MASTERARBEIT | MASTER'S THESIS

Titel | Title

KI-Halluzinationen: Wie helfen Warnsignale?

verfasst von | submitted by

Yusuf Kiziler BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 840
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Psychologie
Betreut von | Supervisor:	Univ.-Prof. Dipl.-Psych. Dr. Arnd Florack

Danksagung

Mein Dank gilt meiner Familie und Freunden, die mich auf dem Weg bis hier hin unterstützt haben. Auch dankbar bin ich der Stadt und Universität Wien, die mir das Studium und die Lebensqualität hier ermöglicht haben. Durch die zwei Jahre hier durfte ich mich weiterentwickeln und tolle Personen kennenlernen. Ich bedanke mich ebenfalls bei Univ.-Prof. Dr. Florack für die kompetente sowie zuvorkommende Unterstützung im gesamten Prozess der Masterarbeit. Auch dankbar bin ich für die finanzielle Unterstützung zur Durchführung der Studie, die mir ermöglicht wurde.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abstract

Durch die hohe Präsenz von KI-Chatbots besteht die Gefahr von häufigem KI-Halluzinieren: Dabei werden von KI-Chatbots falsche Informationen wiedergegeben und als wahr dargestellt. Das kann verheerende Folgen haben, beispielsweise Fehldiagnosen im Medizinkontext. Daher stellt sich in dieser Masterarbeit die Frage, wie Warnsignale effektiv gestaltet und formuliert werden können, damit Menschen solche KI-Halluzinationen häufiger erkennen und korrigieren. Das Extended Parallel Process Model von Witte (1992) diente als theoretische Grundlage zur Entwicklung eines solchen Warnsignals. Im Experiment wurden diesbezüglich KI-Halluzinationen mit einer vorprogrammierten ChatGPT-Seite simuliert. Dabei entstanden nach einem Between-Subject Design drei Bedingungen: Ein fehlendes Warnsignal, das allgemeine Warnsignal von ChatGPT selbst und ein Furchtappell mit Verhaltensempfehlung. ProbandInnen (n = 127) beantworteten mit den vorprogrammierten ChatGPT-Seiten zehn Wissensfragen, von denen fünf Antworten falsch vorprogrammiert wurden. Es wurde erwartet, dass der Furchtappell am stärksten wirkt und ein fehlendes Warnsignal am schwächsten. Ebenfalls wurde in Abhängigkeit der Warnsignale die Akzeptanz nach dem Technologieakzeptanzmodell erfasst: Es wurde erwartet, dass durch die stärkere Wirkung des Furchtappells die wahrgenommene Einfachheit und Nützlichkeit für ChatGPT am stärksten sinkt. Es zeigte sich für den Furchtappell im Gegensatz zu einem fehlenden Warnsignal ein signifikant höheres Erkennen und Korrigieren von KI-Halluzinationen. Eine mögliche Limitation der Studie ist es, dass das verwendete Warnsignal keine Furcht bei den ProbandInnenen ausgelöst haben könnte. Es wurden somit erste Impulse für die Forschung gesetzt und Empfehlungen für KI-Chatbot Betreibende formuliert.

Keywords: KI-Chatbots, KI, ChatGPT, KI-Halluzinationen, Warnsignale, Warntexte, Warnhinweise, Furchtappelle mit Verhaltensempfehlung, Technologieakzeptanzmodell.

KI-Halluzinationen: Wie helfen Warnsignale?

Einleitung

Generative künstliche Intelligenzen sind allgegenwärtig – besonders der im November 2022 durch OpenAI veröffentlichte Chatbot namens ChatGPT ist oft eingesetzt: Allein bis November 2023 wurden monatlich über 1,5 Milliarden BesucherInnen generiert, was es zu einem der meistbesuchten Webseiten weltweit macht (Frąckiewicz, 2023; OpenAI, 2023; Shewale, 2023). Das ist nicht verwunderlich, denn ChatGPT hat durch die implementierte künstliche Intelligenz (KI) vielfältige Funktionen: Mit dessen Hilfe können sekundenschnell Programmcodes erstellt sowie umfangreiche Texte verfasst werden und vieles mehr (Rahman & Watanobe, 2023). Jedoch gibt es auch eine klare Gefahr solcher KI-Chatbots, und zwar das Erfinden – auch genannt *Halluzinieren* – von falschen Informationen: Dabei werden die Informationen selbstbewusst als wahr dargestellt, obwohl diese teilweise oder gänzlich falsch sind (Alkaissi & McFarlane, 2023). Aufgrund der hohen Verbreitung von KI-Chatbots kann das fatal sein. In der Praxis wurde zum Beispiel eine hohe Frequenz der halluzinierten Referenzen von ChatGPT von bis zu 93 % im Medizinkontext erfasst, was die hohe Gefahr aufzeigt (Bhattacharyya et al., 2023). Daraus ergibt sich in dieser Arbeit die wichtige Frage: Wie kann die Wahrscheinlichkeit erhöht werden, dass Menschen KI-Halluzinationen vermehrt erkennen und korrigieren, ohne jedoch die Effizienz und den Zugang der KI-Chatbots zu limitieren?

