

Martina Philippi

Reflexionsschulden begleichen: Überlegungen zur kritischen Nutzung von KI

In den wissenschaftlichen und gesellschaftlichen Debatten um künstliche Intelligenz (KI) hat Vertrauenswürdigkeit den Status einer zentralen ethischen Anforderung. Mit guten Gründen wird argumentiert, dass das Erreichen dieses Zieles nicht allein auf technischem Weg möglich ist, sondern auch einer verantwortungsvollen Gestaltung und Regulierung des Einsatzes von KI in konkreten Anwendungsfeldern bedarf. Vertrauenswürdige KI wird demnach so verstanden, dass sie die Autonomie der Nutzenden erhält und unterstützt; von diesen wiederum ist im Zuge eines verantwortungsvollen Einsatzes von KI eine kritische Haltung gefordert.

Unter anderem im Kontext der erklärbaren KI — einem Ansatz, der die Nachvollziehbarkeit KI-generierter Empfehlungen und Einschätzungen gewährleisten und damit die menschliche Entscheidungsautonomie unterstützen soll, dabei die soziale Praxis des Erklärens einbezieht (Rohlfing et al. 2021) und auch im Kontext politischer Entscheidungen nutzbar gemacht werden könnte (Amara-singhe et al. 2023; Papadakis et al. 2024) — wird diese kritische Haltung als angemessenes Vertrauen adressiert. Spezielle Debatten diskutieren Konzepte wie ‚kalibriertes‘ oder ‚evaluatives‘ Vertrauen (Miller 2023) oder auch ‚gesundes Misstrauen‘ (Peters/Visser 2023), um eine passende Beschreibung für diese Haltung zu finden und auf diese Weise zu klären, welche Erwartungen einerseits an Technik, andererseits an ihre Nutzer*innen gestellt werden sollten und können.

Doch noch basaler ist die Frage, unter welchen Bedingungen eine solche Haltung überhaupt eingenommen werden kann. Ich möchte in meinem Beitrag darstellen, dass die Forderung nach einer kritischen Haltung sowohl im Umgang mit Alltagstechnologien als auch in spezialisierten Arbeitswelten unter Zeit- und Effizienzdruck nicht trivial ist und einige Ambivalenzen birgt. Zentral ist hier die Überlegung, dass ein reibungsloser Gebrauch von digitalen Technologien lange Zeit als selbsterklärend und *reflexionsfrei* beworben wurde und wir gewissermaßen Reflexionsschulden aufgebaut haben, die im Umgang mit KI dringend begleichen werden müssen. Anschließend möchte ich darüber nachdenken, welche

Form des Umgangs mit KI uns helfen könnte, diesen Ambivalenzen verantwortungsvoll zu begegnen.¹

1. „Don’t make me think“: Lob der Reibungslosigkeit

„Don’t make me think“ ist der Titel eines Klassikers aus dem Bereich Web Usability, also der Methodik des nutzerfreundlichen Webdesigns (Krug 2000). Dieses Motto bedeutet zunächst nur, dass eine Webseite selbsterklärend und übersichtlich sein soll, sich die Nutzenden also auf Inhalte konzentrieren können und nicht unnötig suchen und probieren müssen. Doch in dem Motto klingt eine implizite Norm des digitalen Technikdesigns an: nämlich, dass eine wünschenswert reibungslose Nutzung eine reflexionsfreie ist.

Tatsächlich macht es den reibungslosen Umgang mit einer Technologie aus, dass sie in ihrem Gebrauch *unthematisch* ist, in dem Sinne, dass man im Gebrauch nicht darüber nachdenkt. Man richtet die Aufmerksamkeit auf das *Werkstück*, nicht auf das *Werkzeug*. Jedoch besteht ein Unterschied zwischen „reflexionsfrei, weil im Gebrauch unthematisch“ und „reflexionsfrei aufgrund von *overtrust*“. Dieser Unterschied wird im Moment des Gebrauchs nicht sichtbar, wohl aber in der dahinterstehenden Haltung: nämlich einem dann fehlenden Problembewusstsein, also der Bereitschaft, mögliche Risiken und Nebenfolgen zu reflektieren. Denn man vertraut eben nicht nur der Technologie, sondern auch ihrem Anbieter, ihrem Betreiber, der damit verbundenen Dateninfrastruktur, der sozialen Praxis („machen alle“), dem Überwiegen des Nutzens. Dadurch entstehen Risiken, die man *eigentlich* in gesonderter Reflexion adressieren müsste, und diese Reflexion müsste über eigens eingerichtete Routinen und Infrastrukturen in die Praxis eingebettet werden.

