Heart-a-Tech Podcast by MaibornWolff

Folge 03:

Der Spion im eigenen Unternehmen:
Wie sicher ist Künstliche Intelligenz im Business-Kontext?

Mit Experte Konrad Schreiber

Intro Konrad Schreiber: Regulatorien gibt es und gab es in der IT immer schon. Diese Technologie wird uns dadurch jetzt nicht wieder weggenommen, sondern wir werden Lösungen finden, wie wir im Rahmen dieser Leitplanken diese Technologie nutzen können.

Brigitte Streibich: Hallo zur dritten Folge von Heart-a-Tech, dem Podcast rund um alles, was du wissen musst, um neue IT-Trends und Technologien erfolgreich in deinem Unternehmen zu implementieren. Dieser Podcast ist happily hostet by MaibornWolff. Mein Name ist Brigitte Streibich und wir sprechen heute über LLM-Applikationen: Wie sicher sind sie? Sind Sie vielleicht der Spion im eigenen Unternehmen? Ich freue mich sehr über meinen heutigen Gast, der in diesem Podcast auch schon bekannt ist. Es handelt sich um Konrad Schreiber, mit dem ich schon in der ersten Folge über den Mehrwert von KI-Sprachmodellen für Teams und Kunden gesprochen habe. Konrad hat langjährige Erfahrung im Bereich Data Science und IT-Architektur. Er ist Experte für die Anwendung von Machine Learning im Unternehmenskontext. Und er ist stellvertretender Bereichsleiter Data and AI bei MaibornWolff. Hallo, Konrad!

Konrad Schreiber: Hallo, Brigitte. Schön, dass ich wieder hier sein kann.

Brigitte Streibich: Ich freue mich ganz besonders auf diese Folge, weil es jetzt wirklich ans Eingemachte geht. Es geht um das Thema Datensicherheit und Datenschutz. In Deutschland ist man ja diesbezüglich sehr vorsichtig. Viele Unternehmen bekommen schon Schnappatmung, wenn sie das Wort „Cloud“ nur hören und wollen da eigentlich gar nicht so richtig den Schritt in die Cloud gehen. Du kennst das wahrscheinlich schon aus vielen Kunden-Projekten. In der Unternehmenswelt herrscht ja viel Skepsis gegenüber Technologien, die nicht aus Deutschland oder aus Europa kommen, insbesondere bei US-Technologien. Wir kennen das alle: Menschen geben etwas bei ChatGPT ein – etwas, das vielleicht noch gar nicht in die Öffentlichkeit gehört. Vielleicht irgendwelche Features eines Produkts, das noch nicht veröffentlicht ist. Das System lernt natürlich und merkt sich diese Dinge und irgendwann tauchen sie ganz woanders wieder auf. Jetzt stellt sich die Frage: Sind LLM-Applikationen Spione im eigenen Unternehmen? Kann das im Unternehmenskontext auch passieren?

Konrad Schreiber: Es handelt sich hierbei um ein absolut breites Thema und du hast jetzt ganz viele Aspekte angesprochen. Ich greife mal einen davon auf. Das werde ich auch immer von Kunden gefragt: *„Wenn ich etwas bei ChatGPT eingebe, sind dann diese Informationen weggeflossen?“* Nun, das stimmt nur zum Teil. Da kann ich durchaus beruhigen, ich muss aber auch warnen. Ja, alles, was man dort eingibt, gerade in öffentlichen Services wie etwa einem Bing Chat oder auch ChatGPT von OpenAI, kann grundsätzlich geloggt werden. Das heißt: Alles, was ich dort eingebe, unter Umständen auch sensible Informationen, weil ich das Sprachmodell darum bitte, etwas für mich zu erledigen – mit einer Exceltabelle, mit einer internen Information: All das wird und kann gespeichert werden. Und es kann auch passieren, dass diese Informationen für das Training der nächsten Modellgeneration verwendet werden. Bei bestimmten Informationen wäre das natürlich ein Desaster. Firmen wie Open AI haben da auch strikte Policies. Bei der Auswahl der Trainingsdaten sind sie sehr

vorsichtig und passen auf, dass sensible Informationen wahrscheinlich nicht eingefasst werden. Als Firma oder Organisation sollte man aber stark aufpassen und die eigenen Mitarbeiterinnen und Mitarbeiter auch schulen, solche Informationen eben nicht in öffentliche Services einzugeben, sondern dann In-House- oder B2B-Lösungen zu nehmen, wie ein Bing-Chat-Enterprise zum Beispiel oder etwas Eigenes zu bauen.

