KI-Ethik

Fragestellungen Datenethik-Board für *High-Risk KI-Systeme* zum Grundsatz «Fairness»

aktualisiert Dezember 2025

swisscom

Ethische Prinzipien als Basis für die Entscheide des Datenethik-Boards

Die sechs Prinzipien konkretisieren das Ethik-Framework von Swisscom im Bereich Daten und KI

Fairness

Nutzen und Mehrwert

Wir nutzen digitale Ressourcen und Daten so, dass Nutzen und Mehrwert für die Kunden und Gesellschaft entstehen. Wir wenden unseren Werte-Rahmen uneingeschränkt bei der Nutzung digitaler Technologien an.

Transparenz

Wir ermöglichen unseren Kunden und der Öffentlichkeit, unseren Einsatz digitaler Technologien und die Verarbeitung von Daten sowie die Risiken der wichtigsten Anwendungsfälle zu verstehen und nachzuvollziehen.

.

Verantwortlichkeit / Rechenschaft

Wir übernehmen die volle Verantwortung für den Einsatz digitaler Technologien und die Verarbeitung von Daten sowie deren Ergebnisse oder Folgen durch Swisscom oder in unserem Auftrag tätige Dritte.

Keine Diskriminierung

Wir verhindern Diskriminierungen und Verzerrungen (Bias). Wir sorgen dafür, dass niemand wegen besonderen Eigenschaften wie Geschlecht oder Hautfarbe etc. benachteiligt oder diffamiert wird.

Informationelle Selbstbestimmung

Wir schützen unsere Kundschaft und die Öffentlichkeit vor einer unbegrenzten Verarbeitung ihrer Daten. Wir geben Betroffenen die Möglichkeit, die Verarbeitung ihrer Daten selbst zu bestimmen, ohne dass ihnen dadurch Nachteile im Verhältnis zu Swisscom entstehen.