Digitale Werkzeuge haben mit klassischen Werkzeugen (etwa aus dem Werkzeugkasten des traditionellen Handwerks: Hammer, Zange et cetera) gemeinsam, dass sie uns etwas erleichtern; doch sie unterscheiden sich von ihnen insofern, dass sie diese ganz bestimmte Art von Reflexionskosten erzeugen, wenn wir sie verantwortungsvoll und letztlich auch zu unserem langfristigen Vorteil nutzen wollen. Das ist zunächst kein unlösbares Problem, insofern man diese Reflexionsschritte leisten und gegebenenfalls in Routinen einbauen kann: etwa Wachsamkeit bezüglich Datenschutz oder Beeinflussung durch kommerzielle oder

1 Der Beitrag entstand im Kontext des SFB/TRR 318 „Erklärbarkeit konstruieren“, Teilprojekt B06 WEthik und Normativität der erklärbaren KI (DFG: Projektnummer 438445824).

politische Interessen; oder auch die Praxis, vor-digitale Fähigkeiten anderweitig zu trainieren, um Kompetenzverlust vorzubeugen. Solche Maßnahmen finden jedoch häufig nicht statt, und dies liegt, so meine These, daran, dass unsere Nutzung digitaler Technologien seit Beginn der Digitalisierung unter dem Einfluss eines kommerziell vorangetriebenen Narrativs steht: dass reibungslose Technik das Leben *einfacher* macht, und „einfach“ im Sinne von „sorglos“, „effizient“, „bequem“ und „schnell“. Bei fortschreitender Digitalisierung ist *scheinbar weniger* Reflexion erforderlich, obwohl sich die Reflexionsanforderungen durch Risiken und unerwünschte Nebenfolgen eigentlich erhöhen. Ich möchte dies anhand von zwei technikhistorischen Entwicklungstendenzen der letzten Jahrzehnte darstellen, die zugleich Gestaltungsziele darstellen, und zwar die zunehmende Niedrigschwelligkeit und die Unauffälligkeit der Nutzung digitaler Technologien als *Designziele*.

2. Niedrigschwelligkeit und Unauffälligkeit digitaler Techniknutzung als Designziele

In den vergangenen Jahrzehnten ist die Nutzung digitaler Technologien in verschiedenen Hinsichten niedrigschwelliger geworden, nämlich erschwinglicher, handlicher und in der Bedienung intuitiver: Rechner sind keine raumfüllenden Riesenmaschinen mehr, zu deren Bedienung man Programmierkenntnisse benötigt, sondern schreibtischkompatible, erschwingliche Geräte mit zunehmend intuitiven Eingabemethoden. Meilensteine dieser Entwicklung sind der *Personal Computer*, die Maus als handliches Eingabegerät und die Entwicklung zunehmend selbsterklärender Benutzeroberflächen: Während zu Beginn noch mit entsprechenden Vorkenntnissen zeilenweise Code eingegeben werden musste, stand bald ein Textmenü mit anwählbaren Optionen zur Verfügung, die dann in klickbare Symbole überführt wurden. Eingabegeräte wie Tastatur und Maus wurden um Touchscreens und Sprachsteuerung ergänzt und teilweise abgelöst. Sprachsteuerung wiederum bewegt die Menüführung grundlegend in die Richtung einer kommunikationsähnlichen Mensch-Maschine-Interaktion, die von Chatbots im Schriftlichen wieder aufgegriffen wird und so eine zunehmende Anthropomorphisierung des digitalen Gegenübers fördert.

