvorstellung und Fachsprache erfordern, um komplexe technische und ethische Probleme über verschiedene disziplinäre Hintergründe hinweg zu verstehen. Zudem ist fraglich, ob ethischer Input zu dauerhaften Veränderungen in den Prozessen führen kann.

*Beispiel: Unser Prototyp funktioniert perfekt, warum weigern sich unsere Nutzer*innen, ihn zu benutzen?*

Formale Ansätze

Formale Ansätze bieten ein standardisiertes Verfahren, das oft aus den technischen Wissenschaften selbst stammt oder in Zusammenarbeit mit ihnen entwickelt wurde. Beispiele hierfür sind Ansätze wie Model Cards und Data Sheets. Diese Ansätze werden häufig aus der Informatik oder Ingenieurwissenschaften heraus entwickelt und basieren auf den üblichen Dokumentationsanforderungen technischer Systeme, d.h. sie erfordern wenig zusätzlichen Aufwand, da sie sich auf etablierte Praktiken in der Informatik und in den Ingenieurwissenschaften beziehen.

Dadurch sind sie in der Regel gut an die fachspezifischen Methoden, Prozesse und Vorgehensweisen dieser Disziplinen angepasst. Dadurch lassen sich diese formalen Ansätze reibungslos in die Arbeitsabläufe in Forschung und Entwicklung integrieren - eben, weil sie von vornherein so konzipiert wurden. Im Vordergrund steht dabei die sachgerechte Dokumentation, d.h. Designentscheidungen, Entscheidungen über die verwendeten Datensätze, Datenherkunft, für welchen Zweck das System entwickelt wurde, und Details über die verwendeten statistischen und informatischen Methoden - ein ‚Beipackzettel‘ für Technik.

Beispiel: Mein Prototyp funktioniert, wie dokumentiere ich Leistungsfähigkeit und Grenzen für die technische Fachcommunity?

2. Fragenkatalog zu ethischen und gesellschaftlichen Aspekten Autonomer Robotik

Was ist das konkrete Ziel des Fragenkatalogs?

Die Dokumentation von Entwicklungsschritten und zentralen Designentscheidungen sorgt für *Nachvollziehbarkeit* für andere Entwicklungsteams und Forschungsprojekte sowie der *Transparenz* gegenüber Stakeholdern und Geldgebern. Sie ermöglicht es, Unsicherheiten und *Entscheidungsprozesse sichtbar zu machen* und trägt dazu bei, *intendierte sowie nicht intendierte Konsequenzen frühzeitig abzuschätzen*.

In der Technikentwicklung existieren häufig mehrere technische Lösungsansätze, deren Wahl sowohl funktionale als auch ethische Implikationen mit sich bringt. Ein Beispiel hierfür ist die Entscheidung zwischen zwei Sensoralgorithmen: Algorithmus A ist dem Entwicklungsteam vertraut, sammelt jedoch mehr Daten als für das System erforderlich. Algorithmus B erfüllt die technischen Anforderungen ebenso, ist aber datensparsamer. Die bewusste Wahl von B vermeidet ein potenzielles ethisches Problem im Umgang mit sensiblen Daten. Solche Reflexionsprozesse sind besonders im Kontext der Informatik und des Ingenieurwesens relevant, da hier eine Tendenz zur Werkzeughörigkeit sowie zu Standardverfahren („Defaultism“) besteht, die oft unkritisch übernommen werden, weil „es schon immer so gemacht wurde“.

Wer sollte diesen Fragebogen ausfüllen?

Der Fragenkatalog kann sowohl von *Einzelpersonen* als auch von *Teams* als Diskussionsgrundlage verwendet werden. Eine Bandbreite von unterschiedlichen Expertisen innerhalb eines Projektteams ist hilfreich, wenn auch nicht zwingend nötig. Idealerweise ist eine Person im Team für diesen Fragenkatalog und Ergebniskontrolle zuständig.

Wie lange dauert es, diesen Fragenkatalog auszufüllen?

Es ist davon auszugehen, dass der größte Teil des Fragenkatalogs innerhalb von *60 Minuten* beantwortet werden kann. Die Klärung einzelner Fragen kann dabei auf einen spezifischen Folgetermin verschoben werden, da erst noch notwendige Detailinformationen eingeholt werden müssen.

