„Update verfügbar – ein Podcast des BSI“

Transkription für Folge 69

Titel: Agentic AI

Moderation: Schlien Gollmitzer und Hardy Röde

Gast: Benjamin Bechtloff, Frauenhofer IAIS

Schlien Gollmitzer: Hardy?

Hardy Röde Röde: Schlien?

Schlien Gollmitzer: Was hältst du davon, wenn wir uns einen Praktikanten holen für den Podcast?

Hardy Röde: Jemand, der uns Kaffee ins Studio bringt, oder was?

Schlien Gollmitzer: Ich finde es ganz lustig, dass du jetzt sofort an eine eher unverantwortungsvolle Aufgabe denkst, die du unserem Praktikanten geben würdest.

Hardy Röde: Also, um jetzt im Bild zu bleiben, Kaffee ist wirklich eine sehr verantwortungsvolle Aufgabe. Aber so ernsthaft und wertschätzend gegenüber allen großartigen Praktikanten und Praktikantinnen, die ich in meinem Leben schon kennenlernen durfte. Ich würde natürlich so jemanden sanft an Aufgaben heranführen und nicht sofort am Tag eins sagen, hallo, hier ist ein Interview mit dem Bundeskanzler, möchtest du das vielleicht führen oder Fakten recherchieren, die hinterher unüberprüft auf Sendung gehen.

Schlien Gollmitzer: Naja, klar, unüberprüft natürlich nicht, aber vorrecherchieren vielleicht schon, aber wir würden dann natürlich unsere journalistisch ausgebildeten Nasen schon auch noch mal darüberlaufen lassen.

Hardy Röde: Verantwortung ja, aber hinterher noch drauf schauen. Dass ist das worauf du raus willst. Da könnte ich mich übrigens auch total damit anfreunden.

Schlien Gollmitzer: Und du wirst staunen. Theoretisch könnten wir uns dafür sogar ganz kostengünstig einen KI-Agenten engagieren. Und um die geht es jetzt heute auch in dieser Folge tatsächlich. Agentic AI, die wiederum noch einen Schritt weitergeht als diese klassische KI, die wir aktuell so im Alltag mittlerweile nutzen.

Hardy Röde: Also ich nehme an wir bleiben jetzt bei bei Agentic AI und nicht bei Praktikantic AI im Lauf dieser Folge. Gut, also wie soll uns ein KI-Agent dann einen Kaffee bringen? Ich habe wirklich klare Vorstellungen. Soll ich es dir kurz erklären oder würde das zu weit gehen?

Schlien Gollmitzer: Vielen Dank. Also tatsächlich bräuchte es ja dann natürlich einen KI-Agenten-Roboter wiederum. Ich bin nicht sicher, ob wir den bewilligt bekommen beim BSI, bloß weil du jetzt zu faul bist in die Küche zu gehen und dir selber Kaffee zu machen.

Hardy Röde: Auch wieder Realtalk. Was glaubst du, wie vielen Praktikanten und Praktikantinnen ich schon Kaffee gemacht hab, weil ich will, dass sie guten Kaffees trinken.

Hardy Röde: Aber wieder im Ernst, so ein Rechercheagent, also KI-Agent letztendlich, der schon mal die Vorarbeit leistet, uns vielleicht hilft, interessante Themen zu identifizieren aus dem Wust da draußen und dann vielleicht eins weitergeht. Also so automatisiert interessante Gäste kontaktiert. Vielleicht sogar sagt, haben sie da Zeit, der Hardy hat da Zeit, die Schlien hat da Zeit. So in die Richtung geht die Denke, nehme ich an. Da kann ich mich damit anfreunden, ja.

Schlien Gollmitzer: Unseren Gast für diese Folge haben wir allerdings jetzt noch ganz menschlich und händisch selbst gefunden, tatsächlich. Benjamin Bechtloff. Benjamin ist Projektleiter am Fraunhofer Institut für intelligente Analyse- und Informationssysteme und erforscht da zum Thema Agentic AI. Sein Forschungsschwerpunkt liegt auf der Frage, wie diese Agenten verantwortungsvoll, sicher und wertschöpfend in Industrie und Gesellschaft eingesetzt werden können. Aber ich würde sagen, bevor wir ins Gespräch mit ihm gehen, wäre es vielleicht doch noch ganz hilfreich aufzudröseln, was genau dieses Agendic AI ist und wie sich es eben von der KI aus unserem Alltag unterscheidet. Hardy, und bitte.

Hardy Röde: Was ist eigentlich Agentic AI?