Zugleich findet eine Miniaturisierung von Technologien bis hin zu jenem äußerst leistungsfähigen, multifunktionalen und sozial bedeutsamen Taschencomputer statt, den das Smartphone heute für uns darstellt (Merchant 2017), das uns als Teil einer persönlichen Infrastruktur in Gestalt eines transportablen

Alltagsgegenstandes neben Schlüssel und Portemonnaie ständig begleitet (und diese teilweise sogar ersetzt). Digitale Technologien sind immer dabei, immer an, sie müssen nicht mehr extra angeschaltet und hochgefahren werden, und ich muss mich als Nutzerin längst nicht mehr in einen Expertenmodus begeben, in dem ich Befehle beherrschen und Ausgaben fachkundig deuten muss, um sie verstehen zu können (auch wenn sich Eingabekompetenzen qua *Prompting* gezielt optimieren lassen). Bedienungserleichterung und Miniaturisierung können in ihrer Effizienz als *Designziele* verstanden werden, also als *beabsichtigte* technische Umsetzung eines Nutzungstrends. Von der daraus resultierenden breiten Verfügbarkeit und Zugänglichkeit lässt sich ein weiteres Designziel heute nicht trennen: die Unauffälligkeit digitaler Techniknutzung.

Dass die Einbettung von Technologien in routinehafte Praxen zu einem Unauffälligwerden der Technologien selbst führt, wurde unter anderem in der philosophischen Phänomenologie ausführlich thematisiert und von Hans Blumenberg 1963 pointiert formuliert: „Ungleich vollkommener als durch das Mimikry der Gehäuse wird das Technische als solches unsichtbar, wenn es der Lebenswelt implantiert ist“ (Blumenberg 2015, S. 190). Technikhistorisch besonders interessant ist, dass sich dieses Phänomen zu einem Designprinzip avanciert hat, das Marc Weiser 1991 in der Zeitschrift ‚The Scientific American‘ als *ubiquitous computing* formuliert hat (Weiser 1991; vgl. Wiegerling 2024). Gemäß diesem Designprinzip ist es nicht nur ein Nebeneffekt, sondern beabsichtigt, dass die Nutzung von EDV in einem weiten Sinne — nicht nur als Desktop PCs, sondern vor allem in anderen Formen von Computern — sowohl unauffälliger, als auch verbreiteter wird, bis hin zur Omnipräsenz. Zentral ist hierbei zweierlei: die unmerkliche, passive Einspeisung von Informationen in Systeme über Sensoren und automatisierte Datenverarbeitung einerseits und die hochautomatisierte Vernetzung der Geräte andererseits, so dass auch der Datenfluss effizient und unmerklich (das heißt prinzipiell kontrollierbar, aber in der Regel ohne bewusste Aufsicht) stattfindet. Smart Devices und das *Internet of Things* sind eine Vision aus diesem Kontext, die etwa in Gestalt von *Smart Homes* inzwischen realisierte und in den Alltag eingebettete Technologien sind.

Diese beiden Designprinzipien erzeugen — sowohl durch die Erleichterung und Unmerklichkeit des Technikgebrauchs als auch durch die sie begleitenden, teils kommerziellen Narrative — eine weitreichende Ambivalenz in der digitalen Techniknutzung: Digitale Technologien sind aufgrund der automatisierten Datensammlung und -verarbeitung mit der Kombination aus Sensoren, Vernetzung und zunehmend auch KI sogar ohne aktiven Aufwand zu nutzen. Zugleich hat diese hochautomatisierte, sensor- und KI-gestützte Datenerfassung, -verarbei-

tung und -speicherung immer weitreichendere Folgen, und darunter auch unerwünschte: Sie wirken sich möglicherweise zu meinem Nachteil als Nutzerin aus, befinden sich aber in der Regel außerhalb meiner Wahrnehmung und zu weiten Teilen sogar meiner Kontrolle.

