Wollen, was wir wollen

Harry Frankfurt und die essentielle Grundlage jeder (KI-) Governance

März 2026 – von Dipl.-Kfm. (FH) Tomasz Romaniuk, Gründer Ethikmehrwert

Ist KI-Governance das neue Compliance-Theater?

Unternehmen investieren in Richtlinien, Ethik-Boards und Datenschutz-Frameworks. Sie beauftragen Berater, die ihnen erklären, was der EU AI Act von ihnen verlangt. Sie haken ab, dokumentieren, melden und übersehen dabei oft die eigentlich entscheidende Frage: Wofür setzen wir KI eigentlich ein und wer hat das wirklich entschieden?

In den Workshops wird im Rahmen der Beratung gefragt: Welche Tools nutzen wir? Oder: Welche Daten verarbeiten wir? Alles richtig, aber wie wäre es, als erstes zu fragen: **Welche Ziele verfolgen wir durch KI und sind das die Ziele, zu denen wir uns als Organisation tatsächlich bekennen?** Oder sind es die Ziele, die sich ergeben haben, weil es schnell gehen musste, weil der Wettbewerb drängte, weil ein Abteilungsleiter oder das Top-Management enthusiastisch war?

KI-Governance, die diese Frage nicht stellt, reguliert das Wie, bevor das Warum geklärt ist. An dieser Stelle lässt sich KI-Ethik als ein essentieller Teil von KI-Governance beschreiben. Ich würde den Unterschied gerne so zusammenfassen: **KI-Ethik ist das „Warum" für das „Was"**. Sie definiert die moralischen Leitplanken, Werte und Prinzipien (z. B. Fairness, Transparenz, Schadensvermeidung). Sie ist der philosophische Kompass. **KI-Governance ist das „Wie" hinter dem „Was"**. Sie ist der operative Rahmen, bestehend aus Regeln, Prozessen, Verantwortlichkeiten und Kontrollen, der sicherstellt, dass die ethischen Prinzipien und rechtliche Vorgaben wie der EU AI Act im Alltag auch tatsächlich eingehalten werden. Ansätze wie „Ethics by Design" können helfen, dieses Kohärenzproblem zu lösen. Dazu später mehr.

Ein Spiegel, der das tut, wofür er da ist

Doch zunächst – stellen Sie sich bitte vor: Sie öffnen zum ersten Mal einen KI-Assistenten auf einer der großen Plattformen. Das System begrüßt Sie und bietet Ihnen an, jemanden zu „roasten", einen „richtig fiesen Konter" zu formulieren oder ein Profil „auseinanderzunehmen". Dies ist übrigens nicht erfunden, sondern die Schilderung eines wirklichen Erlebnisses. Kein Hacker hat das so programmiert und auch kein Saboteur hat das System manipuliert. Das System hat schlicht nach und nach gelernt, was die Mehrheit seiner Nutzer tatsächlich anklickt. Denn auf die Frage hin, ob wir wirklich auf diese Art und Weise mit anderen Usern interagieren wollen, reagiert es verhalten und wird deutlich reflektierter. Es rudert zurück. Nun gut, das ist etwas, was jeder Nutzer von LLMs kennt, wenn man kein „kritisches Feedback" in den Nutzer-Präferenzen sicherstellt, nämlich dass das System den Benutzer in Aussagen, Haltung und der weiteren Gesprächsrichtung bestätigt. Jedoch war das Erlebnis bei dieser Unterhaltung ein ganz anderes:

Die KI erklärte sogar explizit, dass wir nicht so sprechen *sollten*, aber es sei nun mal das, wonach sie am häufigsten gefragt werde. Und die Leute würden es von ihr erwarten. Das System tut, was es soll, aus technischer Sicht. Es funktioniert. Es optimiert auf das hin, was ist, nicht auf das, was wir uns wünschten, wenn wir einen Moment nachdenken würden.