Hardy Röde: KI-Modelle, die wir im Alltag nutzen, funktionieren bislang meist wie ein Assistent auf Abruf. Wir stellen zum Beispiel eine Frage, das KI-Modell gibt eine Antwort und damit ist der Vorgang abgeschlossen. Agentic AI, also agentische KI, geht aber einen Schritt weiter. Sie liefert nicht nur eine Antwort, sondern sie kann durch Zugriff auf andere Systeme und Werkzeuge auch vorgegebene Ziele selbstständig verfolgen. Sie erledigt also in mehreren Schritten komplexe Aufgaben. Ein Beispiel wäre: Buche mir eine komplette Reise nach Italien. Ein agentisches KI-Modell kann dann Flüge vergleichen, Hotels prüfen, Termine abgleichen und, sofern es von der Nutzerin oder dem Nutzer dazu berechtigt wurde, auch Buchungen vorbereiten oder sogar auslösen. Der große Vorteil liegt in der Automatisierung komplexer Abläufe. Aufgaben können dann schneller, effizienter und mit weniger Aufwand erledigt werden. Gleichzeitig entstehen aber auch Risiken. Wenn ein KI-Modell eigenständig handelt, können Fehler passieren. Zum Beispiel falsche Buchungen oder unerwünschte Änderungen oder am Ende unnötige Kosten. Zudem benötigen agentische KI-Modelle häufig Zugriff auf sensible Daten, etwa Kalender, E-Mails oder Konten und kommunizieren mit anderen Systemen, in unserem Beispiel etwa um Flugdaten zu erfragen. Das erhöht natürlich auch Angriffs- und Missbrauchsrisiken.

Schlien Gollmitzer: Benjamin, herzlich willkommen und vielen Dank, dass du dir die Zeit für uns nimmst.

Benjamin Bechtloff: Sehr gerne, grüß dich!

Schlien Gollmitzer: Du hattest mir ja schon erzählt, du bist gerade umgezogen und hast dafür tatsächlich auch im Privatbereich Agentic AI genutzt. Wie lief das ab? Wofür hast du es da genutzt?

Benjamin Bechtloff: Ja, war natürlich ein kleines Experiment, genau, ich bin kürzlich umgezogen, auch vor dem Hintergrund, um mich beruflich intensiver mit dem Thema Agentic AI zu befassen. Da dachte ich, alles klar, dann können die Agenten doch direkt mit anpacken, zumindest digital. Ja, es gibt solche Systeme, die dann eben auch im Internet-Browser laufen. Und ja, dem habe ich dann sozusagen die Aufgabe gestellt, für mich nach einem Umzugsunternehmen zu suchen, was in einer entsprechenden Zeit eben in der Lage ist, mir mal ein Angebot zu machen für meinen Umzug. War sehr, sehr spannend. Ich habe dann sozusagen meinen Tag weiter durchgeführt, war weiter in meinen Arbeitsaufgaben

versunken. Später klingelte dann mein Telefon und am anderen Ende meldet sich dann jemand vom Umzugsunternehmen, der mir dann die Frage gestellt hat, ob die Anfrage von mir kommt, ob ich, ja, ob die Daten alle vollständig sind und hat mich dann gebeten, noch ein paar Fotos von meiner Wohnung zu machen, damit er das Angebot noch mal spezifizieren kann. Das war für mich auf jeden Fall absoluter Wow-Moment, wenn man so möchte, weil ich gemerkt habe, ey, hier hat ein KI-Agent tatsächlich in meinem Alltag eine Aufgabe für mich übernommen. Einiges an Recherche durchgeführt, bis zu einem Punkt, wo dann auf einmal mein Telefon klingelte.

Schlien Gollmitzer: Wo dann tatsächlich die Buchung ansteht, aber das ist auch ein bisschen gruselig, oder? Kam das nicht auch komisch vor?

Benjamin Bechtloff: Es kam mir auf jeden Fall komisch vor. Man muss vielleicht dazu sagen, wenn ich so was nutze, sowohl privat als auch beruflich, dann immer mit einer gewissen Vorsicht, auch vor dem Hintergrund meiner eigenen Daten. Das heißt, ich benutze da eine Wegwerf-Handy-Nummer, eine Wegwerf-E-Mail-Adresse. Genau, aber ich habe auch natürlich berufliche bedingt eine gewisse experimentelle Neugier, wenn ich so was mache. Und ja, das ist gruselig auf der einen Seite, aber vielleicht auch ein Stück weit die Zukunft.

Schlien Gollmitzer: Spannend. Siehst du das wirklich im Alltag von Menschen in Zukunft? Privatpersonen?