Eine naheliegende Möglichkeit wäre es, vor KI-Halluzinationen zu warnen. So gibt OpenAI als Warnsignal schriftlich an, dass ihr KI-Chatbot Fehler machen kann und wichtige Informationen geprüft werden sollten (OpenAI, 2023, siehe Anhang A). Solch ein Warnsignal warnt jedoch lediglich vor dem Auftreten von KI-Halluzinationen, ohne auf konkrete Gefahren hinzuweisen, wodurch eine hohe Diskrepanz zwischen der Warnung und den möglichen Gefahren bestehen könnte. Menschen nutzen möglicherweise mehrere

Informationsquellen, um den Wahrheitsgehalt von Informationen zu überprüfen, zum Beispiel die Einfachheit der Verarbeitung der Informationen (Stanley et al., 2022). KI-Chatbots geben Inhalte sehr flüssig und leserfreundlich wieder, dadurch könnte folglich die Gefahr bestehen, dass Menschen KI-Halluzinationen eher als wahr annehmen.

Welche Warnsignale könnten also zuverlässiger vor den Gefahren von KI-Halluzinationen warnen? In dieser Masterarbeit wird dafür ein Warnsignal nach dem *Extended Parallel Process Model* (EPPM) von Kim Witte (1996) formuliert: Dabei sollten Furchtappelle (hier Warnsignale) persuasiv eine mögliche Gefahr mitteilen und es sollten attraktive Verhaltensempfehlungen angegeben werden, mit der diese Gefahr effizient aufgelöst werden kann (Xiao & Benbasat, 2015): Durch die kommunizierte Gefahr könnte eine mögliche Gefahrenkontrolle beim empfangenden Gegenüber aktiviert werden, um die Gefahr abzuwenden. (Witte, 1996). In der Forschung wurde die Wirksamkeit des EPPM vermehrt aufgezeigt (Tannenbaum et al., 2015; Zonouzy et al., 2018). Jedoch wurden noch nicht Warnsignale nach dem EPPM bei KI-Halluzinationen untersucht – die zu erforschende Frage dieser Masterarbeit ist also, ob Furchtappelle mit Verhaltensempfehlungen nach Witte (1996) bei Menschen dazu führen, KI-Halluzinationen vermehrt zu erkennen und zu korrigieren als bei allgemeineren Warnsignalen wie von ChatGPT selbst und als bei keinen Warnsignalen.

Wenn jedoch solch ein Warnsignal dazu führen könnte, dass Menschen KI-Halluzinationen häufiger erkennen und korrigieren, könnte aufgrund des erhöhten Korrekturaufwandes die wahrgenommene Effizienz des KI-Chatbots geringer bewerten werden. Um das zu überprüfen, wird in dieser Masterarbeit das Technologieakzeptanzmodell (TAM) nach Davis (1989) für neue Technologien herangezogen: Durch einen Fragebogen werden damit die wahrgenommene Effizienz und die Einfachheit der Nutzung von ChatGPT in Abhängigkeit mit den Warnsignalen erfasst.

Theoretischer Hintergrund

Ein wichtiger Aspekt von KI-Halluzinationen ist das Glauben beziehungsweise fehlende Hinterfragen von neuen Informationen: Nach dem *Truth-Bias* - abgeleitet von der *Wahrheits-Default-Theorie* - sind wir nämlich eher dazu geneigt, neue Informationen zunächst als wahr anzunehmen und erst bei einer kognitiven Anstrengung als falsch zu kennzeichnen (McCornack & Parks, 1986; Gilbert et al., 1993; Levine et al., 1999). Es könnte also ein heuristischer Prozess dazu beitragen, dass Menschen neue Informationen aufgrund geringerer kognitiver Anstrengung zunächst als wahr annehmen, wobei erst bei Verdacht diese hinterfragt werden (Stanley et al., 2022). Bezüglich KI-Halluzinationen könnte auch die heuristische Einfachheit der Verarbeitung neuer Informationen relevant sein: Einfach verarbeitbare Informationen könnten nämlich eher als wahr angenommen werden (Stanley et al., 2022). Vor allem zeigt das der sogenannte *Truth-Effect*: Bereits früher verarbeitete Aussagen werden vermehrt als wahr angenommen als unbekannte (Pohl, 2016). Besonders im Kontext von KI-Chatbots ist der Truth-Bias gefährlich: Gerade weil eine Person solche KI-Chatbots benutzen könnte, um schnell Informationen zu erhalten, könnte es nach dem Truth-Effekt sein, dass die Person die Informationen noch weniger hinterfragt und somit weniger kognitive Anstrengungen anwendet. Außerdem geben KI-Chatbots technisch die am wahrscheinlichst passendste Antwort an – mit gängigen Schreibformen und eher benutzerfreundlichem und flüssigem Text – welche nach dem Truth-Effect einfacher zu verarbeiten sind und somit eher als wahr angenommen werden könnten (Greifeneder et al., 2020). Daher ist es bei KI-Chatbots wichtig, vor den Gefahren von KI-Halluzinationen effizient zu warnen, jedoch ohne dessen Nutzen einzuschränken. Oft sind die Gefahren von KI-Halluzinationen nicht direkt ersichtlich, daher könnten Warnsignale dazu führen, dass Menschen die Gefahren konkret einschätzen.