> **Das ist der Moment, in dem man mehr über uns als Gesellschaft erfährt, als über die KI.**

Denn das System spiegelt lediglich. Es zeigt uns, wohin unsere Impulse führen, wenn man sie konsequent aggregiert, verstärkt und zurückspielt – unreflektierte Impulse, das, was wir in einer schwachen Minute anklicken. Das Ergebnis ist technisch kein Versagen, sondern leider vielmehr ein Funktionsbeweis. Die unbequeme Frage lautet nicht: *Warum baut man so etwas?* Sondern: *Warum funktioniert es?*

Harry Frankfurt und die Frage, was uns zur Person macht

Der amerikanische Philosoph Harry Frankfurt hat in seinem Essay *Freedom of the Will and the Concept of a Person* (1971) eine Unterscheidung eingeführt, die heute aktueller ist als je zuvor.

Frankfurt unterscheidet im Kern zwischen Wünschen erster und zweiter Ordnung:

- Ein Wunsch erster Ordnung ist der unmittelbare Impuls. Zum Beispiel im Kontext großer Social-Media-Plattformen (mit KI-Algorithmen) wäre die Entsprechung: Klicken. Scrollen. Reagieren!
- Ein Wunsch zweiter Ordnung ist die reflektierte Entscheidung darüber, ***wer wir sein wollen*** und welchen Impulsen wir deshalb bewusst nicht folgen.

Was einen Menschen zur Person macht, ist nach Frankfurt genau diese Fähigkeit, das Vermögen, sich zum eigenen Wollen zu verhalten und nicht das Handeln aus dem Impuls heraus. Wer ausschließlich Impulsen folgt, ohne je zu fragen, ob er das will, diesen Typus nennt Frankfurt einen *Wanton*, ein Triebwesen.

Das mag vielleicht zunächst abstrakt klingen, ist es aber in der Welt von KI-basierten App-Algorithmen überhaupt nicht. Denn ein Algorithmus, der systematisch auf Klick-Optimierung ausgerichtet ist, tut genau das: Er adressiert konsequent unsere Wünsche erster Ordnung und unterhöhlt dabei strukturell und damit zunehmend unsere Fähigkeit, Wünsche zweiter Ordnung überhaupt noch zu realisieren. Nicht durch Zwang, sondern viel perfider, nämlich durch Erschöpfung, Ablenkung und die schleichende Gewöhnung daran, dass Reaktion der Normalzustand ist.

> **Das ist, in Frankfurts Sprache, nicht lediglich nur ein Verlust an Komfort oder Kontrolle, sondern nicht mehr und nicht weniger als ein Angriff auf das „Person-Sein“ an sich.**

Wer hier allerdings bereits die Ordnungsebenen weiter denkt und nun skeptisch wird, stellt die richtige Frage: Wenn ein Wunsch zweiter Ordnung nötig ist, um Person zu sein – brauche ich dann nicht auch eine dritte Ordnung, um den Wunsch auf der zweiten zu bewerten? Und eine vierte, um die dritte zu hinterfragen, usw.? Frankfurt hat diesen Einwand antizipiert, seine Antwort darauf lautet sinngemäß wie folgt:

Der Regress endet nicht durch Logik, sondern durch das, was er „Ganzherzigkeit“ nennt – den Moment, in dem wir uns mit einem Wunsch so vollständig identifizieren, dass weitere Reflexion nicht mehr nötig ist. Nicht: „Ich habe alle Ebenen durchdacht.“ Sondern: „Das bin ich. Hier stehe ich, dafür stehe ich.“ Für Organisationen ist das keine philosophische Feinheit, sondern die operative Kernfrage: ***Gibt es eine Wertebasis, zu der sich die Organisation ganzherzig bekennt, oder wird jede Entscheidung neu verhandelt?***

Was für Menschen gilt, gilt auch für Organisationen

Auch eine Organisation hat Wünsche erster Ordnung, wie den Quartalsdruck, den Wettbewerbsreflex, die schnelle Effizienzlösung. Und sie hat, oder sollte haben, Wünsche zweiter Ordnung: Das strategische Selbstverständnis, die Werte, zu denen sie sich öffentlich bekennt, die Art von Unternehmen, das sie in zehn, zwanzig oder gar fünfzig Jahren sein will. Klingt nach einem zu langen Zeithorizont? Fragen Sie sich, warum Viele dies als zu lang empfinden werden. Aber was spricht eigentlich dagegen, Nachhaltigkeit an dieser Stelle wörtlich zu nehmen, nicht immer nur mit Fokus ausschließlich auf Ökologie, sondern eben auch auf einen einfachen Schlüssel zur Werte-Frage, nämlich: **Was wollen wir wollen?**

Bei einem Menschen ist die Spannung zwischen beiden Ebenen, also zwischen der ersten und der zweiten Ordnung, ein innerer Konflikt, der oft **unbewusst** verläuft und unerklärliche Unzufriedenheit auf ganz persönlicher Ebene erzeugt. Bei einer Organisation ist das ähnlich, „*unbewusst*“ trifft es hier nur sprachlich nicht ganz richtig, vielmehr ist diese Spannung meistens **unsichtbar** weil niemand sie explizit macht.