Benjamin Bechtloff: Absolut. Ich glaube, da muss man grundlegend unterscheiden. Es gibt, glaube ich, bei jeder neuen Technologie immer die sogenannten Early Adopter, also die, die mit einer gewissen Begeisterung und auch einer gewisse Risikoaffinität an solche Technologien rangehen. Und das sind dann die Personen, die dann die frühen Experimente machen. Wir sehen aber durchaus den Schritt dahin. Dazu, dass dann solche Systeme nach und nach mehr auch im Alltag ankommen. Genau, ich habe dieses Beispiel gemacht. Man kann die Agentic AI jetzt sozusagen in seinem Browser verwenden. Viele Leute nutzen eben solche Systeme wie ChatGPT, wie Gemini, wie Claude und auch das sind, wenn man so möchte, auch schon agentische Systeme. Der erste Chatbot oder die erste Variante von ChatGPT hat einfach aus dem Modell heraus geantwortet. Mittlerweile, wenn ich die Frage stelle: Ich suche für eine Feier am Wochenende noch irgendwie ein schwarzes Hemd. Dann benutzt auch ChatGPT mittlerweile Werkzeuge. Es fängt an zu planen, es fängt eben an abzuwägen, was sind die Möglichkeiten, um dieses Ziel, diese Aufgabe dann noch zu erledigen. Und ja, das sind ebenso Möglichkeiten, wie auch Agentic AI dann nach und nach im Alltag ankommt. Ich würde dennoch sagen, dass wahrscheinlich insbesondere im wirtschaftlichen Sinne diese Systeme vielleicht sogar noch einen Schritt früher letzten Endes den Menschen begegnen werden, weil Unternehmen dann doch das Ziel haben, eine gewisse Produktivität zu haben, Effizienzsteigerung zu erreichen.

Schlien Gollmitzer: Also das heißt, du siehst es schon auch eher so, wie du sagst, Wirtschaft, Im Handel vielleicht sogar. Also etwas, das Unternehmen einbauen, um vielleicht sogar Verbraucherinnen oder Verbrauchern den Einkauf zu erleichtern, als dass ich das jetzt zu Hause hier in meiner Küche mit dem Kühlschrank verwende, der für mich einkauft.

Benjamin Bechtloff: Sowohl als auch. Es ist sicherlich Beides möglich. Es gibt eben die Möglichkeit, dass man sich selber seinen eigenen persönlichen Assistenten baut. Auch mit solchen Produkten wie ChatGPT, Claude, Gemini, aber auch so persönliche Open-Source-Agenten wie OpenClaw, was ja Anfang des Jahres viel durch die Presse gegangen ist. Das sind eben solche Systeme, die vor allem so persönlichen Assistenten sind, wo ich mir, ja, ich möchte mal sagen, so meinen eigenen Assistenten erstelle. Für die Aufgaben und Herausforderungen, die ich im Alltag habe, dann wirklich sehr persönlich und dann Agentic

AI als Produkt eingebaut in unterschiedlichen Use Cases, in unterschiedlichen Systemen, zum Beispiel beim Online-Shopping, zum Beispiel in der Kundenberatung.

Schlien Gollmitzer: Weil du OpenClaw gerade ansprichst, das ist ja eine Open Source Software, die eine Grundlage sozusagen für diese persönlichen KI-Assistenten oder Agenten in dem Fall sein kann. Jetzt wird bei OpenClaw aber trotzdem immer wieder vor mangelnder Sicherheit gewarnt. Was macht denn da die Software so gefährlich?

Benjamin Bechtloff: Ich würde das Wort gefährlich ein bisschen relativieren. Also kurze Einordnung, OpenClaw ist sozusagen ein Open-Source-KI-Agentensystem. Das System läuft auf dem Rechner, also ich kann es mir sozusagen im Internet runterziehen, bei mir installieren und kann dann auch, nachdem ich ein Sprachmodell, also immer noch das KI-Modell als solches ist da Dreh- und Angelpunkt und sozusagen dann das Gehirn des Systems, ohne dass auch nichts läuft, kann das System, wenn ich das dann installiert habe, über einen Messenger wie WhatsApp, Telegram oder auch Signal steuern und kann dann verschiedenste Aufgaben recht flexibel machen. Also das System ist ein Stück weit auch in der Lage, je nachdem welche Berechtigung ich dem System gebe, wie ich es dann auch aufsetze, meinen kompletten Computer unter Umständen zu steuern. Und da stecken natürlich auch ein Stück weit die Risiken. Grundlegend würde ich sagen, es ist so gefährlich, wie es dann letzten Endes auch aufgebaut ist, ganz persönlich bei mir. Wenn ich bestimmte Mechanismen auch beachte, wenn ich darauf achte, die Kontrolle zu behalten, wenn ich dem System auch nicht zu viel Autonomie gebe, dann kann man die Gefahr ein Stück weit managen.

Schlien Gollmitzer: Ja, ich wollte gerade sagen, du hast ja vorhin schon erzählt, dass du eine Wegwerf-E-Mail-Adresse dann in dem Fall lieber verwendest. Hast du dann auch den Wegwerflaptop eigentlich?