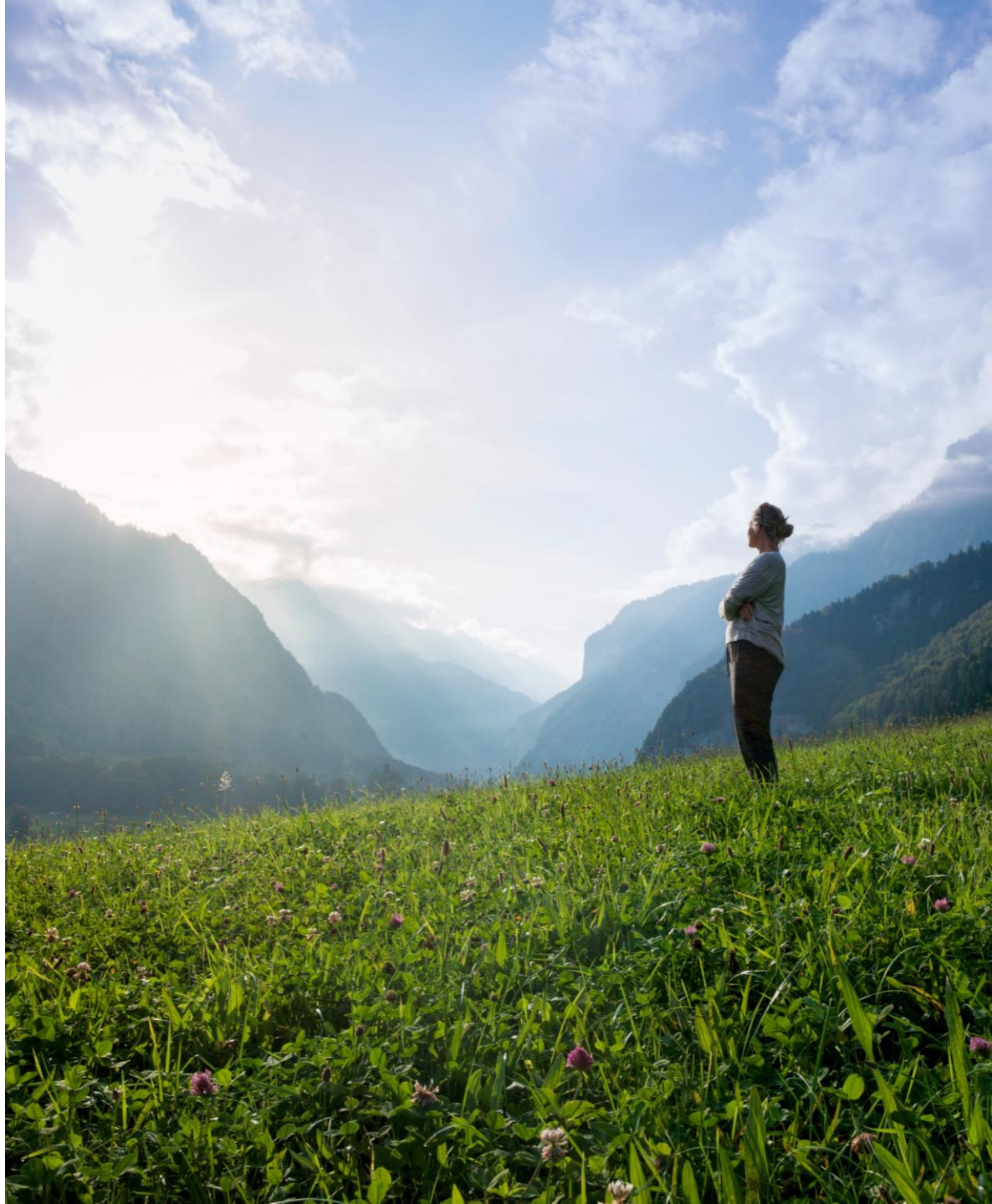
Achtung der Persönlichkeit

Wir achten die Persönlichkeit und Privatsphäre der Menschen. Wir vermeiden die Beeinträchtigung oder Schädigung der körperlichen sowie psychischen Integrität. Wir respektieren das Recht am eigenen Bild und an der eigenen Stimme.

Fairness

Nutzen & Mehrwert
Wir nutzen digitale Ressourcen und Daten so, dass Nutzen und Mehrwert für die Kunden und Gesellschaft entstehen. Wir wenden unseren Werte-Rahmen uneingeschränkt bei der Nutzung digitaler Technologien an.