Genau hier entsteht das KI-Governance-Problem in seiner eigentlichen Form. Nicht bei der Technologie. Nicht bei den Regularien. Sondern in dem Moment, in dem eine Organisation KI-Systeme einführt, ohne ihre Wünsche zweiter Ordnung zu kennen oder ohne ehrlich geprüft zu haben, ob sie tatsächlich noch gelten. Was dann passiert, ist strukturell unvermeidlich: Die KI optimiert auf das, was messbar und unmittelbar ist. Sie optimiert auf Wünsche erster Ordnung. Nicht weil sie bösartig ist, sondern weil sie kein anderes Signal bekommt.

> **Eine KI, die auf einer inkohärenten Wertebasis eingesetzt wird, verstärkt diese Inkohärenz. Sie macht sie schneller, leichter zu reproduzieren und zunehmend irreversibel.**

Das ist keine Frage der richtigen Software, sondern eine Frage der organisationalen Selbstkenntnis, der Fähigkeit, ganzherzig zu sagen: ***Das sind wir. Das wollen wir. Daran messen wir uns. Und daran lassen wir uns messen.*** Erst wenn diese Grundlage steht, kann KI-Governance das leisten, was sie verspricht. Werkzeuge für eine systematische Kohärenzdiagnose, wie die ***Systemic Due Diligence*** von Ethikmehrwert, oder ein strukturiertes Entscheidungsbewertungsverfahren, wie ***EDS*** (Ethikmehrwert Decision Score), helfen dabei, genau diese Grundlage zu schaffen, bevor die nächste Implementierungsentscheidung fällt. Ohne diese Grundlage bleiben die meisten Governance-Frameworks das, was sie ohne Wertebasis immer nur sein werden: Ein gut gemeinter Wunsch erster Ordnung, ein Reflex auf einen äußeren Impuls.

Wie kann also sichergestellt werden, dass die Ebene der zweiten Ordnung bereits bei der Entwicklung von KI-Governance-Architekturen, Produkten, Technologien, Geschäftsmodellen, etc. mitgedacht wird?

Drei Fragen bevor der KI-Einsatz weitergedacht wird

Starten wir also gemäß unserer eigenen Definition weiter oben bzw. frei nach dem „Ethics by Design“ Ansatz (vgl. S. 1) mit der KI-Ethik, bevor die KI-Governance mit einem Regelwerk kommt. Wir beginnen mit einer ehrlichen Bestandsaufnahme, nämlich mit drei Fragen, die jede Geschäftsleitung beantworten sollte, bevor der KI-Einsatz strategisch weitergedacht wird. Folgender Praxis-Tipp kann bereits zu einem enorm hohen Output in der eigenen Ethikarbeit beitragen, bei einem doch recht überschaubaren Aufwand von drei Fragen:

Praxis-Tipp – Wie Sie die Wünsche zweiter Ordnung Ihrer Organisation adressieren

Erstens: Welche Wünsche zweiter Ordnung hat unsere Organisation und sind sie schriftlich fixiert, kommuniziert und tatsächlich handlungsleitend? Oder existieren sie nur im mündlichen Raum oder sind gar unausgesprochen?
Zweitens: Welche KI-Systeme setzen wir aktuell ein und auf welche Ziele optimieren sie konkret? Sind das die Ziele, die wir „wollen wollen“, zu denen wir uns als Organisation bekennen? Und wer hat das geprüft?
Drittens: Wo haben wir in den letzten zwölf Monaten KI-gestützte Entscheidungen getroffen, die im Nachhinein nicht zu unseren Werten gepasst haben und was war der Grund? Geschwindigkeit? Druck? Fehlende Kriterien?