Benjamin Bechtloff: Ich habe mir tatsächlich einen eigenen Computer gekauft, weil es mir dann schon wichtig war, dass das System nicht auf all meine Daten zugreift. Es gibt aber auch so technisch gelagerte Risiken, also in dem Moment, wo das System zum Beispiel mit anderen Systemen interagiert, im Internet sich austauscht, vielleicht auch mit anderen KI-Agenten sich austauscht. Da gibt es so Möglichkeiten von Prompt-Injections, also sozusagen versteckte Anweisungen, die in mein KI-System gefüttert werden, mit dem Ziel, Schaden anzurichten, das System ein Stück weit umzuprompten oder Daten zu exfiltrieren. Gibt aber auch Risiken, die daraus entstehen, die, wenn ich dem System auch zu viele Berechtigungen gebe. Ich habe in der Presse und in Social Media auch von solchen Fällen gehört, wo Menschen den Mut hatten, dem System Zugriff auf Zahlungsmittel zu geben und dann kam die eine oder andere Rechnung. Ich glaube, davon ist definitiv abzuraten.

Schlien Gollmitzer: Aber weil du gerade gesagt hast, die Agenten, die KI-Agenten tauschen sich gegenseitig auch aus. Anfang des Jahres ging ja mit Moltbook ein erstes Internetforum für KI-Agenten an den Start. Was hat es denn damit auf sich? Ist das tatsächlich wie so ein Social-Media-Netzwerk, Social-Netzwerk nur für KI-Agenten?

Benjamin Bechtloff: Wenn man so möchte, ja. Inspiriert ist Moltbook von dem sozialen Netzwerk Reddit mit dem Clou, dass das eine reine Plattform ist, wo sich KI-Agenten austauschen. Die Menschen sind in dem Social Network tatsächlich nur die Beobachter. Es ist natürlich durch die Presse gegangen, weil dann relativ schnell verschiedenste Beiträge kamen, wie Karriereagenten gründen, Religionen, tauschen sich über ihre Besitzer aus sozusagen. Und das war natürlich sehr spannend und hat für große Schlagzeilen gesorgt.

Schlien Gollmitzer: Was mein Mensch neulich wieder für eine Reise buchen wollte.

Benjamin Bechtloff: Genau. Er wurde sozusagen ein bisschen gelästert und das hat natürlich für ein Stück weit auch berechtigterweise für Unbehagen gesorgt. Ich glaube, man kann das aus einer heutigen Perspektive auch ein Stückweit entzaubern. Also vieles davon ist von Menschen gesteuert, denn hinter jedem KI-Agenten steckt immer noch ein Prompt. Und den muss jemand schreiben. Und man konnte auch in verschiedenen Forschungen mittlerweile feststellen, dass hier sozusagen ein eher menschlicher Fingerabdruck auch zu sehen ist. Die Frage, die wir uns dann auch stellen oder die ich mir dann auch stelle, wie wertvoll ist das tatsächlich für die Forschung? Der eine oder andere hat gesagt, das ist ein erstes großes Feldlabor, wo sich KI-Systeme untereinander austauschen, um das Verhalten vieler Agenten dann eben zugleich auch zu beobachten, zu schauen, wie sie sich koordinieren. Was man tatsächlich lernt, ist, dass Modelle vielleicht auch einfach nur Trainingsmuster nachspielen. Bei all der Aufregung um die Gründung von KI-Religion sehe ich da eher das Nachahmen menschlicher Muster dessen, wie Menschen sich auf Social Media verhalten. Aber ja, wenn man sich das aus wissenschaftlicher Perspektive anschaut, ist es eben kaum unterscheidbar, was dann wirklich vom Agenten, was dann auch vom Menschen stammt.

Schlien Gollmitzer: Jetzt hast du schon angesprochen, es kommt immer auf den Zugriff an. Wir waren gerade bei Zahlungsmitteln. Also angenommen, wir hatten vorhin das Beispiel einer Reise. Angenommen, ich würde jetzt meinen Agenten darauf ansetzen, mir eine wunderschöne Urlaubsreise zu buchen. Dann muss ich mein Zahlungsmittel im Grunde dann schon zur Verfügung stellen. Jetzt sucht sich die KI oder der Agent in dem Fall dann raus, ja die Schlien macht jetzt dann schön Urlaub auf Gran Canaria und bezahlt folgendes Hotel. Und das wird auch schon mal alles gebucht und jetzt bin ich aber nicht zufrieden damit, was der Agent da gemacht hat. Wie ist das denn dann mit dieser Zahlung, mit dieser Buchung? Im Grunde steckt dahinter ja die Frage, wer haftet bei Fehlern?