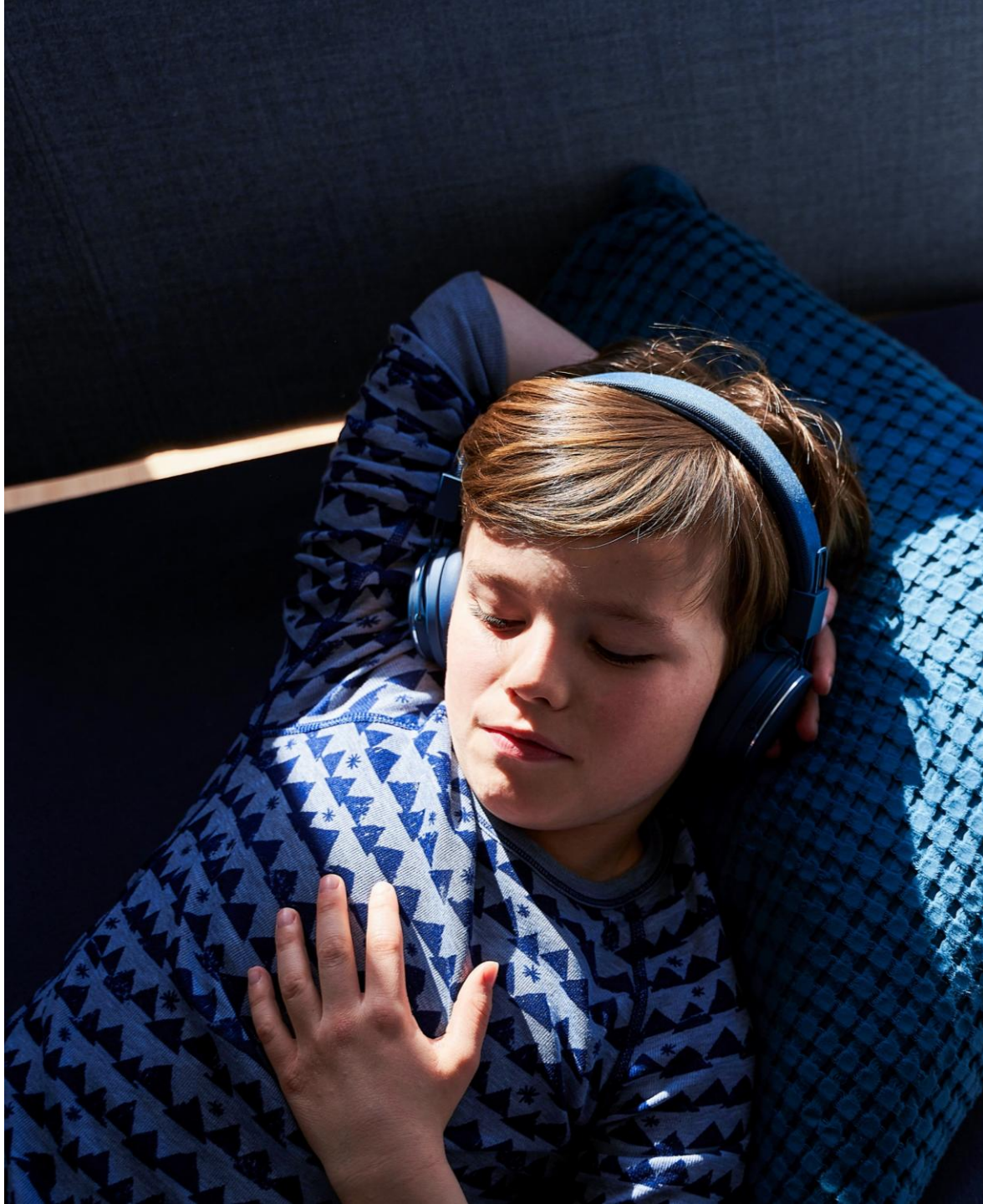
1. Was ist der Verwendungszweck des eingesetzten KI-Systems bzw. welches Problem wird damit gelöst?
2. Welchen Nutzen und Mehrwert schafft das KI-System für Mitarbeitende, Kunden und Gesellschaft?
3. Welche Überlegungen zu Umweltbelastungen (Energie- und Wasserverbrauch) des KI-Systems haben stattgefunden?



Fairness

Keine Diskriminierung
Wir verhindern Diskriminierungen und Verzerrungen (Bias). Wir sorgen dafür, dass niemand wegen besonderen Eigenschaften wie Geschlecht oder Hautfarbe etc. benachteiligt oder diffamiert wird.

1. Wie wurde das KI-System (vom Hersteller oder von Swisscom) auf mögliche Verzerrungen/Bias* der Vorhersagen analysiert?
2. Wie ist sichergestellt, dass das KI-System keine Echokammern schafft und keine bestehenden Stereotypen verstärkt?
3. Welche Massnahmen wurden (vom Hersteller oder von Swisscom) ergriffen, um Modellverzerrungen* vorzubeugen?
4. Welche Trainingsdaten wurden oder werden (vom Hersteller oder von Swisscom) für das Modell angewendet?
5. Welche Massnahmen wurden (vom Hersteller oder von Swisscom) ergriffen, um sicherzustellen, dass diese Daten fair und gerecht (bias-free) sind? Falls das KI-System eingekauft wurde: Gibt es zusätzliche Massnahmen von Seiten Swisscom?
6. Welche Daten produziert das KI-System?