3. Das Problem der Reflexionsschulden

Zwar haben wir Gestaltungsmöglichkeiten, die auch eine verantwortungsvolle *routinemäßige* Praxis erlauben, in der die Aufmerksamkeit auf etwas anderes gerichtet wird (nicht auf das Werkzeug, sondern auf das Werkstück). Doch was diese Ambivalenzen so brisant macht, ist die Verbindung von erfolgreich umgesetzten Designzielen, in denen Visionen ohne eine Reflexion auf ihre möglichen negativen Folgen und Begleiterscheinungen eingebettet waren, und wirkmächtigen Narrativen. Narrative besitzen gesellschaftliche Deutungsmacht und können hierdurch wirtschaftliche Interessen durch scheinbare Reflexion legitimieren (Daub 2020); eine kritische Haltung erscheint dann im Extremfall als technikskeptisch, fortschrittsfeindlich und unnötig. Verantwortungsvolle Techniknutzung bedeutet auch, dem wirtschaftlich aufrechterhaltenen Narrativ der wünschenswerten Reflexionslosigkeit zu widerstehen. Denn dieses Narrativ ließ eine gesellschaftlich getragene Reflexion lange unnötig erscheinen — bis zum Auftreten der Nebenfolgen von Social Media, von „Filterblasen“ (Pariser 2012), von übermäßigem Medienkonsum und Aufmerksamkeitsverlusten (Newport 2016) über die kommerzielle Ausbeutung persönlicher Daten (Zuboff 2018) bis hin zu Desinformation (Mafi-Gudarzi 2019) und Kompetenzverlusten in der ‚Wissensarbeit‘ (Lee et al. 2025).

Auch wenn diese Themen jetzt intensiv adressiert werden, konnten über die Jahre digitale Nutzungsgewohnheiten und Angebote entstehen und wachsen, die jetzt zwar reguliert und mit Vorsichtsmaßnahmen versehen werden, die jedoch zusätzlichen Aufwand bedeuten (zum Beispiel Ablehnung von Cookies, bewusster *Digital Detox*). Hinzu kommt, dass die Nutzung digitaler Technologien durch zunehmende Reibungslosigkeit und Komfort sowie wirtschaftlichen und administrativen Zielen dienende Effizienz in vielen Kontexten ein Standard geworden ist, der laut Roberto Simanowski zum „Verschwinden“ der Computer aus der Gesellschaft (Simanowski 2018) führt. Dies spitzt das technikphilosophische und medienwissenschaftliche Grundphänomen des Unsichtbarwerdens von Technologien in ihrem Gebrauch zu: Die Technik wird als solche nicht mehr nur in ihrer jeweiligen Nutzung nicht mehr gesehen, sondern *gar nicht mehr* — einfach, weil die situative Nutzung eine permanente geworden ist. Die

Technologien verschwinden nicht aus der Praxis, sondern aus dem gesellschaftlichen Diskurs, der gesellschaftlichen Aufmerksamkeit. Anders ausgedrückt: Eine Reflexionspause — also ein *situatives* Nichthinterfragen — geht in ein generelles Nicht-mehr-Befragen sowohl der Technologie als auch ihrer Nutzungspraxis über. Die für einen verantwortungsvollen Umgang erforderliche Reflexion ist in einen Reflexionsraum verschoben worden, der, plakativ gesprochen, nahezu gleichzeitig wegrationalisiert wird; eine niedrigschwellige Benutzung, ‚Selbstverständlichkeit‘ der Technologien im Alltag (Gewöhnung, Allgegenwart und Unmerklichkeit) und Effizienzanforderungen (Zeit- und Leistungsdruck in Alltag und Arbeitswelt) führen zu einer Art pragmatischem *Overtrust*, der durchaus neben einer reflektierteren Haltung („*Eigentlich* sollte ich mir den KI-Output kritischer ansehen... / Cookies ablehnen / auf Social Media vorsichtiger sein...“) bestehen kann, die man sich aber nicht leisten kann oder möchte. Analog zu einer Formulierung von Hartmut Rosa, der bezüglich geplanter, aber aus Zeitmangel nicht mehr erledigbarer Aufgaben von „Zeitschulden“ spricht (Rosa 2019), kann man hier von ‚Reflexionsschulden‘ sprechen. In einem Extremszenario müssten wir Reflexionsinsolvenz anmelden — eine verantwortungsvolle und kontrollierte Techniknutzung wäre dann nicht mehr möglich. Dies ist, wie gesagt, ein Extremszenario. Irgendjemand wird immer Fragen stellen. Die Frage ist aber, ob diese Fragen gesellschaftlich gehört und aufgegriffen werden, ob es noch einen gesellschaftlichen Reflexionsraum für sie geben wird. Denn angesichts fortschreitender Entwicklungen ist eine ein für alle Mal funktionierende verantwortungsvolle Technikgestaltung nicht realistisch; vielmehr muss ein gesellschaftlicher Reflexionsraum offengehalten werden, auch über den aktuellen „KI-Hype“, also die sehr intensive Thematisierung der KI-Nutzung, hinaus. Denn gerade dieser Reflexionsraum ist erforderlich für eine gesellschaftliche Verständigung über eine kritische Nutzung digitaler Technologien. Und während sich bei Social Media oder Datenschutz allmählich immer stärker herauskristallisiert, was das bedeuten kann, ist nicht nur die eingangs erwähnte akademische Reflexion, was eine kritische Haltung im Umgang mit KI sein soll, noch im Fluss. Das hängt auch mit einer epistemischen Eigenheit von KI-generierten Informationen und ihrer Einschätzung zusammen.