Benjamin Bechtloff: Am Ende des Tages haftet schon immer noch der Betreiber, also wenn das mein eigener persönlicher Agent ist, den ich dazu befähigt habe, mit meiner Kreditkarte durchs Internet zu ziehen, dann ist das tatsächlich meine eigene Verantwortung. Es ist auch nicht zu empfehlen, zumindest nicht im aktuellen Stand der Technik, das würde ich ganz klar sagen. Es empfiehlt sich bei allen Aktionen, die irgendwie eine, ich sag mal, Echtweltkonsequenz haben, die eine Art von Transaktion bedingen, hier sozusagen den Human in the Loop zu haben. Als Mensch selber noch, mal mindestens auf Annahme zu drücken, wenn der Vorschlag kommt, diese Reise zu buchen.

Schlien Gollmitzer: Um die Kontrolle zu behalten einfach, ne?

Benjamin Bechtloff: Um die Kontrolle zu behalten über die Auswirkungen. Letzten Endes gilt das für mich als Privatperson, das gilt aber auch für Unternehmen. Das heißt, wenn ein Unternehmen so ein System auf seiner Website einsetzt, dann muss es sich darüber im Klaren sein, das ist für den Output des Systems verantworten muss.

Schlien Gollmitzer: Also es ist am Ende des Tages im Grunde trotzdem der Praktikant. Es ist kein ausgebildeter Mitarbeiter, so darf man es als Unternehmen nicht verstehen, sondern es ist ein Praktikant, dem man eine Aufgabe gegeben hat, aber am Schluss sollte dann doch nochmal jemand drauf schauen.

Benjamin Bechtloff: Also es ist definitiv ein sehr, sehr, sehr fähiger Praktikant, aber gerade bei Aktionen mit einer großen Konsequenz. Wir reden da eben von so Art Risikoklassifizierung auch ein Stück weit, das machen wir auch, wenn wir solche Systeme für Unternehmen entwickeln. Dann schauen wir uns an, welche Art von Aktion kann welche Konsequenz haben, wie teuer ist ein potenzieller Schaden, der eben bei einer Fehleinschätzung, Fehlaktion dann auch entsteht und da kann man vielleicht in verschiedenen Kategorien auch denken. Und es kommt ein Stück weit auch auf die eigene

Risikoaffinität an. Die Konsequenz, eine E-Mail an die falsche Person zu schicken, ist vielleicht unangenehm. Hat aber sicherlich nicht dieselbe Konsequenz, wie wenn hier eine Reise für viele tausend Euro gebucht wird. Und ich denke auch, ein Stück weit wird in den nächsten Jahren hier auch die Technik definitiv abgesicherter und die Forschung auch deutlich besser. Aber letzten Endes ist schon zu empfehlen, vor tatsächlichen Aktionen immer noch mal auch darauf zu schauen.

Schlien Gollmitzer: Viele, sagen wir mal, technische Entwicklungen auch im Netz sind ja im Grunde daraus entstanden durch Menschen wie dich. Early Adopter, die damit einfach mal rumspielen, die vielleicht auch gewisse Risiken eingehen. Wie siehst du denn da so den zeitlichen Rahmen? Also danke erst mal an alle Early Adopter, aber wann kommt es zu uns?

Benjamin Bechtloff: Ich würde sagen, man braucht hier nicht von Jahren zu reden, wahrscheinlich auch nicht von Monaten. Jeder, der solche Systeme wie ChatGPT oder Claude benutzt, der verwendet auch schon agentische Systeme. Ich denke, wir reden hier von wenigen Monaten. Meine persönliche These ist, dass so der ganz, ganz große Durchbruch dann kommt, wenn solche Systeme auf dem Smartphone tatsächlich nutzbar sind.

Schlien Gollmitzer: Jetzt hast du aber ja auch schon gesagt, eigentlich sicherheitstechnisch sind wir noch nicht so weit. Also die Early Adopter können damit umgehen. Du benutzt einen Wegwerfrechner, um das zu nutzen. Und auch Anthropic hat ja zuletzt Schlagzeilen gemacht, du hast das gerade genannt, mit der Forderung nach einer weltweiten Pause in der KI-Entwicklung. Was würdest du jetzt sagen? Geht das alles ein bisschen zu schnell?