Forschung bezüglich Warnsignale und Falschinformationen

Warnsignale werden bei Fehlinformationen (unbeabsichtigte falsche Informationen – also auch bei KI-Chatbots) und Desinformationen (intentionale falsche Informationen) oft als Prebunkingstrategie angewendet (Ecker et al., 2022). Prebunkingstrategien sollen präventiv vor möglichen falschen Informationen schützen, also bevor diese Informationen vernommen werden. Debunkingstrategien wiederum sollen erst nach dem Vernehmen der falschen Informationen diese richtig stellen und Maßnahmen umsetzen, wie mit solchen falschen Informationen umgegangen werden soll (Ecker et al., 2022). Für beide Strategien zeigen sich aus der Forschung positive sowie fehlende Effekte (Ecker et al., 2022): Zum Beispiel zeigten Ecker et al. (2010), dass allgemeiner gehaltene Warnsignale bei falschen Informationen eher weniger effektiv sind (siehe Anhang A für den allgemeineren Warnsignal von Ecker et al., 2022). ChatGPT benutzt ein ähnliches, allgemeiner gehaltenes Warnsignal (siehe Anhang A, OpenAI, 2023). Ein Literaturreview von Martel und Rand (2023) führt wiederum aus, dass Warnsignale als Prebunkingstrategie im Kontext von Desinformationen allgemein effektiv sein können, wenn auch nur moderat.

Im Kontext von Desinformationen zeigen Warnsignale mit der Zeit ebenfalls verringerte Effekte: In einer Studie von Street und Richardson (2014) wurde gezeigt, dass der Truth-Bias in Kontexten vorhanden sein kann, in denen Lügen und somit Desinformationen erwartet werden müssten. Street und Richardson (2014) haben den ProbandInnen Videos von Personen gezeigt, die über ihren Urlaub berichtet haben. Dabei wurde in verschiedenen Bedingungen mitgeteilt, dass entweder 20 %, 50 % oder 80 % der Personen aus den Videos die Wahrheit sagen würden. Die Probanden sollten jederzeit bewerten, ob sie denken, dass die Personen die Wahrheit sagen würden, dabei gaben sie am Ende der Videos ein abschließendes Urteil. Zwar haben sie den Personen der 20 % Bedingung – verglichen mit der 50 % und 80 % Bedingung – anfangs weniger geglaubt, mit der Zeit haben sie ihnen

jedoch immer mehr geglaubt, was aufzeigt, dass Warnsignale mit der Zeit ihren Effekt verlieren können (Street & Richardson, 2014). Ein weiterer Grund für die Ineffektivität bestimmter Warnsignale könnte sein, dass Personen sich mit der Zeit an diese erinnern und somit kognitive Anstrengungen aufzeigen müssten, was vor allem bei unauffälligen Warnsignalen schwieriger ist (Stanley et al., 2022). Kritisch bei bestimmten Warnsignalen über mögliche falsche Informationen ist auch, dass diese als Debunkingstrategie erst am Ende der Informationen gegeben werden: Diese erscheinen also erst nach der Verarbeitung der Informationen, welche nach dem Truth-Bias also bereits möglicherweise als wahrhaftig verarbeitet wurden – die rückwirkende Einstufung der Informationen als mögliche Fehlinformation ist somit eher schwächer und möglicherweise schneller vergessen als die falschen Informationen selbst (Stanley et al., 2022).