* mögliche Gründe für Bias siehe nächste Seite

Mögliche Gründe für Bias aufgrund der Daten für das Trainieren, Validieren und Testen des KI-Systems

[Quelle: Florent Thouvenin, Stephanie Volz; WHITE PAPER Diskriminierung, Juni 2024]

Historical Bias: Es wurden Daten verwendet, welche auf veralteten, diskriminierenden Realitäten beruhen. Gesellschaftliche Veränderungen wurden möglicherweise vernachlässigt, wodurch Probleme der Vergangenheit reproduziert werden.

Representation Bias: Es wurde ein unausgewogener Datensatz verwendet, der nicht alle Personengruppen gleichermassen berücksichtigt. Dies kann z.B. sein, weil für gewisse Personengruppen weniger Daten verfügbar sind als für andere.

Label: Die verwendeten Daten wurden unausgewogen gekennzeichnet, z.B. wurden Personendaten mit geschützten Merkmalen negativ gelabelt.

Rückkoppelung: Bei selbstlernenden Systemen kann es zu einer problematischen Rückkoppelung («feedback loop») kommen, wenn die Ergebnisse des KI-Systems als Teil einer neuen Datenbasis für das weitere Training des KI-Systems verwendet werden. Dies kann eine Diskriminierung aufrecht erhalten oder sogar verstärken.

Transfer Context Bias: Das KI-System wird in einem falschen Kontext verwendet. Wurde der Algorithmus z.B. zur Beurteilung einer bestimmten Personengruppe entwickelt, kann er bei der Anwendung auf eine andere Personengruppe zu ungenauen oder falschen Ergebnissen führen.