Benjamin Bechtloff: Ich würde sagen, der Stand der Technik schreitet wirklich sehr, sehr schnell fort und vielleicht auch schneller als die Gesellschaft, vielleicht auch die Regulation solcher Systeme da auch hinterherkommt. Die technologischen Zyklen, die haben sich massiv verkürzt, auch vor dem Hintergrund, dass solche Systeme im Bereich der Softwareentwicklung eingesetzt werden können. Vor dem Hintergrund, ja, würde ich schon sagen, dass es Sinn machen könnte, mal kurz innezuhalten und sich zu überlegen, okay, was ist denn Sinn und Zweck des Ganzen und wer profitiert letzten Endes davon, welche Risiken kann das haben. Aber wenn wir über Risiken von Agentic AI sprechen und auch das Tempo, dann muss man schon noch sagen, das es sehr, sehr vielschichtig letzten Endes ist. Also wir reden hier von einer System- und Technik-Ebene. Wir haben eben schon mal das Thema Prompt Injection angesprochen, das Thema Datensouveränität, die Konsequenz von Fehlhandlungen. Das sind natürlich irgendwie kaskadierende Fehler, die aus dem System heraus entstehen. Dann gibt es auch die Sorge, und ich glaube, die schwingt so ein Stück weit auch bei Anthropic mit. Dass diese Modelle einfach ein Stück weit zu potent geworden sind oder so potent, also gerade im Cyber Security Bereich eine Potenz erreicht haben, dass sie in der Lage sind, Sicherheitslücken in nahezu jedem System zu finden. Und der Gedanke, warum zum Beispiel auch Anthropic mit ihrem neuesten Modell, stärksten Modell Mythos gesagt hat, okay, das wird jetzt nicht mal morgen auf die Öffentlichkeit losgelassen, sondern wir müssen jetzt vor allem erst mal den Unternehmen Zugriff darauf geben, die Teil der kritischen Infrastruktur sind, damit sie diese Sicherheitslücken bei sich selber finden und dann auch erst mal schließen können. Und ja, das ist mal so ein Aspekt. Wenn man dann noch ein Stück weit weiterdenkt, dann habe ich eben auch darüber gesprochen, dass diese Systeme in der Lage sind, viele Aufgaben zu übernehmen. Das heißt, wir reden hier von einem recht hohen Automatisierungspotenzial. Zuerst, vielleicht erst in der Bürowelt. Also wenn so ein System nahezu jede Aufgabe auf einem gewissen Niveau auf meinem Computer durchführen kann, dann kann man sich auch überlegen, was das für mögliche Auswirkungen auf den Arbeitsmarkt haben kann. In den nächsten Jahren wird das Thema Physical AI, also Robotik, fortschreiten. Das heißt während wir jetzt ein großes

Automatisierungspotenzial gerade in der digitalen Welt haben, wird das in der physischen Welt ebenso fortschreiten.

Schlien Gollmitzer: Dann räumt endlich der Roboter meine Spülmaschine aus.

Benjamin Bechtloff: Dann räumt der Roboter die Spülmaschine aus, übernimmt vielleicht auch in der Logistikhalle das Sortieren von Paketen. Haushaltsrobotik ist sicherlich auch ein Thema, was dann in den nächsten Jahren kommen wird. Und wir merken, wenn man so diese Themen zusammennimmt, dann hat das Schritt für Schritt weitere Folgen und Effekte. Das sind natürlich auch Fragen, die mich auch um meine Arbeit beschäftigen, die mir auch schon mal die eine oder andere Stunde Schlaf gekostet haben. Weil, ja, die Auswirkung auf den Arbeitsmarkt ist das eine. Ich glaube, psychologisch ist das auch eine sehr, sehr relevante Frage. Ich glaube, viele Menschen gehen nicht nur zur Arbeit, um Geld zu verdienen, sondern eben auch, um für sich eine Identität zu haben. Ihren Selbstwert darüber auch zu definieren, der dann ein Stück weit auch an der Arbeit hängt, das Gefühl gebraucht zu werden. Und das sind schon elementare Risiken, die dann aus dieser gesamten Transformation fortschreiten. Und am Ende des Tages muss man sich schon die Frage stellen, ist diese Transformation ein Selbstzweck? Ist KI ein Selbstzweck? Also nicht die Technik per se ist ja das Ziel, wenn man so möchte, sondern es muss ja schon der Nutzen für Mensch und Gesellschaft ein Stück weit garantiert sein. Und ich denke, darauf muss sich dann konzentriert werden in den nächsten Schritten bei der Entwicklung solcher Systeme.

Schlien Gollmitzer: Du hast jetzt gesagt, es dauert nicht mehr lange, bis das Ganze hier auf unseren Smartphones tatsächlich aktiv genutzt wird, auch wenn es im Hintergrund schon so ein bisschen genutzt wird aktuell. Aber kannst du uns vielleicht so ein paar kleine Tipps geben, wenn uns die Agentic AI begegnet? Worauf sollten wir achten oder woran merken wir überhaupt, dass die Agentic AI gerade für uns arbeitet, dass wir sie tatsächlich schon nutzen? Also ich habe mir jetzt schon mal gemerkt, Wegwerf-Handy auf jeden Fall besorgen. Wegwerflaptop schon mal bereitlegen für sowas. Aber wie ist es denn so in der Nutzung für uns?