Es existieren also viele mögliche Gründe, weshalb Falschinformationen geglaubt werden und weshalb Warnsignale nicht immer effektiv wirken: Zum einen eben durch psychologische Effekte wie dem Truth-Bias und Truth-Effect, zum anderen aber auch durch Gewöhnungseffekte, durch unauffällige und unpräzise Warnsignale, aber auch durch fehlendes Kommunizieren einer möglichen Gefahr (Ecker et al., 2022; Gilbert et al., 1993; Martel & Rand, 2023; Stanley et al., 2022; Street & Richardson, 2014).

Furchtappelle mit Verhaltensmepfehlungen

Um Menschen vor KI-Halluzinationen zu warnen, wird in dieser Masterarbeit ein Warnsignal als Furchtappell mit Verhaltensempfehlung nach dem EPPM formuliert. Witte (1996) entwickelte ihr Modell aus bestehenden Theorien der Furchtappellforschung: Die *Triebtheorie* nach Hovland et al. (1953), das *Duale Prozessmodell* nach Leventhal (1971) und die *Schutzmotivationstheorie* nach Rogers (1975). Dabei definiert Witte (1992) die Variablen aus dem EPPM wie folgt: *Furchtappelle* werden als furchtinduzierende Botschaften beschrieben, welche selbst eine klare Gefahr suggerieren, zum Beispiel gesundheitliche

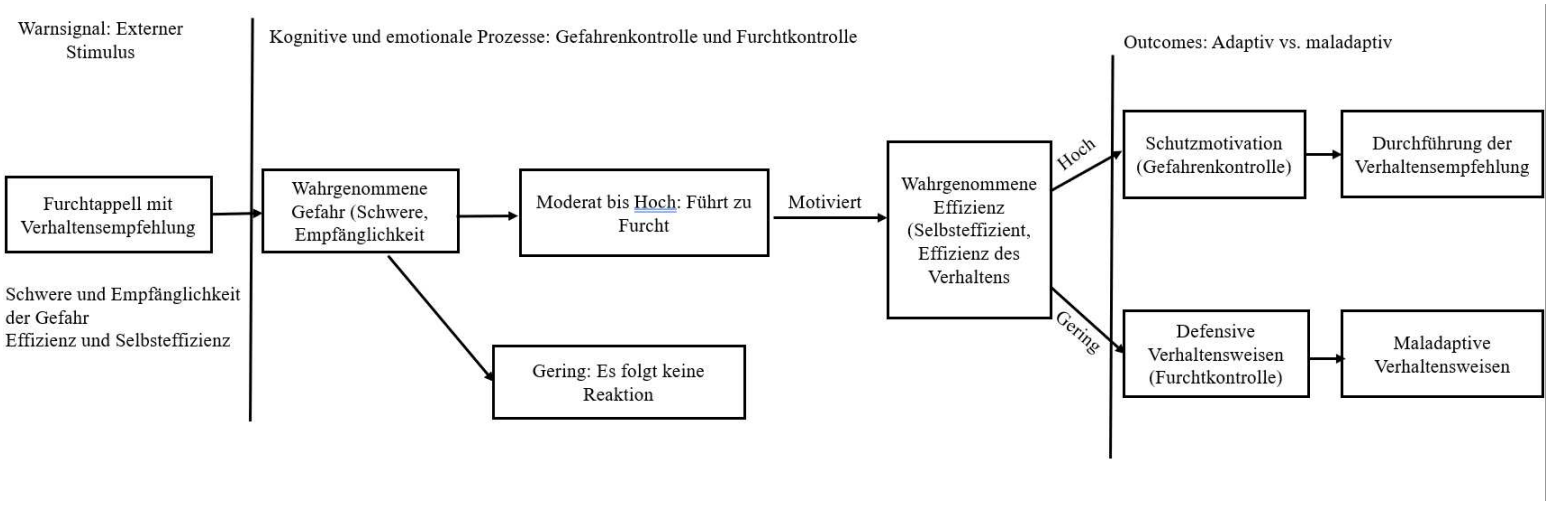
Schäden. *Furcht* selbst definiert Witte als negative Emotion mit einem hohen Zustand an Erregung – ausgelöst durch eine wahrgenommene Gefahr, die potenziell bedrohlich ist (Witte, 1992). Die furchtinduzierende *Gefahr* wird von Witte (1992) als ein externer Stimulus definiert, welches unabhängig von der Person existiert: Sobald die Person nun die Gefahr kognitiv beurteilt, könnte möglicherweise Furcht ausgelöst werden. Das hängt jedoch davon ab, wie schwer die Gefahr wahrgenommen wird und wie wahrscheinlich die Gefahr auf die Person zutreffen könnte. Definiert ist das als die *wahrgenommene Schwere (Perceived Severity)* und *wahrgenommene Empfänglichkeit (Perceived Susceptibility)* der Gefahr (Witte, 1992). Ferner wird die *Effizienz* der Verhaltensempfehlung definiert durch die Effektivität, Machbarkeit und Einfachheit, welches empfohlen wird, um die Gefahr abzuwenden (Witte, 1992; Witte, 1996). Die *wahrgenommene Effizienz (Perceived Efficacy)* wird von der Person kognitiv gebildet durch: *Die wahrgenommene Effizienz des Verhaltens (Perceived Response Efficacy)*, die den Glauben der Person beschreibt, wie sehr die Verhaltensempfehlung die Gefahr abwenden könnte und die *wahrgenommene Selbsteffizienz (Perceived Self-Efficacy)*, die die eigenen Kompetenzen bezüglich der Durchführung der Verhaltensempfehlung bewertet. Als Ergebnisvariablen ergeben sich dadurch die *Akzeptanz des Warnsignals (Message Acceptance),* definiert als die Akzeptanz der Gefahr und Durchführung der Verhaltensempfehlung (Witte, 1992). Jedoch gibt es auch maladaptive Verhaltensweisen, auch genannt *defensive Verhaltensweisen (Defencive Avoidance)*, die das Ignorieren oder Unterdrücken der Gefahr beschreibt (Witte, 1992; Witte, 1996).