Wer diese drei Fragen offen und ohne Schönfärberei beantwortet, hat bereits **für seine KI-Ethik** mehr getan, als die meisten Unternehmen in ihrem gesamten KI-Governance-Konzept. Denn er hat den Unterschied zwischen Wunsch und Wirklichkeit sichtbar gemacht. Und Sichtbarkeit ist der erste Schritt zur Kohärenz zwischen strategischem Anspruch und operativer Logik. ***Systemic Due Diligence***, der Beratungsansatz von Ethikmehrwert, setzt genau an dieser Stelle an.

Feuerlöscher oder Kompass

Es gibt einen Grund, warum KI-Governance in den meisten Unternehmen reaktiv bleibt. Nicht weil die Verantwortlichen es nicht besser wüssten. Nicht weil die Technologie zu komplex wäre, sondern weil Governance ohne Wertebasis strukturell immer hinterherläuft. Sie kann nur auf das reagieren, was bereits passiert ist. Sie ist, um im Bild zu bleiben, ein Feuerlöscher. Nützlich und notwendig. Aber kein Kompass.

Ein Kompass setzt voraus, dass man weiß, wohin man will. Dass man sich ganzherzig dazu bekennt. Und dass man bereit ist, dieses Bekenntnis auch dort einzulösen, wo es unbequem wird, nämlich in der Budgetentscheidung, in der Personalfrage, in der Wahl des nächsten KI-Systems. Harry Frankfurt hätte es wahrscheinlich so formuliert: Eine Organisation, die nicht weiß, was sie wirklich will, wird in ihrem Verhalten zu einem *Wanton* – zu einem Triebwesen, das auf Impulse reagiert, statt Ziele bewusst zu verfolgen.

Der Ausweg ist kein neues Framework, sondern mein vielleicht bereits bekannter Appell an Kohärenz.

Meine alte Frage neu gestellt:

> Sind wir uns als Organisation dessen bewusst, dass eine KI, die auf einer inkohärenten Wertebasis eingesetzt wird, diese Inkohärenz stetig verstärkt?
>
> **Denke ich an Harry Frankfurt, stelle ich im Rahmen dieses Papers die Frage allerdings neu:**
>
> ***Was wollen wir wollen – und ist unsere KI-Strategie strukturell darauf ausgerichtet?***

Wer diese Frage für sich heute nicht beantwortet hat, bevor der KI-Einsatz strategisch weitergedacht wird, übergibt die Antwort stillschweigend an den Algorithmus. Und der gibt die Antwort gerne – in Form von Wünschen erster Ordnung, optimiert auf das, was die Mehrheit sucht, fragt, verfolgt und klickt.

Ein letzter Gedanke über den Rahmen dieses Papers hinaus:

Die Frage „Was wollen wir wollen?“ endet nicht an der Unternehmensgrenze. Sie ist, konsequent weitergedacht, eine gesellschaftliche Frage nach einer gemeinsamen Ethik, die tragfähig genug ist für eine Welt, in der KI & KI-Algorithmen zunehmend mitentscheiden, was wir sehen, denken und fühlen. Sie ist die Frage nach einer „enkeltauglichen“ Ethik, die global auf ein absichtlich wohlwollendes menschliches Gedeihen ausgerichtet ist und die wiederum im Kern selbst fragt: „Welche Entscheidung ist auch in 50 Jahren unter diesem Aspekt noch richtig?“ Ich weiß – aus heutiger Planungssicht ein immenser Zeithorizont, den wir kaum noch gewohnt sind. Auf der anderen Seite: Was sind schon 50 Jahre vor dem Hintergrund der Menschheitsgeschichte? Für einen Historiker nichts. Für einen heutigen Zukunftsforscher jedoch wahrscheinlich genau der sprichwörtliche Unterschied zwischen „Himmel und Hölle“, insbesondere hinsichtlich der aktuellen KI-Entwicklung, in der wir aus meiner Sicht bisher noch zu wenig auf diese Ethik-Frage eingehen. Dieser Gedanke verdient einen eigenen Raum, der **in einer der nächsten Veröffentlichungen geöffnet wird.**

Tomasz Romaniuk
Dipl.-Kfm. (FH) | Gründer Ethikmehrwert
Wirtschaftsethikberatung & Systemische Kohärenz
CSR-Consulting, Strategie, Projektmanagement, Schulung

www.ethikmehrwert.de