Wann ist der ideale Zeitpunkt, um diese Fragen zu beantworten?

Grundsätzlich ist der ideale Zeitpunkt für ethische Reflexion *so früh wie möglich* in einem Entwicklungsprozess und bevor eine wichtige *neue Phase* innerhalb des Projekts beginnt. So bietet es sich etwa an, vor der Designphase eines Prototyps zu prüfen, ob ethische und gesellschaftliche Aspekte berücksichtigt werden müssen, in der Testphase, und bevor ein Projekt abgeschlossen wird. Ethische Reflexion kann daher punktuell, aber auch durch den Projektprozess fortlaufend stattfinden.



Was passiert, wenn Sie Fragen nicht beantworten können?

Grundsätzlich ist es nicht auszuschließen, dass diese konkrete Frage für Ihr Projekt nicht relevant ist. Häufig ist es in einem Projekt nicht klar, welche Person das Wissen oder die Zuständigkeit hat, eine bestimmte Frage zu beantworten. Bitte bestimmen Sie eine verantwortliche Person im Team, die sich der offenen Fragen annimmt.

Was passiert nach dem Ausfüllen des Fragebogens?

Analysieren Sie erst, welche Fragen noch nicht zufriedenstellend beantwortet wurden, und legen Sie ein Meeting und eine verantwortliche Person fest, indem diese Aspekte weiterentwickelt werden sollen. Das ausgefüllte Dokument kann in der von Ihnen verwendeten Projektdokumentation abgelegt werden. Dieses Dokument kann im Folgenden für die Vorbereitung von Veröffentlichungen für Journale und Konferenzen verwendet werden, die ethische Analysen in ihren Richtlinien vorschreiben, oder es kann an eine interne oder externe Ethikkommission weitergegeben werden. Dieses Dokument ebenfalls dient dem Wissensmanagement für die beteiligten Institutionen, Wissenschaftler*innen und Projekte. Wenn in einem zukünftigen Projekt ähnliche ethische und soziale Aspekte auftreten, kann darauf zurückgegriffen werden.



Ethische und soziale Aspekte Autonomer Robotik

Ein Fragenkatalog zur ethischen Gestaltung autonomer Systeme im Rahmen des Projekts EGAR: Ethische und gesellschaftliche Aspekte Autonomer Robotik

Frage 1: Grundüberlegungen

Diese Fragen zielen darauf ab, auch vagen Ideen Ausdruck zu verleihen und zu überlegen, ob es Möglichkeiten gibt, diese technisch einzubinden.

- a) Welche grundsätzlichen ethischen Bedenken haben Sie, wenn Sie an Ihr Projekt denken?

- b) Welche Schritte kommen Ihnen spontan in den Sinn, wie Sie bereits in der Entwicklung ethische Risiken minimieren könnten? Nennen Sie diese explizit.

Frage 2: Datensätze und Datenherkunft

Diese Fragen decken grundlegende Themenkomplexe der Datenverarbeitung, Datenherkunft und Datenqualität ab.

- a) Welches Unternehmen hat die Software oder Teile davon erstellt, welche Komponenten wurden von Dritten übernommen?

- b) Welche möglichen Schwierigkeiten könnten daraus resultieren?



- c) Woher stammen die verwendeten Daten? Wurden die Daten im Projekt generiert oder wurde ein externer Datensatz verwendet (z.B. wiederverwendete Forschungsdaten aus einem anderen Projekt, lizenzierte Datensätze)? Wenn beides der Fall, wie sind die Anteile an selbst erzeugten Daten und externen Daten?

- d) Wurden bereits Daten der Stakeholder verwendet (Endnutzer*innen, einsetzende Unternehmen)?

- e) Welche Maßnahmen zur Sicherung der Datenqualität wurden angewendet?

Frage 3: Systemeigenschaften

Diese Fragen behandeln grundlegende Eigenschaften des Systems, wie (autonome) Entscheidungen des Systems zustande kommen und welche technischen Verfahren dabei zum Einsatz kommen.