4. Epistemische Ambivalenzen von KI auf Inhalts- und Metaebene

Die Forderung nach einem kritischen Gebrauch digitaler Technologien ist also schon ohne KI in der Praxis nicht unproblematisch, denn wir sind sogar an

die Gewöhnungseffekte gewöhnt. ‚Kritisch‘ mit KI umzugehen, würde erfordern, Nutzungsgewohnheiten aufzugeben beziehungsweise umzugestalten und zweitens *noch zusätzlich* dem epistemischen Problem von KI-Output Rechnung zu tragen: die Ergebnisse mit Vorsicht zu genießen. Denn gemessen am Maßstab des traditionellen Werkzeugs, das einen bestimmten Zweck und unsere diesbezüglichen Erwartungen erfüllt, kann KI als ein unzuverlässiges, ambivalentes Werkzeug erscheinen. Die Ambivalenz besteht darin, dass sich in vielen Fällen nicht sicher bestimmen lässt, ob die Ausgabe korrekt ist. Wie kann ein unzuverlässiges Werkzeug Nutzen bringen? Wie ist die Information „kritisch“ zu verwenden, insbesondere unter Zeit-, Effizienz- und Wettbewerbsdruck — immer unter Vorbehalt? Und wäre ein solcher Vorbehalt nicht gerade das Auslagern von Reflexionsbedarf, dessen Klärung dann unter Umständen nie wirklich stattfindet? Auch die Forderung nach Vertrauenswürdigkeit mit Aspekten wie Autonomieerhalt, Verantwortbarkeit oder Transparenz löst das epistemische Problem nicht vollständig, da dann die Lösung vornehmlich auf der technischen Seite verortet würde; die Gestaltung vertrauenswürdiger KI nach diesen Aspekten kann die *Möglichkeit* einer verantwortungsvollen Nutzung gewährleisten, jedoch nicht ihre Realisierung. Insbesondere in professionellen Kontexten birgt dies das Risiko von Vertrauensverlust; zugleich ist KI trotz dieser Spannung aus mangelnder Zuverlässigkeit und Transparenz einerseits und praktischen Erfordernissen andererseits schon in zahlreiche alltägliche Praxen wie selbstverständlich eingebettet.

Zusammenfassend gesagt, birgt KI epistemische Risiken auf der inhaltlichen und der Metaebene: Auf der Inhaltsebene besteht das Risiko von Irrtum und unzureichender Prüfung von KI-generierten Einschätzungen und Entscheidungen; auf der Metaebene sind der Verlust von Reflexionskompetenz und -raum möglich. Denn ironischerweise werden ausgerechnet Kompetenzbildung und Reflexionsraum durch den zunehmenden Gebrauch von KI im Sinne eines Kompetenzverlusts gefährdet — so berichtet eine aktuelle Studie von „Auswirkungen von generativer KI auf kritisches Denken“ als ‚Verringerung des kognitiven Aufwands‘ und Auswirkungen auf Selbst- und Technikvertrauen (Lee et al. 2025). In diesem Sinne ist es nicht die einzige und vielleicht nicht einmal die größte Gefahr hinsichtlich menschlicher und maschineller Leistungsfähigkeit, dass KI uns überholen könnte (Grunwald 2018), sondern dass wir durch die zunehmende Delegation von informationsbezogenen Aufgaben hinter unserem eigenen geistigen Potenzial zurückbleiben.