Benjamin Bechtloff: Also vielleicht, um auf deinen Aspekt vorerst mal einzugehen, man muss nicht unbedingt einen Wegwerf-Laptop- oder PC besorgen. Es gibt auch so Möglichkeiten, das in einer sicheren Umgebung, so einer Art Sandbox auf seinem Gerät laufen zu lassen. Das ist definitiv zu empfehlen.

Schlien Gollmitzer: Sandbox ist immer so ein bisschen der Spielkasten, der Sandkasten, in dem man ein bisschen ausprobieren darf, bevor man die ganz große Burg baut.

Benjamin Bechtloff: Genau, und da gibt es eben technische Möglichkeiten, dass so ein Agent dann nicht direkt Zugriff auf den gesamten Computer hat und sozusagen die Konsequenzen in der Sandbox dann auch stattfinden. Ich würde sagen, letzten Endes ist mir eigentlich eine Sache sehr, sehr wichtig. Also wenn man das verwendet, dann sollte man schon immer noch schauen, dass man selber im Loop bleibt und dass man sich mit den Themen dennoch beschäftigt. Denn es ist, glaube ich, schon wichtig, zu überblicken und zu verstehen, was passiert denn? Also wenn so eine KI einfach mal einen halben Tag arbeitet, dann ist das schon sehr, sehr schwer, das alles zu nachzuvollziehen und zu überblicken. Aber ich glaube, es ist schon in unserem eigenen Interesse, dass man immer noch selber im Thema bleibt und das ist auch schon wichtig. Ja, das würde ich für dich als einen der wichtigsten Aspekte hier tatsächlich auch nennen.

Schlien Gollmitzer: Nicht jeden Quatsch glauben und den Rechenweg nachvollziehen.

Benjamin Bechtloff: Genau das ist der Punkt. Ja, und letztendlich, wenn ich mit meinen Freunden darüber spreche und sage ich eigentlich, in erster Linie ist es total wichtig, sich

überhaupt erstmal mit der Thematik zu befassen. Ja, probiere einfach mal aus, aber nicht ohne Sinn und Verstand, sondern beachte einfach die wichtigsten Regeln, um ja, hier keinen Schaden dann letzten Endes auch für sich selber anzurichten. Aber, dass das Thema relevant ist, ist außer Frage, dementsprechend sollte sich glaube ich jeder mal damit befassen.

Schlien Gollmitzer: Auch die 80-jährige Nachbarin? Nein…

Benjamin Bechtloff: Ich denke schon. Letzten Endes kommt es immer darauf an, was ist mein Ziel? Was ist meine Aufgabe? Ja, wir hatten ja auch schon mal im Vorgespräch darüber gesprochen, wenn wir gerade im Bereich Handel oder Online-Shopping darüber sprechen. Es ist kein Selbstzweck. Also der eine hat Spaß daran und will einfach letzten Endes das Ziel haben, also das T-Shirt letzten Endes auf dem Tisch haben und hat keine Lust auf den Prozess und lagert solche Aufgaben dann gerne an KI-Agenten aus. Aber dann gibt es natürlich auch Leute, die haben einfach Spaß am Prozess und gehen selber gerne einkaufen und das ist auch vollkommen in Ordnung. Aber ja, es wird ein relevantes Thema und man sollte sich schon damit befassen.

Schlien Gollmitzer: Benjamin, ganz herzlichen Dank dir. Was für ein spannendes Gespräch, sehr interessant. Ich bin gespannt, wann es auf meinem Smartphone tatsächlich ankommt und ich anfange, die blöden Aufgaben an den Agenten weiterzugeben.

Benjamin Bechtloff: Super gerne, hat mir sehr viel Spaß gemacht. Dankeschön.

Hardy Röde: Also ich nehme aus eurem Gespräch, Schlien, eine Message mit, die wir schon oft hier hatten, in „Update verfügbar“, nämlich ganz simpel Up-to-date bleiben, wissen, was passiert in dieser komischen Maschine, die so magische Dinge für uns erledigt, auf die wir irgendwie total stehen. Aber wissen, was in der Box so rund geht, das kann uns niemand abnehmen.

Schlien Gollmitzer: Das ist tatsächlich so. Ja, wir hatten das schon öfter hier im Podcast. Also wer sich mit den Grundlagen auskennt, der ist auf jeden Fall schon mal eher geschützt als jemand, der einfach total den Überblick verloren hat und nicht mehr mitkriegt. Gerade eben, weil diese Entwicklung so schnell geht, wie Benjamin ja auch gesagt hat. Das ist ja eh klar. Also ich würde sagen, das setzen wir mal auf jeden Fall auf die Eins unserer Checkliste.