Ich werde auch häufig gefragt: *„Lernen diese Modelle diese Informationen, nur indem ich mit ihnen interagiere?“* Automatisch erst mal nicht. So ein Sprachmodell ist statisch. Das haben wir in diesem Podcast auch bereits besprochen. Das entwickelt sich nicht dadurch weiter, weil mit ihm gesprochen wird, sondern es bleibt – so salopp gesagt – auf der Infrastruktur liegen. Alles, was dieses Modell scheinbar aus einer Konversation lernt, passiert immer nur im Kontext des Sprachmodells und der liegt immer beim User und wird nach einer Konversation üblicherweise auch weggeschmissen. Du kennst das vielleicht, Brigitte. Wenn du ChatGPT öffnest, werden dir ja deine ganzen Konversationen angezeigt. Das heißt: Sie sind irgendwo gespeichert. Also Vorsicht an der Stelle! Das Sprachmodell selber, das LLM, hat die Informationen allerdings nicht in sich drin. Dazu müsste man die Konversation jetzt nehmen und für das Training der nächsten Generation dann wieder verwenden. An der Stelle sind wir schon mal sicher.

Brigitte Streibich: Verstehe. Doch vor allem dann, wenn es um solche neuen Technolgien geht, muss ja auch irgendwo von außen ein gesetzlicher Rahmen gegeben werden. Die Diskussion ist bekannt: Muss KI jetzt irgendwie reglementiert werden? Übernehmen die Systeme die Weltherrschaft, wenn hier nicht mal irgendjemand eingreift? Wie weit sind wir denn mit gesetzlichen Regelungen und mit dem Thema Datenschutz hier in Deutschland und in Europa?

Konrad Schreiber: Zum Thema Datenschutz fallen mir unmittelbar zwei Regelwerke ein. Zum einen die DSGVO, und zum anderen der EU AI Act. Ersteres kümmert sich um Datenschutz und Datensicherheit und das zweite reglementiert die Entwicklung von intelligenten Systemen und von künstlicher Intelligenz. Die DSGVO sieht vor, dass unsere Kunden ihre Daten nicht ohne Weiteres in Drittländer schicken dürfen, beispielsweise in die USA. Das Problem ist nur: Genau das tun wir, wenn wir mit ChatGPT interagieren. Das GPT-Modell ist auf einem Server in den USA gespeichert, bei OpenAI handelt es sich dabei auch um einen Microsoft-Server, aber eben in den USA. Das heißt: Alles, was ich das Modell frage, alle Daten, die ich dort hinschicke – all das geht direkt in die USA und wird dort wie gesagt unter Umständen auch geloggt und dann kommt halt die Antwort zurück. Man hat in so einem Fall die Daten einfach an einen Drittstaat abgegeben, was man laut DSGVO eigentlich nicht darf oder nur in bestimmten Fällen darf. Und genau aus diesem Grund hat Microsoft reagiert und gesagt: Wir nehmen diese OpenAI-Modelle und stellen die auch in Europa bereit, also legen sie wirklich auf europäischen Rechenzentren ab. In Deutschland wären das zum Beispiel Berlin und Frankfurt, aber auch in Frankreich beziehungsweise eben auch europaweit. Dort wird auch das Sprachmodell gespeichert, sodass die Daten innerhalb von Europa oder sogar innerhalb von Deutschland bleiben.

Brigitte Streibich: Das heißt, da wären Unternehmen dann auf der sicheren Seite. Gibt es denn eine Alternative zu einem Hosting mit Microsoft Azure, AWS, Google Cloud…?