5. Irritierbare Nutzer, befragbare Systeme: Dialogbasierte Mensch-Maschine-Interaktionsansätze als Chance für eine kritische Nutzung von KI

Eine kritische Nutzung von KI erfordert also erstens die *Bereitschaft*, den KI-Output zu hinterfragen, zweitens die *Kompetenz*, das zu tun, und drittens den *Impuls* dazu (entgegen *Overtrust*-Tendenzen wie etwa dem *automation bias*). Insbesondere der dritte Punkt führt zu Reflexionsanforderungen an eine verantwortungsvolle, das heißt auch reflektierte Nutzungspraxis:

- 1) Man muss sich Reflexion leisten können. Insofern hängt eine verantwortungsvolle Technikgestaltung von verantwortungsvoller Regulation ab, zugleich aber auch von Reflexionsraum und Kompetenzbildung.
- 2) Wenn Reflexion für einen kritischen Gebrauch unabdingbar ist, darf sie der intuitiven Nutzungspraxis nicht widersprechen. Das bedeutet: Die Routine der Nutzungspraxis darf dem Bedarf an Kritik nicht zuwiderlaufen; sie muss Kritik ermöglichen, vielleicht sogar integrieren.

Was also ist eine kritische Nutzung, und wie kann sie unter den beschriebenen Voraussetzungen sinnvoll realisiert werden? Mein experimenteller Vorschlag ist nun, den geforderten Reflexionsraum teilweise in die Mensch-Maschine-Interaktion zu verlagern. Das kann geschehen, indem ein derzeit sich ohnehin ausbreitender Trend genutzt wird: die dialogbasierte Interaktion mit KI. Der derzeit meistverbreitete Einsatz, der das Thema KI zugleich in nie dagewesener Intensität in den Fokus gesellschaftlichen Diskurses gebracht hat, ist die nicht unumstrittene Nutzung des OpenAI-Tools ChatGPT. Speziellere, gezielter gestaltbare Nutzungsszenarien finden sich im Kontext der erklärbaren KI (Kovari 2025). Diese Form der Interaktion bietet die Möglichkeit, KI-Output nicht nur zu *hinterfragen*, sondern auch zu *befragen*.

Auch eine dialogbasierte KI-Nutzung birgt Risiken. Zu nennen sind insbesondere Effekte der Anthropomorphisierung, bei denen unsere Zuschreibungen — die KI tritt wie eine Expertin auf, sie scheint „kompetent“ zu sein — auf ein Verhalten von KI treffen, das man als Kompetenz-Mimikry bezeichnen kann. Übersehen wird hierbei der Faktor der menschlichen Erfahrung und Urteilkraft. Dies kann zu einer *Opazitätsverschiebung* führen: Zwar scheint das System transparenter, doch gibt es auf der Userseite Kommunikationserwartungen und auf der Systemseite im Rahmen eines *User Modelling* Vorannahmen über die menschliche Dialogpartnerin. Wenn die Erwartungen an zwischenmenschliche Kommunikation unmodifiziert in die Mensch-Maschine-Interaktion transferiert werden, kann dies zur Folge haben, dass die Möglichkeit der Befragung nicht

hinreichend in Anspruch genommen wird, obwohl Bereitschaft und Fähigkeit eigentlich vorliegen — es fehlt dann aber der *Impuls* zum Fragen.

Diese Risiken gilt es sorgfältig und gesondert zu adressieren. Zugleich jedoch eröffnet eine kollaborative und gegebenenfalls dialogbasierte Interaktion mit KI die Chance, uns aus den zu reibungslos gewordenen Denk- und Nutzungsmustern bezüglich digitaler Technologien zumindest ein Stück weit herauszuholen und zugleich unsere besondere Kompetenz, Fragen zu stellen und Irritation zu formulieren, in die Nutzungspraxis einzubinden. Denn bezüglich der Gewöhnungseffekte muss damit gerechnet werden, dass Reflexion vernachlässigt wird, wenn sich eine komfortable reflexionsfreie Nutzung nach dem gewohnten Werkzeugschema anbietet. Man kann ein Werkzeug im traditionellen Sinne nicht „bewusst“ in einer solchen Ambivalenz nutzen; man kann jedoch andere Routinen etablieren. Mein Vorschlag lautet daher, diese besondere Art von Technik nicht als Werkzeug zu nutzen, bei dem man sich — entgegen dem Design und den intuitiven Nutzungsroutinen — immer wieder bewusst daran erinnern muss, dass es nicht ganz zuverlässig ist; sondern als eine *neuartige* Quelle von Hinweisen, Einschätzungen und sogar Ideen und Problemlösungsansätzen. KI kann uns überraschen, und das soll sie auch; sie fordert unser Denken heraus; wir müssen uns zu ihren Ergebnissen mithilfe unserer Urteilkraft verantwortungsvoll verhalten. In diesem Sinne ist auch eine epistemisch zunächst nicht entscheidbare Ausgabe nicht nutzlos: Sie kann ein Anhaltspunkt sein, etwas zu sehen und näher zu betrachten, was uns vorher nicht aufgefallen ist. Es geht dann nicht darum, „Informationen“ zu erhalten und zu verwerten, sondern KI als epistemische Erweiterung der eigenen Fähigkeiten zu verstehen. KI kann uns helfen zu priorisieren; zu unterscheiden; zu klassifizieren; genau hinzuschauen; in allen diesen Fällen gilt, dass sie unser eigenes Potenzial unterstützen kann — und das ist der entscheidende Punkt: ohne es zu ersetzen oder zu beschränken.

