Abstract Momentum Kongress 2023: Human in Command

Prof. Dr. Doris Aschenbrenner

Aktuell wird der EU AI Act1 in Brüssel beraten. Dieses Diskussionspapier möchte die Diskussion zu den zu erwartenden Vorgaben dieser politischen Rahmenverhandlungen umgesetzt werden können. So fordert die europäische Ebene (Commission, 2021) im Draft für den EU AI Act explizit „Human agency and oversight“ und spezifiziert: “These shall include measures enabling users to understand, monitor, interpret, assess and intervene in relevant aspects of the operation of the AI system.” Analog fordert die Deutsche Normungsroadmap Künstliche Intelligenz an verschiedenen Stellen (DIN, 2022) den Menschen als Teil des Systems in allen Phasen des KI-Lebenszyklus zu begreifen (vgl. u.a. Empfehlung 3, Bedarf 5-19, 5-20).

Dies startet die in Forschungskreisen bereits rege geführte Diskussion (z.B. unter dem Stichwort „Human in Command“ oder auch „Meaningful Human Control“) über Kontrollierbarkeit von Anwendungen mit künstlicher Intelligenz und das Machtverhältnis zwischen immer ausgereifteren Algorithmen und Maschinen und uns Menschen auf der anderen Seite.

Gerade bei der Einführung komplexer Systeme, die sich ggf. im Verlauf von neuen Daten-Input (zumindest teilweise) weiterentwickeln können, stellen sich einige Grundfragen, die sich grob in das Spannungsfeld „Wer hat die Kontrolle: Mensch oder Maschine?“ einordnen lassen. Es sei vorangestellt, dass es i) keine eindeutige Definition von Künstlicher Intelligenz gibt (vgl. (Legg & Hutter, 2007)), ii) auf absehbare Zeit keine dem Menschen überlegene Künstliche Intelligenz („starke KI“) entwickelt werden wird und dass iii) auch andere „klassische“ Algorithmen die oben genannte Frage aufwerfen, ohne KI-Methoden einzusetzen – vgl. dazu ausführlich (Aschenbrenner, 2021). Konkrete Spannungsszenarien werden in der öffentlichen Debatte diskutiert: Hier gibt es zum Beispiel die automatisierte Auswahl von Bewerbungsunterlagen (s.o. Fall Amazon, vgl. (Pumhösel, 2020). Zwar ist die automatisierte Vorsortierung der Bewerbungsunterlagen das gewünschte Verhalten der Algorithmen, eine unvorsichtige Auswahl der Datenlage oder des Lernverfahrens kann aber hier zu einem „falschen“ Ergebnis mit einem sogenannten Bias führen.

Die Frage der Verortung der Kontrolle wird nicht zum ersten Mal gestellt. So entzündete sich diese Diskussion insbesondere bei der Einführung von Flugassistenzsystemen: Wann soll der Mensch die Kontrolle haben, wann die Maschine? Dazu hat (Fitts, 1951) den „HABA-MABA“-Ansatz (Humans are better at, Machines are better at) beschrieben, in dem Aufgaben generell für besser für eine der beiden Parteien geeignet eingeordnet wurden. Für einen hervoragenden Überblick der darauffolgenden Diskussion im Kontext der Diskussion über autonome Flugasssistenzssysteme siehe (Inagaki, 2003). Jüngere Arbeiten (vgl. zum Beispiel (Bradshaw, et al. 2012)) kritisieren diesen Ansatz hauptsächlich weil es zumindest eine zeitliche oder örtliche Dimension der Beantwortung mit einbezogen werden muss, also fragen nicht nur nach: Wer macht was? Sondern eher nach Wer macht was wann bzw. wo. Die Diskussion bewegt sich aktuell eher in die Richtung der sogenannten „hybrid intelligence“ (Dellermann et al., 2021) die eine Verschmelzung der beiden Parteien andeutet. Zentral ist dieser Ansatz auch im „Operator 4.0“ Konzept (Romero, 2020), in dem die Erweiterung der menschlichen Fähigkeiten durch Einsatz von technologischen Hilfsmitteln (also in Kontext klassischer Assistenzsysteme) diskutiert wird.

