VERBAND

NACHVOLLZIEHBARKEIT VON ALGORITHMEN.

1 Ausgangslage

Der zunehmende Einsatz von Algorithmen in allen möglichen Lebensbereichen wirft eine Reihe von Fragen und Forderungen in Gesellschaft und Politik auf. Einerseits führt er auf gesellschaftlicher Ebene zu einer gewissen Skepsis, weil nicht klar ist, welche Informationen wie und zu welchem Zweck genutzt werden. Es besteht ein Informations- und Machtgefälle zwischen Anbietenden und den Einzelnen, das wiederum Misstrauen erwecken kann. Andererseits hat der Fortschritt bei der Entwicklung von Künstlicher Intelligenz (KI) die Politik insbesondere auf EU-Ebene aktiviert, welche nun eine hohe Dichte an technologiespezifischen Regulierungen anstrebt.

Aus Sicht von Swico und der ICT-Industrie besteht auf gesellschaftlicher und politischer Ebene ein Interesse daran, das **Vertrauen** in die Digitalisierung und deren Anwendung nachhaltig zu stärken. Ein Element ist die **Transparenz** über den Einsatz und die Nachvollziehbarkeit von Algorithmen für Kundinnen und Kunden bzw. deren Endkunden. Dieses Merkblatt dient als Hilfsmittel, um den Einsatz und die Funktionsweise von Algorithmen zu erklären und damit für Ihre Kunden und Endnutzerinnen nachvollziehbar zu machen.

Die Schaffung von Transparenz ist ein wesentlicher Beitrag zu einer **ethischen** Digitalisierung. Der Zusammenhang zwischen Transparenz gegenüber unterschiedlichen Zielgruppen und dem ethischen Umgang mit Daten wird auch in der Swico Charta für den ethischen Umgang mit Daten beschrieben.

Nachfolgend werden die Schritte aufgeführt, welche Swico für eine bessere Transparenz und damit die Nachvollziehbarkeit von Algorithmen für Ihre Kundschaft und deren Endkundinnen und -Kunden empfiehlt (S. 2). Das Merkblatt wird mit einigen Instrumenten (S. 3) sowie drei Beispielen (S. 4) ergänzt.

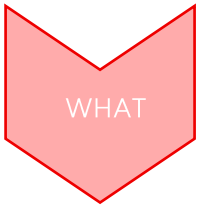
Swico
Lagerstrasse 33
CH-8004 Zürich

Tel. +41 44 446 90 90
www.swico.ch
info@swico.ch

2 Elemente für die Schaffung von Transparenz von Algorithmen

2.1 Adressaten

Definieren Sie, **wem gegenüber** Sie Informationen zum Algorithmus übermitteln wollen. Berücksichtigen Sie als ICT-Anbieter die gesamte Wertschöpfungskette bis hin zu m End-User. Algorithmische Entscheide, die Endkundinnen und -Kunden betreffen, sollten auch für diese nachvollziehbar sein.

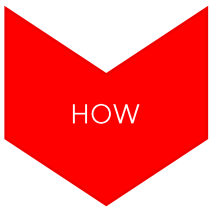


2.2 Deklaration

Schaffen Sie Transparenz, indem Sie deklarieren, **dass** ein Entscheid auf einem automatischen Algorithmus oder einer KI-Applikation beruht. Die End-User sollten wissen, ob ein sie betreffender Entscheid von einem Algorithmus oder einem Menschen getroffen wird. Idealerweise ist dies in internen Richtlinien vorgegeben.

2.3 Begründung

Erklären Sie, **weshalb** Sie den Algorithmus einsetzen, resp. **wozu** er dient und worin der **Nutzen** für die Kunden/Konsumenteninnen/Gesellschaft besteht.



2.4 Funktionsweise

Erklären Sie, auf welcher Grundlage der Algorithmus Entscheide fällt und **wie er funktioniert**.
Dieser Punkt kann besonders anspruchsvoll sein, weshalb wir ihn in einem separaten Kapitel weiter unten detaillierter behandeln.

2.5 Korrekturmöglichkeit

Geben Sie Ihren Kundinnen und Kunden sowie den End-Usern die Möglichkeit, auf einen fehlerhaften (bspw. diskriminierenden) algorithmischen Entscheid hinzuweisen. Sehen Sie Massnahmen vor, um fehlerhafte Entscheide zu korrigieren.