Konkrete Chancen und zugleich Anforderungen bestehen in einem solchen kollaborativen oder hybriden Ansatz erstens darin, Erwartungen zu klären, anhand derer die Qualität des KI-Outputs bewertet und gegebenenfalls durch Training verbessert werden kann, und dies durchaus in Analogie zu menschlichen Rollen (Assistenz, Beratung et cetera); zweitens Grenzen des Systems bewusst wahrzunehmen und anzuerkennen, wie man es auch bei Einschätzungen eines menschlichen Gegenübers, das nie allwissend und unfehlbar sein kann, machen würde; und drittens auf diese Weise sowohl von der menschlichen, als auch der durch KI durchgeführten Form der Ergebnisfindung zu profitieren und dabei ihre Unterschiedlichkeit anzuerkennen, ohne zugleich Deutungshoheit (und somit Autonomie) abzugeben. Wesentliche Anforderung an die nutzende Person

ist, durch den KI-Output *irritierbar* zu sein, denn schließlich ist es gerade der Unterschied zum traditionellen „Werkzeug“, dass KI uns durch ihre besondere Art der Erkenntnisfindung überraschen kann und soll; es gilt zugleich, diese Irritation auszuhalten, also sie nicht in eine der beiden Richtungen epistemisch aufzulösen (im Sinne des *automation bias* — KI hat im Zweifelsfall recht — oder des *Undertrust* — KI ist unzuverlässig), sondern als Ausgangspunkt für weiteres Fragen zu verwenden. Denn Fragen zu stellen, muss nicht dazu führen, sich mit einer Antwort zufriedenzugeben; auf der Basis dieser Antwort können weitere Fragen gestellt werden, und die Kunst ist, zu wissen, wann man zufrieden ist, also „verstanden“ hat *und* nach einer kritischen Prüfung des Outputs überzeugt ist. Denn gemeint ist hier nicht, eigene Recherchen mit Suchmaschinen und unterschiedlichen Informationsquellen durch ChatGPT-Anfragen zu ersetzen, wie es heute schon gängige (obwohl umstrittene) Praxis ist. Die Bedingung für eine kritische Prüfung ist, neben dem Zurateziehen weiterer Quellen, das Vertrauen auf die eigene Urteilskraft ebenso wie eine Selbstverständigung im Vorfeld: Ich muss für eine kritische KI-Nutzung nicht nur Klarheit über die Grenzen des Systems haben, sondern auch über meine eigene Position, meine Erwartungen und meinen Informationsbedarf (wo genau kann KI mir helfen?).

In diesem Sinne ist der Ansatz der kommunikativen Interaktion mit KI und der darauf basierenden kollaborativen Erkenntnisfindung eine klare Chance für Einsätze in disziplinspezifischen Expertendomänen: Mediziner*innen beispielsweise wissen, wonach sie suchen und welches Ergebnis ihnen realistisch erscheint. Dies ist schwieriger für Einsatz zum Beispiel in *evidence-based policies*, wo die Ergebnisse in einem breiteren Beurteilungsrahmen liegen. Doch auch hier kann KI uns helfen, unsere Nutzungsgewohnheiten zu überschreiten, die wir jahrelang auf der Basis des Designparadigmas der wünschenswert reibungslosen und unauffälligen Interaktion entwickelt haben. Wir müssen uns die kritische Nutzung nicht nur zurückerobern, sondern überhaupt erst ausbilden, indem wir lernen, in der Interaktion mit digitalen Technologien und insbesondere KI irritierbar und wach zu sein.