1 Draft standardisation request to the European Standardisation Organisations in support of safe and trusworthy artificial intelligence https://ec.europa.eu/docsroom/documents/52376?locale=en

Wie kann man sich „die Kontrolle am besten teilen“? Zentral ist an dieser Stelle die Wahl geeigneter Autonomiestufen zwischen „komplett manuell“ und „komplett automatisch“ sowie eine dezidierte Übergabe der Kontrolle bei Unsicherheiten oder Gefahren. Aktuelle Lösungen dieses Dilemmas bieten bspw. definierte Autonomiestufen im autonomen Fahren (eine ausführliche Kritik dazu in (Inagaki & Sheridan, 2019) sowie Forschung zu „Handover“-Strategien, wenn beispielsweise der Mensch das Steuer wieder übernehmen muss (wie im H2020 Projekt „Mediator“ (Grondelle, 2020)) sowie die Einführung ähnlicher Autonomiestufen für Künstliche Intelligenz in der industriellen Fertigung (Plattform "Industrie 4.0", 2021). Fazit dieser Entwicklung: Die Frage nach der Kontrolle und Kontrollierbarkeit über automatisierte Systeme bleibt wichtig und muss bei der Neuentwicklung von Systemen grundsätzlich bedacht werden. Dies ist ein zentraler Baustein dafür, dass wir KI-Systeme entwickeln können, in denen Mensch und Maschine produktiv miteinander arbeiten und sich gegenseitig verstehen.

Aschenbrenner, D. (2021). Zukunft der Arbeit in Zeiten von Industrie 4.0 und künstlicher Intelligenz. In I. N. W. Beck (Hrsg.), Theologie und Digitalität - Ein Kompendium (S. 496-516). Herder Verlag.

Bradshaw, J. M., Dignum, V., Jonker, C., & Sierhuis, M. (2012). Human-agent-robot teamwork. IEEE Intelligent Systems, 27(2), 8-13.

Cavalcante Siebert, L., Lupetti, M.L., Aizenberg, E. et al. (2022). Meaningful human control: actionable properties for AI system development. AI Ethics (2022).

Colloseus, C.; Aschnebrenner (2023, forthcoming) Normung und Standardisierung von KI-Systemen aus soziotechnischer Sicht. Tagungsband Kann ein Algorithmus im Konflikt moralisch kalkulieren? (**KAIMo**) Ethik und digitale Operationalisierung. München (2023).

Commission, E. (2021). Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Brüssel, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=DE.

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2021). The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. arXiv preprint arXiv:2105.03354.

DIN (2022). Deutsche Normungsroadmap Künstliche Intelligenz, https://www.din.de/resource/blob/891106/57b7d46a1d2514a183a6ad2de89782ab/deutsche-normungsroadmap-kuenstliche-intelligenz-ausgabe-2--data.pdf

Fitts, P. (1951). *Human engineering for an effective air-navigation and traffic-control system.* National Research Council, Div. of.

Grondelle, E. e. (2020). HMI Functional Requirements, D1.5 of the H2020 project MEDIATOR. https://mediatorproject.eu/deliverables: Online.

Hirsch-Kreisen, H. (2018). Einleitung: Digitalisierung industrieller Arbeit. In H. Hirsch-Kreinsen, P. Ittermann, & J. Niehaus, Digitalisierung industrieller Arbeit: die Vision Industrie 4.0 und ihre sozialen Herausforderungen (S. 13-32). Frankfurt/Main: Nomos Verlag.

Inagaki, T. (2003). Adaptive automation: Sharing and trading of control. In E. Hollnagel, Handbook of cognitive task design (S. 147–169). Hillsdale, New Jersey: Lawrence Erlbaum Associates Publishers.

Inagaki, T., & Sheridan, T. B. (2019). A critique of the SAE conditional driving automation definition, and analyses of options for improvement. Cognition, technology & work, S. 569-578.

Ittermann, P., & Niehaus, J. (2018). Industrie 4.0 und Wandel von Industriearbeit – revisited. Forschungsstand und Trendbestimmungen. In Hirsch-Kreinsen, I. P. H., & J. Niehaus, Digitalisierung industrieller Arbeit (S. 33-60). Nomos Verlagsgesellschaft mbH & Co. KG.

Legg, S., & Hutter, M. (2007). A collection of definitions of intelligence. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, S. 17-24.

Plattform "Industrie 4.0". (2021). *Technologieszenario "Künstliche Intelligenz in der Industrie 4.0".* https://www.plattform-i40.de/IP/Redaktion/DE/Downloads/Publikation/KI-industrie-40.html: Online Plattform Industrie 4.0.

Pumhösel, A. (15. 03 2020). Gender-Bias: Schlechtere Jobchancen für Frauen durch Algorithmen. *Der Standard*, https://www.derstandard.de/story/2000115720676/gender-bias-schlechtere-job-chancen-fuer-frauen-durch-algorithmen.

Romero, D., Stahre, J., & Taisch, M. (2020). The Operator 4.0: Towards socially sustainable factories of the future. Computers & Industrial Engineering, 139, 106128.