Hardy Röde: Die ewige Checkliste und rumspielen mit diesen Tools, da macht es natürlich grundsätzlich Bock. Also es ist einfach wirklich eine total interessante, gute Vorstellung und es lädt dazu ein, aber dass man es in einem sicheren Rahmen machen soll. Also wirklich nur so viel Daten rüber schiebt, so viele Berechtigungen freigibt, wie nötig. Und auch so ein Thema, was wir das Öfteren hatten, hier in “Update verfügbar“, vielleicht nicht die gleiche E-Mail-Adresse für alle Konten und Accounts verwenden. Das könnte mit so einem Agentic AI-Tool vielleicht noch ein bisschen schiefer gehen, als es ohnehin schon manchmal ist.

Schlien Gollmitzer: Ja, ja, ja. Ich habe mir das jetzt tatsächlich zu Herzen genommen, was du immer und immer wieder sagst. Und ich habe jetzt tatsächlich auch mal so zwei, drei neue E-Mail-Adressen mir zugelegt. Eine davon ist tatsächlich eine sogenannte Wegwerf-E-Mails-Adresse, um genau solche Rumspielereien einfach mal auszuprobieren. Mit denen kann ich da auch mal ein paar nicht ganz so gesicherte Dinge im Netz ausprobieren.

Hardy Röde: Ja, das finde ich natürlich super. Lass dir von einem Early Adopter der Wegwerfadresse, der seit einem Jahr, seit wir es hier im Podcast davon haben, praktiziert sagen, man wirft die leider nicht weg. Die bleiben einfach immer am Leben.

Schlien Gollmitzer: Okay, dann brauchen wir aber noch unseren dritten Punkt auf der Checkliste. Und der sollte auf jeden Fall auch sein: Lass mal den KI-Agenten noch nichts Kriegsentscheidendes alleine ausmachen und buchen. Also, bevor der KI- Agent uns den Kaffee tatsächlich kochen darf, soll er vielleicht doch erstmal einen Tee machen oder zum Vorkosten. Und am Ende der Recherche- oder Buchungskette sollte dann doch immer noch ein Mensch zur Kontrolle sitzen und den letzten Schritt tun. Nachdem er den Rechenweg der KI aber auch geprüft hat. Das klingt alles erstmal so ein bisschen mühsam, ist aber tatsächlich sicherheitsrelevant.

Hardy Röde: Ja, ja, weil am Ende die Kontrolle und das Verständnis sollten wir Menschen immer behalten, auch gerade bei solchen hochautomatisierten Systemen, weil wir letztlich auch die Verantwortung haben. Also da schließt sich der Kreis irgendwie. Selbst informiert bleiben, wissen, was das Tool tut und nicht blind vertrauen.

Schlien Gollmitzer: So ist das. Wir haben zu diesem Thema übrigens noch einige weiterführende Links vom Fraunhofer-Institut und natürlich auch dem BSI für euch in die Show Notes gestellt, wie jedes Mal.

Hardy Röde: Und das war's dann auch mit unserer Folge 69 von Update verfügbar. Alle Informationen zum Umgang mit Agentic AI und sonstigen raketenwissenschaftlich generierten Tools kriegt ihr, wie Schlien gesagt hat, auf der Seite des BSI und bei uns in den Show Notes. Und da packen wir euch auch ein paar coole Tipps für richtig guten Kaffee rein.

Schlien Gollmitzer: Oder Praktikanten-Ausschreibungen.

Hardy Röde: Sag mal, was hast du jetzt für ein Ding mit den Leuten, die gerade ihre ersten Füße in diesen zynischen Medienbetrieb setzen?

Schlien Gollmitzer: Tatsächlich ist das ja nicht so. Eigentlich sind sie ja die Stütze unserer Gesellschaft, weil nur die halten die ganzen Läden am Laufen.

Hardy Röde: Mehr Informationen, wie wir mit Agentic AI verantwortungsvoll umgehen, hat Schlien schon gesagt, die kriegt ihr auf der Seite des BSI und bei uns in den Show Notes. Und ihr könnt uns Kommentare hinterlassen, ob da auch so Tipps für richtig guten Kaffee rein sollen.

Schlien Gollmitzer: Ja gut, Hardy, da musst du dich dann allerdings drum kümmern, um diese Kommentare, solange wir noch keine Praktikantin und keinen Praktikanten haben.

Hardy Röde: Ich bin dabei.

Schlien Gollmitzer: Folgt uns bitte unbedingt auch auf Instagram, auf Mastodon, auf TikTok oder YouTube für mehr Infos zu allen sicherheitsrelevanten Fragen im Netz. Und abonniert unbedingt diesen Podcast, damit ihr auch ja keine neue Folge verpassen könnt. Wir sagen Tschau und bis zum nächsten Update.

Hardy Röde: Bis zum nächsten Update.