Konrad Schreiber: Man kann natürlich auf Open-Source-Modelle setzen. Da bin ich sehr gespannt, vor allem auf Llama 2. Das ist gerade veröffentlicht worden, 150 Gigabyte groß und vom Funktionsumfang wirklich nah an GPT 3 dran. Das ist also ein Modell, bei dem man sich überlegen kann, es selbst in einer beliebigen Infrastruktur zu hosten. Das kann eine deutsche Infrastruktur sein, das kann aber auch ein deutscher Anbieter wie die Telekom Cloud sein. Man wird dort die Modelle nicht weiterentwickeln, nicht weiter trainieren, aber sehr wohl Inferenz drauf laufen lassen, sprich Anfragen hinschicken. Man braucht dafür auch ordentlich GPU-Power. Das sind spezielle Maschinen, die dort in Rechenzentren stehen müssen und das sind tatsächlich relativ teure Ressourcen, die man dann anzapfen muss. Ist aber eine Alternative und für einige Kunden sicherlich auch ein gangbarer Weg. Entwickelt wurde das Modell von Meta, es ist aber lizensiert und auch für den Business-Use

freigegeben worden. Es gibt da so ein paar kleine Details in der Lizenz Vereinbarung, aber grundsätzlich können Unternehmen das auch In-House nutzen.

Brigitte Streibich: Jetzt haben wir natürlich ganz viel über US-amerikanische Anbieter gesprochen. Gibt es denn im Bereich LLM und GPT auch Alternativen hier in Europa?

Konrad Schreiber: In Europa gibt es auch ein paar Firmen, die an Foundation Models und Sprachmodellen arbeiten. Die Business-Reife, dass wir es auch wirklich schnell in IT-Systeme einbauen und nutzen können, die sehe ich da jetzt gerade noch nicht. Das ist ein spannendes Thema für uns, weil ich schon sehe, dass wir unseren europäischen Markt auch stärken sollten – vor allem auch den europäischen AI-Markt. Die amerikanischen Firmen waren diesbezüglich einfach ein bisschen schneller und haben die Systeme rausgebracht, die leicht, schnell und ohne große Hürden zu nutzen sind und die gut funktionieren. Wenn man also ein qualitativ hochwertiges Modell nutzen möchte, kommt man aktuell um GPT 4 von OpenAI eigentlich nicht drum herum.

Brigitte Streibich: Wir haben ja eben schon kurz den EU AI Act angesprochen. Kannst du darauf nochmal eingehen?

Konrad Schreiber: Klar! Da geht es weniger um Datenschutz und Datensicherheit, sondern da stellt sich eher die Frage: Wie sieht die Zukunft der KI, der KI-Forschung und auch der angewandten KI im Unternehmen dann allgemein aus? Die Frage, die mir Kunden dann oft stellen, lautet meist: *„Kann es sein, dass so harte regulatorische Maßnahmen auf uns zukommen, dass wir diese Technologie der Large Language Models gar nicht mehr oder nur noch unter ganz großen regulatorischen Auflagen verwenden können?* Da wird sich sicherlich was tun. Wir müssen beim Einsatz der Technologie ein paar Aspekt beachten. Ganz viel davon ist aber auch gesunder Menschenverstand. Eine der Forderungen im EU AI Act ist, dass es menschliche Kontrollmechanismen gibt, dass es dort, wo KI-Systeme eingesetzt werden, Feedback Mechanismen gibt. Oder wenn ein KI-System eine Entscheidung trifft oder eine Information erzeugt, dass dann ein menschlicher User immer noch sagen kann, ob das, was gerade passiert, gut oder schlecht ist. Das ist ein laufender Prozess. Der ist noch nicht ganz durch und da wird auch gerade viel Lobbyarbeit betrieben. Regulatorien gibt es und gab es in der IT immer schon. Diese Technologie wird uns dadurch jetzt nicht wieder weggenommen, sondern wir werden da Lösungen finden, wie wir im Rahmen dieser Leitplanken diese Technologie nutzen können. Und ich gehe fest davon aus, dass zum einen sich auch die KI weiterentwickeln wird, dass das weiterhin möglich ist – auch in Europa. Und dass wir sie zum anderen auch im Business-Kontext einsetzen können, wenn wir eben bestimmte Dinge beachten.

Brigitte Streibich: Was genau müssen wir denn beachten?