All diese Komponenten ergeben das EPPM (Witte, 1992): Zunächst wird ein Furchtappell mit der vermittelten Gefahr wahrgenommen, diese selbst enthält die Schwere und Empfänglichkeit der Gefahr, aber auch die Effizienz der Verhaltensempfehlung und durch die Person selbst beurteilt die Selbsteffizienz diesbezüglich. Zunächst wird die wahrgenommene Gefahr nach der Schwere und Empfänglichkeit beurteilt, sobald diese

schwach ausfällt, wird das Warnsignal nicht weiter verarbeitet und zeigt somit keine Wirkung. Wenn diese jedoch moderat bis stark ausfällt, wird Furcht ausgelöst, diese motiviert zu einer weiteren kognitiven Bewertung: Die Effizienz der Verhaltensempfehlung sowie die Selbsteffizienz, diese durchzuführen (Witte, 1992). Wird die Effizienz nun als hoch bewertet, folgt der Prozess der Gefahrenkontrolle (Danger Control): Nach der Schutzmotivation steigt nun das Bedürfnis, die Verhaltensempfehlung durchzuführen, um die Gefahr abzuwenden (Witte, 1992). Wenn nun die wahrgenommene Schwere und Empfänglichkeit der Gefahr hoch ist, die wahrgenommene Effizienz der Verhaltensempfehlung oder die Selbsteffizienz jedoch gering, folgt die Furchtkontrolle (Fear Control): Daraus ergeben sich maladaptive Verhaltensweisen, um mit der Furcht umzugehen, ohne jedoch die Gefahr abzuwenden (Witte, 1992). Die wahrgenommene Schwere und Empfänglichkeit der Gefahr führt letzlich dazu, wie intensiv auf das Warnsignal reagiert wird. Die wahrgenommene Effizienz der Verhaltensempfehlung und Selbsteffizienz bestimmt wiederum, ob eine Gefahrenkontrolle und somit ein Abwenden der Gefahr stattfindet oder eine Furchtkontrolle und somit maladaptive Verhaltensweisen stattfinden (Witte, 1992). Für eine optimale Wirkung des Warnsignals sollten also die wahrgenommene Schwere und Empfänglichkeit der Gefahr sowie die Effizienz der Verhaltensempfehlung und Selbsteffizienz möglich hoch sein, um die Gefahrenkontrolle und somit Schutzmotivation zu aktivieren (Wiite, 1992). Dieser Prozess ist in Abbildung 1 nach Witte (1992) dargestellt.

Abbildung 1

Extented Parallel Process Modell nach Kim Witte (1992)



Wichtig zu erwähnen ist jedoch, dass eine Objektivierung der einzelnen Komponenten nicht möglich ist, denn individuelle Erfahrungen, kulturelle Hintergründe, emotionale Zustände und vieles mehr führen zu verschiedenen Bewertungen von Gefahren und Verhaltensempfehlungen (Witte, 1992). Zum Beispiel kann die Verhaltensempfehlung, mögliche falsche Informationen im Internet zu recherchieren, bei Personen als wenig effizient bewertet werden, die kaum Erfahrungen mit gängigen Suchmaschinen haben und somit möglicherweise eine geringe Selbsteffizienz diesbezüglich aufweisen.