- a) Inwiefern handelt der Roboter selbstständig?

- b) Kann man von autonomen Entscheidungen sprechen, z.B. im Sinne eines Sense-Plan-Act Modells?



- c) Reagiert das System auf die Umwelt (z.B. durch Verfahren Maschinelles Lernens), oder handelt es entlang von vorgehenden Entscheidungsbäumen (decision trees)?

- i. Wenn ein ML-System verwendet, wird: Welche Annahmen und Gewichtungen werden im ML-Verfahren angewandt und warum wurde diese gewählt?

- ii. Wie wurde das Verfahren gegen Diskriminierung und andere ungewollte Einflüsse aus den Trainingsdaten gesichert? (z.B. Proxyvariablen)

Frage 4: Nachvollziehbarkeit des Systemverhaltens

Diese Frage bezieht sich auf die Nachvollzieh- und Erklärbarkeit des Systems durch technisches und nicht-technisches Personal, Nutzer*innen und weitere Stakeholder.

- a) Werden Komponenten von Third-Party-Providern und deren dependencies ausreichend dokumentiert?

- b) Wie transparent ist der Entscheidungsweg des Systems?



- c) Wie ist das Verhältnis von falsch-positiven und falsch-negativen Schlussfolgerungen?

- d) Wie hoch ist die Fehlerquote?

- e) Welche Fehlerquote gilt als akzeptabel?

- f) Bietet das System eine Erklärung an, wie eine spezifische Entscheidung zustande gekommen ist?

- i. Wenn Ja: Bietet das System eine Übersicht an, welche Faktoren zu dieser Entscheidung geführt haben?

- ii. Wenn Ja: Bietet das System eine Übersicht an, welche Faktoren zu einer anderen Entscheidung geführt hätten? (Counterfactual explanations)

Frage 5: Schadensbegrenzung

Dieser Fragenkomplex behandelt Themen von möglichen Personen- oder Sachschäden, der rechtlichen Produktklasse sowie Haftungsfragen

- a) Welche Art von Schaden könnte verursacht werden?

- b) Fällt das System unter gesetzliche Rahmenbedingungen wie die Maschinenrichtlinie, das Medizinproduktegesetz oder der EU-AI-Act? Wenn ja, wie wurden diese Gesetze im Entwicklungsprozess beachtet?

- c) Wurden bereits Haftungsrichtlinien für den Roboter, dessen Design und mögliche negative Konsequenzen formuliert? Wenn ja, wer ist haftbar?

Frage 6: Mensch-Maschine Interaktion und geteilte Entscheidungen

- a) In welchem Kontext wird der Roboter angewendet werden und welche Menschen werden den Roboter benutzen?

- b) Wurden die Nutzer*innen in die Entwicklung miteinbezogen (z.B. Stakeholder-Analyse, participatory design)?



- c) Gibt es im System einen ‚Human-in-the-loop‘? Können automatische Entscheidungen korrigiert werden? Wenn ja, von wem?

- d) Inwiefern treffen Mensch und Roboter gemeinsam Entscheidungen? (human-in-the-loop, human-on-the-loop)

- e) Inwiefern beeinflusst der Roboter Handlungen der User*innen?

Frage 7 Privatheit & Zustimmung

- a) Inwiefern erlaubt das System von Teilnehmenden Einblicke in ihre Entscheidungen, Verhalten und Beziehungen, etwa durch Logdaten, Sensordaten, und deren Verarbeitung?

- b) Werden vulnerable Gruppen nach AGG den Roboter verwenden oder mit ihm interagieren? Wenn ja: Wie wird sichergestellt, dass keine systematischen Verzerrungen vorliegen?



- c) Welche sensiblen Rückschlüsse sind grundsätzlich aus den gesammelten und verarbeitenden Daten möglich?

- d) Ist für den Einsatz absehbar, dass Daten von Menschen gesammelt, die nicht beteiligt sind? (z.B. Passant*innen)

- e) Wurden die Teilnehmenden/Mitarbeitenden schriftlich darüber informiert, dass Daten von ihnen gesammelt werden und haben sie explizit zugestimmt, etwa bei Fortbildungen oder Informationsveranstaltungen?