Kondrad Schreiber: Ein paar Punkte habe ich gerade schon mal versucht zu benennen. Gerade diese Kontrollmechanismen, Feedback-Mechanismen wie Human-in-the-Loop für Endnutzer. Je nachdem, welche Kritikalität so ein AI-System dann hat, gibt es eine Einordnung in verschiedene Kritikalitätsbereiche. Das überlasse ich dann dem geneigten Hörer, sich das selbst mal durchzulesen, der AI Act ist ein recht umfangreiches Dokument. Es existieren unterschiedliche Regulatorien, welche Mechanismen man im Unternehmen je nach Kritikalitäts-Level einrichten muss.

Brigitte Streibich: Gibt es künstliche Intelligenz oder GPT ohne die Cloud, wenn Unternehmen auf keinen Fall in die Cloud möchten? Kann man das auch On-Premise nutzen?

Konrad Schreiber: Ja, ich denke schon. Möglich ist alles. Man kann sich natürlich ein Rechenzentrum selber einrichten, mit einigen modernen GPUs und darauf Open-Source-Modelle oder auch komplett eigene Modelle laufen lassen. Allerdings ist das ein wirklich sehr teures, ressourcenintensives Unterfangen. Da braucht man dann plötzlich nicht nur Data Scientists, die diese Modelle trainieren, sondern auch noch Ingenieure und Techniker, die sich um die Hardware dafür kümmern. Hyperscaler

bieten uns da schon durch die Infrastruktur auf jeden Reifegrad – egal, auf welchem Reifegrad ich so ein Modell einsetzen möchte – tolle Möglichkeiten, das zu nutzen. Und dann auch vertragliche Möglichkeiten im Business Kontext, wie etwa die Datensicherheit, sicherzustellen. Also ich würde jedem empfehlen, der auf Language Modus und auf KI setzen möchte, sich da einen guten Partner zu suchen und einen guten Cloud-Anbieter – wir haben da AWS, Google Cloud und Microsoft Azure als die großen drei – und sich da eben eine Partnerschaft zu suchen und die vertraglich so zu gestalten, dass es zu den eigenen Ansprüchen passt.

Brigitte Streibich: Falls man intern die gleiche Sicherheits-Infrastruktur aufbauen wollen würde, wie es jetzt die Hyperscaler machen, müsste man ja einen Riesenaufwand betreiben. Also im Zweifel sind die ja viel, viel sicherer, weil sie viel mehr Ressourcen und viel mehr Know-How aufwenden, als das jetzt ein Mittelständler machen könnte.

Konrad Schreiber: Absolut. Aus diesem Grund haben in den letzten Jahren nach und nach so gut wie alle Unternehmen gesagt: *„Wir lassen das mit den eigenen Rechenzentren, die sind viel zu teuer. Wir suchen uns eine Partnerschaft und schauen, dass wir dort Verträge aufsetzen, dass wir dort Vertrauen aufbauen."* Die Rechenzentren stehen in Deutschland oder es ist zumindest möglich, sich welche in Deutschland zu nehmen und darauf zu arbeiten.

Brigitte Streibich: Wie gewährleistet ihr von MaibornWolff – wenn ihr in die Unternehmen geht und mit den Kunden sprecht – die Datensicherheit bei LLM-Projekten?

Konrad Schreiber: Zum einen setzen wir wirklich sehr gerne auf unsere Partnerschaft mit Microsoft und verwenden die API für das OpenAI-Modell. Da kann man alle Logging-Mechanismen und auch Sicherheitsmechanismen deaktivieren. Sicherheitsmechanismen bedeutet jetzt nicht, dass die Modelle dadurch unsicher werden, sondern gewisse Guardrails, die den Missbrauch vermeiden sollen. Man kann das alles deaktivieren, wenn man die Berechtigung dazu hat und dann ist schon mal sichergestellt, dass die Firmen – wie Microsoft jetzt in diesem Fall – keinen Zugriff auf die Daten und auf die Anfragen, die an das Modell geschickt werden, bekommen.