Zwar existiert eine Studie zu Warnsignalen und KI-Chatbot-Halluzinationen mit positiven Ergebnissen (siehe Nahar et al., 2024), jedoch ist dabei das Warnsignal stark an ChatGPT angelehnt und vermittelt somit keine klare Gefahr. Mit dieser Studie soll die Forschung diesbezüglich um Furchtappelle mit Verhaltensempfehlung nach dem EPPM fortgeführt werden, da diese eine aktuelle Theorie der Furchtappellforschung darstellt und auch in verschiedenen Studien positive Wirkungen aufzeigt (Basil et al., 2013; Birmingham et al., 2015; Gharlipour et al., 2015; Roberto et al., 2019; Tannenbaum et al., 2015; Zonouzy et al., 2018). Aus der vorgestellten Theorie und damit der vermuteten effektiveren Wirkung von Furchtappellen nach dem EPPM ergibt sich die erste Hypothese:

Hypothese 1a: Furchtappelle mit Verhaltensempfehlung führen zu einem erhöhten Erkennen und Korrigieren von KI-Chatbot-Halluzination als keine Warnsignale und allgemeine Warnsignale.

Entwicklung eines Warnsignals für die Studie

Es existieren keine Untersuchungen zu Furchtappellen mit Verhaltensempfehlungen bei generativen Chatbots und dem relativ neuen Phänomen des Halluzinierens – nach dieser Überlegung könnte also in diesem Kontext mit der Furchtappellforschung ein simples Warnsignal wie der von ChatGPT effektiver gestaltet werden, wenn mögliche furchtinduzierende Konsequenzen von halluzinierten Inhalten aufgegriffen werden (zum Beispiel die Verbreitung falscher Informationen) und dass konkret durch die Verwendung von alternativen Medien wie Google die Inhalte schnell und effizient überprüft werden können. Dennoch ist es noch wichtig, Folgendes zu beachten: Betrachtet man nämlich das EPPM, wird dabei als erster Schritt vorausgesetzt, dass das Warnsignal zuverlässig wahrgenommen wird, jedoch bildet das nicht immer die Realität ab. Betrachtet man zum Beispiel den offiziellen Warnsignal von ChatGPT, könnte hierbei die Gefahr bestehen, dass dieses erst gar nicht wahrgenommen wird, weil es sehr klein und unauffällig präsentiert wird. Nach Conzola und Wogalter (2001) gibt es bei der Formulierung von Warnsignalen nämlich Folgendes zu beachten: Warnsignale müssen die Aufmerksamkeit der Person wecken – diese müssen sich dementsprechend vom Hintergrund abheben und mit anderen Signalen konkurrenzfähig sein. Außerdem muss die Aufmerksamkeit der Personen beibehalten werden, es darf folglich auch nicht zu viel Text verwendet werden. Ferner ist die Verständlichkeit relevant, das Warnsignal soll also so einfach, kurz und prägnant wie möglich formuliert sein. Warnsignale müssen zudem zum Glauben der Personen passen: also keine absurden Behauptungen vorzeigen (Conzola & Wogalter, 2001). Dementsprechend könnte zum Beispiel ein textliches Warnsignal rot markiert und größer hervorgehoben

präsentiert werden – mit nicht zu viel Text sowie mit präzisen und weniger absurden Inhalten.

Ebenfalls wird das allgemeiner gehaltene Warnsignal von ChatGPT getestet, welches kein Furchtappell und keine Verhaltensempfehlung enthält (Siehe Anhang A). In dieser Masterarbeit ist mit einem allgemeinen Warnsignal gemeint, dass diese keine klare Gefahr kommuniziert und keine Verhaltensempfehlung enthält. Nach der Empfehlung von Wogalter und Conzola (2001) wird das Warnsignal jedoch rot markiert und größer dargestellt, um zu vermeiden, dass es nicht wahrgenommen oder aus Gewohnheit ignoriert wird. Dabei wird aufgrund der präsentierten Forschung und der empfohlenen Hervorhebung nach Wogalter und Conzola (2001) des allgemeineren Warnsignals folgende zweite Hypothese aufgestellt:

Hypothese 1b: Allgemeine Warnsignale führen zu einem erhöhten Erkennen und Korrigieren von KI-Chatbot-Halluzinationen als keine Warnsignale.

Akzeptanz von KI-Chatbots

Es ist zudem möglich, dass durch ein gesteigertes Erkennen und Korrigieren von KI-Halluzinationen zugleich die wahrgenommene Effizienz und Einfachheit der Nutzung des KI-Chatbots reduziert wird: Müssen AnwenderInnen nämlich zusätzlich recherchieren und korrigieren, entsteht ein Mehraufwand, was Zeit und Ressourcen kostet. Diese zusätzliche Anstrengung durch effektivere Warnsignale könnte also dazu führen, dass KI-Chatbots als weniger effizient und bedienungsfreundlich bewertet werden. Um das zu erfassen, wird in dieser Masterarbeit das Technologieakzeptanzmodell (TAM) nach Davis (1989) herangezogen: Das Modell beschreibt die Akzeptanz von neuen Technologien mithilfe der *wahrgenommenen Nützlichkeit* - wie sehr die Technologie bezüglich der eigenen Leistung hilfreich ist – sowie der *wahrgenommenen Einfachheit der Nutzung* – wie frei von Anstrengung die Nutzung der Technologie ist (Davis, 1989, Shuhaiber & Mashal, 2019). Nach dem TAM ist für die Akzeptanz neuer Technologien also relevant, dass diese einfach