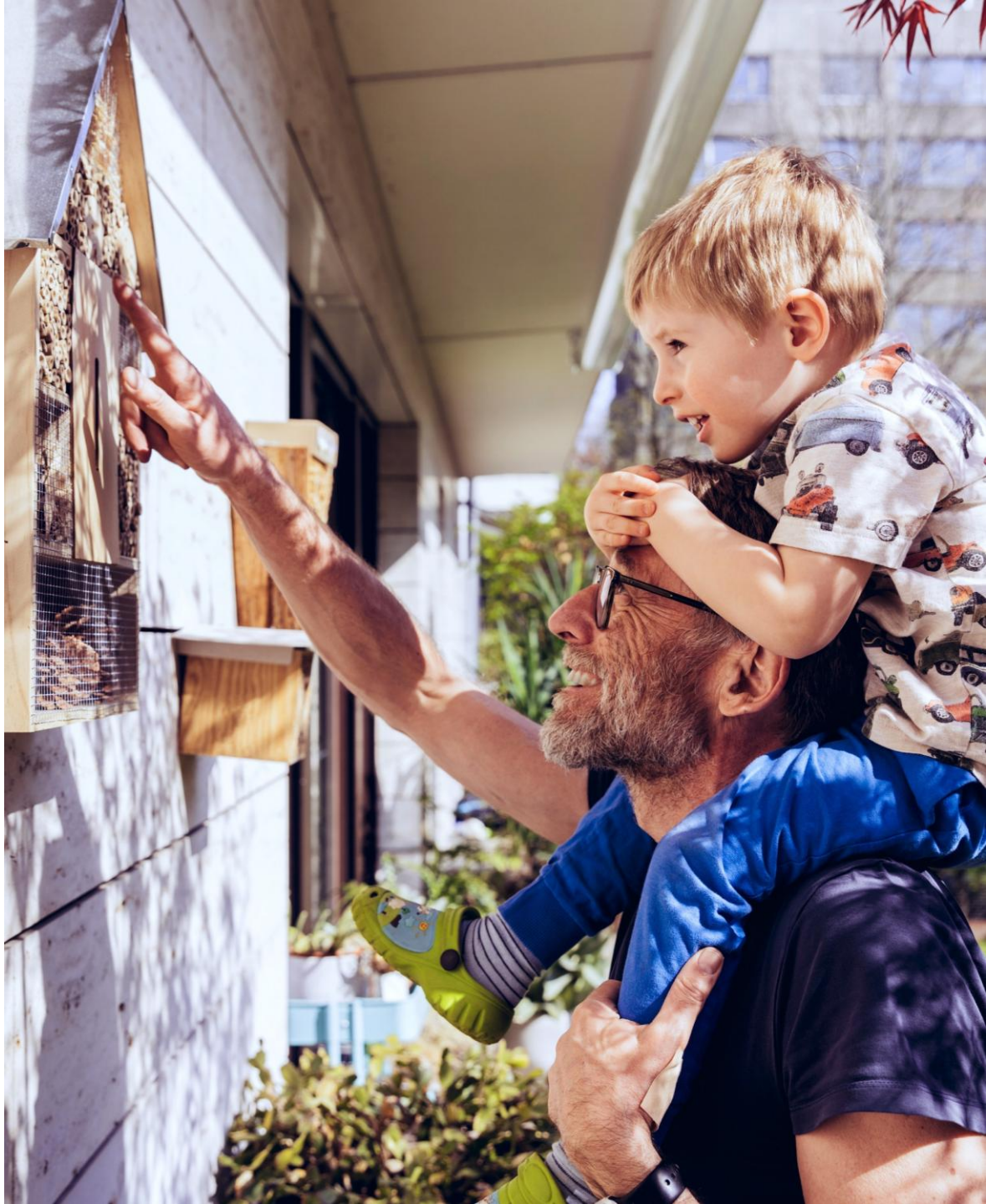
Aggregation Bias: Es werden Rückschlüsse aus einem Einheitsmodell auf Personengruppen gezogen, die unterschiedlich betrachtet werden sollten.

Automation Bias: Den Resultaten des KI-Systems wird zu viel Vertrauen entgegengebracht, weil die Qualität der Entscheidungen überschätzt wird. Z.B. werden Entscheidungen des KI-Systems übernommen, obwohl dieses nur Vorschläge unterbreitet («decision support system»).

Fairness

Achtung der Persönlichkeit
Wir achten die Persönlichkeit und Privatsphäre der Menschen. Wir vermeiden die Beeinträchtigung oder Schädigung der körperlichen sowie psychischen Integrität. Wir respektieren das Recht am eigenen Bild und an der eigenen Stimme.

1. Wie ist die menschliche Kontrolle über die Ergebnisse des KI-Systems und seiner Nutzung vorgesehen?
2. Welche Massnahmen gibt es, die den Benutzer*innen die Interpretation der Ergebnisse (relevante Bestimmungsfaktoren) erleichtern?
3. Welche Auswirkungen hat das KI-System auf Menschen im Rahmen der beabsichtigten Nutzung und welche Auswirkungen könnte es im schlechtesten Fall haben?
4. Welche Auswirkungen kann das KI-System auf Menschen haben, wenn es manipulativ** oder falsch verwendet wird?
5. Wurden Vorkehrungen getroffen, um ein blindes Vertrauen der Nutzer*innen in das KI-System zu verhindern (z.B. kritische Betrachtung einer Leistungsbeurteilung)?



** Typische Beispiele von Manipulation liefern sogenannte «Dark Patterns», welche Nutzer*innen zu Handlungen verleiten sollen, die ihren eigenen Interessen entgegenstehen. Um das zu erreichen, werden Anwender*innen getäuscht oder anderweitig massgeblich in ihrer Fähigkeit beeinträchtigt, freie Entscheidungen zu treffen. Dark Patterns sind allerdings nicht ein spezifisches Problem von KI-Systemen.