2.6 Ethische Grundsätze

Führen Sie aus, an welchen ethischen Grundsätzen Sie sich orientieren, resp. ob Sie sich einem spezifischen Kodex anschliessen. An dieser Stelle verweisen wir auf die Swico Charta zum ethischen Umgang mit Daten.

2.7 Funktionsweise – how does it work

Ein spezielles Augenmerk setzen wir auf die Funktionsweise – das Wie – des Algorithmus: Oft ist die vollständige Offenlegung des Algorithmus nicht vereinbar mit den Geschäftsinteressen. Meist kann jedoch zumindest erklärt werden, wie er grundsätzlich funktioniert und **auf welcher Grundlage er Entscheide fällt**. Wichtig ist, dass Anwendende und Betroffene die Folgen ihres Handelns einschätzen können.

Folgende Fragen sollten dabei beantwortet und transparent gemacht werden können:

- Wie funktioniert der Algorithmus grundsätzlich?
- Wie resp. aufgrund welcher Daten wurde der Algorithmus trainiert?
- Welche wissenschaftlichen Modelle und Methoden verwendet der Algorithmus?
- Was ist der Zusammenhang zwischen dem Input und der Vorhersage? Warum kommt der Algorithmus zu einem spezifischen Entscheid?
- Wie lernt das KI-System?
- Auf welche Informationen greift das System zu? Welche Daten kann es verarbeiten und wie?

Bei besonders komplexen Modellen wie neuronalen Netzen (Deep Learning) ist es für Menschen nicht mehr nachvollziehbar, wie ein Ergebnis zustande kommt – der Algorithmus ist nicht direkt erklärbar. Deshalb wird auch von einer **Black Box** gesprochen. In vielen Bereichen wie Sprach-, Text- und Bilderkennung werden neuronale Netze bereits eingesetzt, weil sie eine besonders leistungsstarke KI-Methode sind. In diesen Fällen werden Hypothesen oder Annäherungen gebildet, um diese komplexe Modelle zu erklären.

Mögliche Lösungsansätze

- **Visualisierung der Datengrundlage**:
  Dank der visuellen Abbildung der Daten wird erkennbar, ob die gesamte Datengrundlage ausreichend und ausgewogen oder umgekehrt mangelhaft und einseitig ist. Damit können allfällige Vorurteile (Bias) vermieden werden.
- **Visualisierung eines neuronalen Netzes**:
  Sie bildet ab, welche Pixel oder Pixelgruppe mutmasslich zu einer Vorhersage geführt hat. Ein bekanntes Beispiel ist die Klassifizierung von *Wolf und Husky*. Die Untersuchung hat gezeigt, dass der Hintergrund (Schnee) und nicht das Tier zu einer Klassifizierung geführt hat, was eine Korrektur der Gewichtung von Bildteilen erfordert. S. auch https://arxiv.org/pdf/1602.04938.pdf
- **Beipackzettel zu einem KI-System**:
  Die öffentlich verfügbaren Model Cards von Google sind ein gutes Beispiel, wie alle relevanten Informationen zu einem bestimmten Algorithmus aufgeführt werden. Dazu gehören Herkunft, vorgeschlagene Verwendungszwecke, Einschränkungen sowie eine ethische Bewertung. IBM schlägt AI Fact Sheets vor, wobei zu den Fakten insbesondere Zweck und Risikoeinschätzung des Modells zählen, ebenso wie die Charakteristika des Datensets, Modells oder Services oder ethische Assessments, die während der Entwicklung durchgeführt wurden Eine Website liefert Hintergrundinformationen und Beispiele solcher Factsheets: AI FactSheets 360 (mybluemix.net) Auch das Institute for Responsible AI ein solches Instrument zur Verfügung.
- **Open Source Toolkits**:
  IBM offeriert das AI Explainability 360 Toolkit, das State-of-the Art Algorithmen und Metriken bietet, die die Erklärbarkeit und Interpretierbarkeit von ML Modellen fördern:
  Introducing AI Explainability 360 | IBM Research Blog
- **Der Algorithmus erklärt sich selbst**:
  DARPA, eine Forschungsagentur für das US-Militär, verfolgt die Idee, dass der Algorithmus die Begründung seiner Entscheidung selber erstellt. So liefert der Algorithmus nicht nur den Entscheid (z.B. Kredit verweigert oder Kredit bewilligt), sondern auch die Erklärung für den Entscheid. Solche Ansätze befinden sich noch im Forschungsstadium. Die Herausforderung liegt neben der technischen Machbarkeit in einer für Menschen nachvollziehbaren, verständlichen Information.