zu bedienen sind und effizient funktionieren. In dieser Mastarbeit wird angenommen, dass ein Furchtappell mit Verhaltensempfehlung zu einem höheren Erkennen und Korrigieren von KI-Halluzinationen führt, was jedoch Mehraufwand bei den AnwenderInnen bedeuten könnte und nach dem TAM somit möglicherweise die Akzeptanz bezüglich des KI-Chatbots senken könnte. Daraus ergeben sich folgende Hypothesen:

Hypothese 2a: Furchtappelle mit Verhaltensempfehlung führen zu einer geringeren Akzeptanz von KI-Chatbots als keine Warnsignale und simple Warnsignale.

Da der allgemeine Warnsignal nach Hypothese 1b dennoch – auch aufgrund der Hervorhebung – eine Wirkung zeigen sollte, wenn auch theoretisch schwächer als der Furchtappell mit Verhaltensempfehlung, sollte ebenfalls die Akzeptanz gesenkt werden im Vergleich zu keinem Warnsignal:

Hypothese 2b: Allgemeine Warnsignale führen zu einer geringeren Akzeptanz von KI-Chatbots als keine Warnsignale.

Hypothesen zusammengefasst

Zentrale Forschungsfrage: Führen Furchtappelle mit Verhaltensempfehlung bei KI-Chatbots dazu, halluzinierte Inhalte vermehrt zu erkennen und zu korrigieren als bei allgemeineren Warnsignalen wie von ChatGPT selbst und als bei keinen Warnsignalen?

Weitere Forschungsfrage: Wird durch Furchtappelle mit Verhaltensempfehlung die Akzeptanz (hier primär Einfachheit und Effizienz der Nutzung) für KI-Chatbots mehr reduziert als durch allgemeine Warnsignale und keine Warnsignale?

Hypothese 1a: Furchtappelle mit Verhaltensempfehlung führen zu einem erhöhten Erkennen und Korrigieren von KI-Chatbot-Halluzination als keine Warnsignale und allgemeine Warnsignale.

Hypothese 1b: Allgemeine Warnsignale führen zu einem erhöhten Erkennen und Korrigieren von KI-Chatbot-Halluzinationen als keine Warnsignale.

Hypothese 2a: Furchtappelle mit Verhaltensempfehlung führen zu einer geringeren Akzeptanz von KI-Chatbots als keine Warnsignale und simple Warnsignale.

Hypothese 2b: Allgemeine Warnsignale führen zu einer geringeren Akzeptanz von KI-Chatbots als keine Warnsignale.

Beitrag zur Forschung und für Betreibende von KI-Chatbots

Mithilfe der Konzipierung eines Furchtappells mit Verhaltensempfehlung soll die Forschung bezüglich der Verbreitung von Fehlinformationen durch allgegenwärtig präsente und immer häufiger eingesetzte generative KI-Chatbots angeregt werden. Ebenfalls sollen erste Daten in diesem Kontext geliefert werden, die zum einen verschiedene Warnsignale bei KI-Halluzinationen erforschen und zum einen die Akzeptanz und Intention der Weiternutzung erfassen – das heißt, ob solche Warnsignale zur Reduzierung der Nutzung führen. Dadurch könnten praktische Implikationen für Betreibende gegeben werden und auch erste Impulse für die Weiterentwicklung und weitere Erforschung von Warnsignalen gesetzt werden, da das Phänomen des Halluzinierens technisch nicht lösbar ist (Xu et al., 2024).

Methode

Vor der Durchführung der Studie erfolgte eine Präregistrierung erfolgte mit der Nummer *#168800* und dem Link *https://aspredicted.org/2dq9r.pdf*. Dazu wurde die Online-Plattform AsPredicted von der Wharton School der Universität Pennsylvania verwendet (*Home | AsPredicted*, o. D.). Ebenfalls hat das *Departmental Review Board* der Universität Wien die Eignung des Experiment vor der Durchführung überprüft und genehmigt mit der DRB-Nummer *2024/S/008*.