Es gibt aber ein paar mehr Sachen, die man beachten muss. Nehmen wir mal das Beispiel von einem Callcenter-Usecase, das habe ich jetzt auch mit einer Kundin wieder besprochen. Wir haben Call Center Mitarbeiter, das sind angelernte Kräfte, die haben im Hintergrund eine Wissensdatenbank mit all unseren Problemen und Lösungen dazu. Und wenn dann Leute anrufen, dann können die sich aus dieser Wissensdatenbank bedienen und auf Lösungssuche gehen und Support geben. Könnte man das nicht durch einen Sprachcomputer ersetzen? Grundsätzlich wäre das sogar möglich. Allerdings muss man da beachten, dass man dann diesem Modell plötzlich Zugriff auf diese gesamte Wissensdatenbank gibt. Ein Angreifer von außen könnte jetzt dieses Callcenter dann nutzen oder missbrauchen, um sich Interna über das Unternehmen rauszuholen, die ein Mitarbeiter oder eine Mitarbeiterin im Call Center niemals weitergeben würde. Es würde sich etwa auch für einen Konkurrenten lohnen, einfach im Call Center anzurufen und mal nachzufragen: *„Sag mal, was sind denn eigentlich so eure Top Ten Probleme?"* Und das Sprachmodell würde natürlich zunächst mal mit Guardrails – mit solchen Leitplanken – ausgestattet sagen: *„Es tut mir leid, ich bin eine KI. Ich werde auf diese Frage jetzt nicht antworten."* Aber durch geschicktes Prompt Engineering, durch geschicktes Nachfragen – und Angreifer können da schon eine erstaunliche Energie entwickeln – kann man dann dieses Modell in eine Richtung drängen, dass es dann doch die gewünschte Antwort liefert. Und da muss man wirklich aufpassen.

Die Empfehlung an dieser Stellte lautet, wirklich auf das zu setzen, was auch vernünftig ist, nämlich weiterhin Menschen den Job machen zu lassen. Aber diesen Menschen – den Call-Center-Mitarbeitenden in dem Fall – kann man dann Assistentensysteme an die Hand geben, die dabei helfen, die Antworten möglichst schnell für den Kunden zu generieren, zu suchen, zu finden und dann allerdings mit menschlichem Verstand zu entscheiden. *„Ist das jetzt ein legitimer Anrufer? Ist das ein*

legitimes Interesse, was ich da habe, und welche Informationen gebe ich jetzt heraus und welche behalte ich für mich?“

Brigitte Streibich: Sehr spannend. Bleibt abzuwarten, wo die kriminelle Energie noch hinführt. Zum Abschluss würde ich dir noch gerne eine Heart-a-Tech typische Frage stellen. Wobei bekommst du Herzrasen, wenn es um das Thema Datensicherheit geht? Was würde bei dir einen Heart Attack auslösen?

Konrad Schreiber: Es wäre natürlich mit das Schlimmste, wenn unternehmensinterne Daten aufgrund von Unachtsamkeit irgendwo abfließen. Nicht richtig geschulte Mitarbeiter, die irgendwo ein öffentliches Tool verwenden. Und davon gibt es gerade wirklich viele, wie Sand am Meer. Wenn irgendwo interne, sensible Informationen raus fließen, lauert gerade dort die Gefahr. Da ist auch einfach viel Schulungs- und Compliance-Arbeit in den Firmen notwendig, um für die entsprechende Sicherheit zu sorgen. Und wenn so was doch mal passiert, dann ist es fast so unumkehrbar wie ein Bild oder Video, das es einmal ins Internet geschafft hat. Das kriegst du nicht mehr rausgelöscht. Das ist dann einfach da, rechtliche Regelsysteme hin oder her. Da sind dann einfach die technischen Möglichkeiten dergestalt, dass es fast unmöglich ist, solchen Content wieder loszuwerden. Da kommt das Herzrasen.

Brigitte Streibich: Du hast also nicht unbedingt Angst vor der KI, sondern eigentlich vor dem menschlichen Versagen. In dem Fall also, dass die Menschen unachtsam mit der neuen Technologie umgehen.

Konrad Schreiber: Ja, das ist ein weiteres spannendes Thema. Wo geht die KI-Entwicklung hin? Rennt die uns jetzt wirklich davon? Wird sie gefährlich für uns? An welcher Stelle passiert das? Dieses Thema können wir uns gerne auch mal für eine andere Podcast-Folge aufsparen. Im Moment ist tatsächlich – wie so oft – das größte Problem der Mensch. Das können wir aber auch sehr gut angehen. Durch Schulungen, durch Informationen, einfach durch gesunden Menschenverstand an ganz vielen Stellen.

Brigitte Streibich: Ich freue mich, wenn du ein drittes Mal vorbeikommst für diese Themen. Vielen Dank und bis bald.