Fairness

Informationelle Selbstbestimmung
Wir schützen unsere Kundschaft und die Öffentlichkeit vor einer unbegrenzten Verarbeitung ihrer Daten. Wir geben Betroffenen die Möglichkeit, die Verarbeitung ihrer Daten selbst zu bestimmen, ohne dass ihnen dadurch Nachteile im Verhältnis zu Swisscom entstehen.

1. Besteht eine Opt-out Möglichkeit für die Anwender*innen, die nicht mit dem KI-System interagieren wollen? Welche (möglichst gleichwertige) Alternative wird den Anwender*innen geboten?
2. Können die Anwender*innen selbst bestimmen, ob und falls ja, wie ihre Daten verwendet werden (z.B. für das Training von KI)? Wie benutzerfreundlich sind diese Möglichkeiten? Können die gewählten Präferenzen resp. die Zustimmung auch später noch angepasst resp. widerrufen werden?
3. Inwiefern kann das KI-System manipulativ** sein (z.B. um mehr Daten als notwendig zu sammeln)?

** Typische Beispiele von Manipulation liefern sogenannte «Dark Patterns», welche Nutzer*innen zu Handlungen verleiten sollen, die ihren eigenen Interessen entgegenstehen. Um das zu erreichen, werden Anwender*innen getäuscht oder anderweitig massgeblich in ihrer Fähigkeit beeinträchtigt, freie Entscheidungen zu treffen. Dark Patterns sind allerdings nicht ein spezifisches Problem von KI-Systemen.

Fairness

Transparenz
Wir ermöglichen unseren Kunden und der Öffentlichkeit, unseren Einsatz digitaler Technologien und die Verarbeitung von Daten sowie die Risiken der wichtigsten Anwendungsfälle zu verstehen und nachzuvollziehen.

1. Wie einfach ist für die (End)Nutzer*innen ersichtlich und verständlich (z.B. angemessene und verständliche Information), dass sie mit einem KI-System interagieren?
2. Frage für Anbieter*innen: Inwiefern sind die für den Betreiber/Deployer zur Verfügung gestellten Informationen (Betriebsanleitung) zugänglich und verständlich?
3. Frage für Betreiber*innen: Inwiefern werden den (End)Nutzer*innen adressatengerechten Informationen zur Verfügung gestellt?
4. Wie ist sichergestellt, dass die Transparenzanforderungen regelmässig überprüft werden?

Fairness

Verantwortlichkeit & Rechenschaft
Wir übernehmen die volle Verantwortung für den Einsatz digitaler Technologien und die Verarbeitung von Daten sowie deren Ergebnisse oder Folgen durch Swisscom oder in unserem Auftrag tätige Dritte.

1. Welches Missbrauchspotenzial besteht beim verwendeten KI-System, wenn dieses ausserhalb des bestimmten Zwecks angewendet wird?
2. Wie wurde das Risiko bewertet, dass das KI-System durch sorgfältig konstruierte Eingaben in die Irre geführt werden kann und somit ungewollte Ergebnisse erzielt?
3. Wohin können Fragen und Kommentare von (End)Nutzer*innen zum KI-System gesendet werden?
4. Welche Unterweisung oder Schulung ist beim Einsatz des KI-Systems für wen (z.B. Endnutzer*innen) vorgesehen (KI-Kompetenz)?
5. Welche geeigneten Prozesse und Mechanismen wurden (vom Hersteller oder Swisscom) eingerichtet, die im Falle des Auftretens von Schäden oder nachteiligen Auswirkungen greifen (Incident-Prozess, Ersatzlösung, Kommunikation, Wiedergutmachung)?
6. Inwiefern wird (vom Hersteller oder von Swisscom) eine regelmässige Überprüfung des KI-Systems auf Nutzen, Funktion, Schadensvermeidung etc. sichergestellt?