2.8 Beispiele

Anhand existierender Use Cases sollen die Elemente für eine bessere Nachvollziehbarkeit veranschaulicht werden.

2.8.1 Einsatz von KI für die Personalrekrutierung

Eingegangene Online-Bewerbungen werden für die Prüfung durch die Rekrutierinnen und Rekrutierer vorselektiert.

Folgende Informationen führen zu einem besseren Verständnis:

- **Adressaten**
  vier Ansprechgruppen: Endbetroffene (= Kandidatinnen und Kandidaten), Anwendende (= Rekrutiererinnen und Rekrutierer; Personen in Entscheidungsfunktionen (jene, die entscheiden, das System zu kaufen und in der Firma einzusetzen); Technikerinnen und Techniker. Informationen zum Einsatz und zur Funktionsweise müssen bis zu den Kandidatinnen und Kandidaten gelangen.
- **Zeitpunkt** der Transparenz:
  - Für alle Adressaten: bereits beim Absenden der Bewerbung, nicht erst bei der Verkündung des Resultats,
  - Für Personen in Entscheidungsfunktion: Für diejenigen, das das KI-System kaufen, wird vor der Kaufentscheidung verständlich erklärt, was das System leisten kann und wo die Grenzen sind.
  - Für Anwendende: Vor der Inbetriebnahme wird verständlich erklärt, wie man das System korrekt benutzt, inwieweit man sich auf die Empfehlungen verlassen kann und was schiefgehen könnte (insbesondere bei Falschbenutzung).
- **Autonomiegrad** des Algorithmus: Der Algorithmus fällt keinen definitiven Entscheid, sondern bietet eine Empfehlung für die Rekrutiererin bzw. den Rekrutierer.
- **Kausalität:** Zusammenhang zwischen Ursache und Wirkung, zwischen Bewerbung und Auswahl.
- **Weshalb** wird ein Algorithmus eingesetzt? Das können Effizienzgründe sein, bei sehr zahlreichen Bewerbungen, oder aber auch Bestrebungen, menschliche Vorurteile auszublenden oder zu korrigieren.

2.8.2 Personalisierte Werbung aufgrund von Videobildern

Bilder von 3D Kameras werden aufgrund festgelegter Kriterien analysiert und zu anonymisierten Metadaten zu Werbezwecken genutzt. Folgende Informationen führen zu einer höheren Akzeptanz:

- **Datensparksamkeit**: Festhalten, dass das System nicht dazu befähigt ist, einzelne Menschen zu identifizieren. Das Interesse des Systems gilt dem Bild, nicht dem Einzelnen.
- **Datennutzung**: Welche Information aus dem Bild wird tatsächlich genutzt? Etwa Geschlecht, Alter, Hautton, Haltung etc.
- **Weshalb** wird ein Algorithmus eingesetzt? Nicht, um die Daten zu speichern, sondern um relevante Werbung anbieten zu können.

2.8.3 Automatische Erkennung von verdächtigen Versicherungsforderungen

Eine Krankenkasse analysiert eingegangene Versicherungsforderung und weist sie bei Verdacht einer Sachbearbeiterin bzw. einem Dachbearbeiter zu. Folgende Informationen schaffen Transparenz:

- **Adressaten**: Mögliche Ansprechgruppen sind die Sachbearbeitende und End-User (= Versicherte)
- **Datennutzung**: Um systematische Betrugsfälle, etwa durch Ärztinnen oder Ärzte, aufzuspüren, werden Muster analysiert; es werden nicht die Informationen einzelner Versicherungsnehmenden beurteilt. Die Offenlegung der Datengrundlage dient der Vermeidung von Diskriminierungen oder verzerrten Ergebnissen.
- **Autonomiegrad**: Der Algorithmus erkennt gewisse Muster, fällt jedoch keine Entscheidungen. Verdachtsfälle werden einem Menschen zur Beurteilung zugewiesen.
- **Funktionsweise** des Algorithmus: in diesem komplexen Case wird ein Erklärungsansatz im SHAP gesucht. Dabei handelt es sich um eine nachträgliche Annäherung, die den Output eines Algorithmus erklären soll.

– **Weshalb** wird ein Algorithmus eingesetzt? Hier geht es um den Schutz von rechtmässigen Versicherungsansprüchen und die Verhinderung von Versicherungsbetrug, was wiederum den Versicherten resp. Endkunden dient.

Für Rückfragen:

Judith Bellaiche
judith.bellaiche@swico.ch
SWICO