Literatur

- Amarasinghe, K.; Rodolfa, K. T.; Lamba, H.; Ghani, R. (2023): Explainable Machine Learning for Public Policy: Use Cases, Gaps, and Research Directions. In: *Data & Policy* 5(2023): e5. <https://doi.org/10.1017/dap.2023.2>
- Blumenberg, H. (2015): *Schriften zur Technik*. Berlin
- Daub, A. (2020): *Was das Valley Denken nennt*. Berlin

- Grunwald, A. (2018): *Der überlegene Mensch*. München
- Kovari, A. (2025): Explainable AI Chatbots towards XAI ChatGPT: A Review. In: *Heliyon* 11(2): e42077. <https://doi.org/10.1016/j.heliyon.2025.e42077>
- Krug, S. (2000): *Don't Make Me Think. A Common Sense Approach to Web Usability*. Indianapolis
- Lee, H.-P.; Sarkar, A.; Tankelevitch, L.; Drosos, I.; Rintel, S.; Banks, R.; Wilson, N. (2025): The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama, S. 1–22. <https://doi.org/10.1145/3706598.3713778>
- Mafi-Gudarzi, N. (2019): „Desinformation“. In: *Zeitschrift für Rechtspolitik*, 52(3), S. 65–68
- Merchant, B. (2017): *The One Device. The Secret History of the iPhone*. London
- Miller, T. (2023): Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven decision support. arXiv: <https://doi.org/10.48550/ARXIV.2302.12389> [Letzter Aufruf: 25.5.25]
- Newport, C. (2016): *Deep Work. Rules for Success in a Distracted World*. Little
- Papadakis, T.; Christou, I. T.; Ipeksidis, C.; Soldatos, J.; Amicone, A. (2024): „Explainable and Transparent Artificial Intelligence for Public Policymaking“. *Data & Policy* 6: e10. <https://doi.org/10.1017/dap.2024.3>.
- Pariser, E. (2012): *Filter Bubble. Wie wir im Internet entmündigt werden*. München
- Peters, T. M.; Visser, R. W. (2023): The Importance of Distrust in AI. In: Longo, L. (Hg.) (2023): *Explainable Artificial Intelligence: Communications in Computer and Information Science*, Band 1903. Cham, S. 301–317. https://doi.org/10.1007/978-3-031-44070-0_15
- Rohlfing, K. J.; Cimiano, P.; Scharlau, I.; Matzner, T.; Buhl, H. M.; Buschmeier, H.; Esposito, E.; Grimminger, A.; Hammer, B.; Häb-Umbach, R.; Horwath, I.; Hüllermeier, E.; Kern, F.; Kopp, S.; Thommes, K.; Ngonga Ngomo, A.-C.; Schulte, C.; Wachsmuth, H.; Wagner, P.; Wrede, B. (2021): Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. In: *IEEE Transactions on Cognitive and Developmental Systems* 13(3), S. 717–728. <https://doi.org/10.1109/TCDS.2020.3044366>
- Rosa, H. (2019): Tut mir leid, ich bin zahlungsunfähig. Reflexionen über die Temporalinsolvenz. Online unter <https://www.recherche-online.net/texte/hartmut-rosa-tut-mir-leid-ich-bin-zahlungsunfaehig/> [Letzter Aufruf: 16.6.25]
- Simanowski, R. (2018): *Stumme Medien. Vom Verschwinden der Computer in Bildung und Gesellschaft*. Berlin
- Weiser, M. (1991): „The Computer for the 21st Century“. *Scientific American* 265(3), S. 94–104
- Wiegerling, K. (2024). Ubiquitous Computing und informatische Vernetzung. In: Gutmann, M.; Wiegerling, K.; Rathgeber, B. (Hg.) (2024) *Handbuch Technikphilosophie*. Stuttgart, S. 345–354
- Zuboff, S. (2018): *Das Zeitalter des Überwachungskapitalismus*. Frankfurt am Main