Design

Ein Between-Subject-Design wurde angewandt: ProbandInnen wurden zufällig einer von drei Bedingungen zugeordnet. Zur Randomisierung der Bedingungen wurde durch einen Zufallsgenerator eine Zufallsreihenfolge von eins bis drei generiert (siehe Anhang B). Die

ProbandInnen bekamen als unabhängige Variable eine Variation der Warnsignale auf einer vorprogrammierten inoffiziellen ChatGPT-Seite. Dabei haben alle ProbandInnen denselben Fragebogen ausgefüllt und zehn offene Wissensfragen über eigenständige Recherchen beantwortet, wobei fünf Antworten bestimmter Fragen über die vorprogrammierte ChatGPT-Seite gezielt falsch wiedergegeben wurden - die Beantwortung jener fünf Fragen wurde als abhängige Variable behandelt. Im Anschluss wurden Fragen aus dem TAM zur Beurteilung der Akzeptanz von ChatGPT als weitere abhängige Variable gestellt.

Stichprobe und Teststärkenanalyse

Um für die statistische Analyse einen angemessenen Stichprobenumfang zu ermitteln, wurde eine A-priori-Teststärkeanalyse mithilfe des Statistikprogramms G*Power durchgeführt (Version 3.1.9.6, Faul et al., 2009): Dabei wurde die hauptsächlich zu untersuchende Between-Subject Hypothese 1a herangezogen: Für die Teststärkenanalyse wurde eine einfaktorielle Varianzanalyse ohne Messwiederholung und ohne Bezug auf Interaktionen verwendet mit drei Bedingungen, einer in der psychologischen Forschung üblichen Power von 0,8, einem Alpha-Niveau von 0,05 und einer erwarteten mittleren Effektstärke von Cohens *f* von 0,25 (Muller & Cohen, 1989). Eine mittlere Effektstärke wurde ausgewählt, da bezüglich der Forschung von Warnsignalen bei KonsumentInnen (z.B. von Tabak) nach einer Metaanalyse von Tannenbaum et al. (2015) sowie Purmehdi et al. (2017) kleine und mittlere Effektstärken gefunden wurden. Die Teststärkenanalyse für eine mittlere Effektstärke ergibt einen Strichprobenumfang von $N = 159$.

An zwölf Erhebungstagen vom 13.05.2024 bis zum 24.06.2024 haben an der Studie 139 ProbandInnen teilgenommen: Dabei wurden 12 ProbandInnen von der statistischen Analyse ausgeschlossen. Bezüglich der Instruktion wurden je ProbandIn vier Verständnisfragen erfasst, ein/e ProbandIn hat eine Verständnisfrage falsch beantwortet, weshalb sie aus der statistischen Datenanalyse ausgeschlossen wurde. Ebenfalls wurden

sieben ProbandInnen ausgeschlossen, die bei mindestens einer der zehn Frage ChatGPT nicht als Erstquelle angegeben haben und somit der Instruktion nicht gefolgt sind. Zudem wussten drei ProbandInnen bereits bestimmte Antworten der Wissensfragen, sie wurden dementsprechend ebenfalls ausgeschlossen. Ein/e ProbandIn beendete die Studie unüblich schnell und gab als Begründung an, dass sie es eilig hätte, daher wurde diese Person ebenfalls ausgeschlossen.

Die finale Stichprobe betrug damit $n = 127$. Das durchschnittliche Alter betrug dabei 24 Jahre mit einer Standardabweichung von 5,4 Jahren. Von den 127 ProbandInnen waren 95 weiblich, 31 männlich und eine Person machte keine Angabe zum Geschlecht. ProbandInnen wurden überwiegend durch ein universitätsinternes System rekrutiert, dem Laboratory Administration for Behavioral Sciences (Labs): Studierende sowie Forschende der Psychologie können sich in Labs registrieren und somit psychologische Studien verwalten. Es gab für die Teilnahme an der Studie drei Kredits sowie die Möglichkeit zur Teilnahme an einem Gewinnspiel für einen 20 Euro Amazon-Gutschein. Außerdem gab es eine Studienvariante mit einer monetären Belohnung von zehn Euro ohne die Vergabe von Kredits und ohne ein inkludiertes Gewinnspiel. Die Rekrutierung erfolgte somit über das Labs E-Mailsystem ab dem 03.05.2024 bis zum 19.06.2024, wobei automatisiert an verschiedenen Tagen insgesamt 11245 E-Mails verschickt wurden. Dabei wurden Ansprachen verwendet, die eine kurze Beschreibung der Studie sowie die Belohnungen der Teilnahme beinhalteten (siehe Anhang H). Durch das Labs-System wurde letzlich die Anwesenheit und die Vergabe der Kredits koordiniert. Zusätzliche Daten wie die Erhebungstage oder die Versuchsleitung sind im Versuchsleiterprotokoll aufgeführt (siehe Anhang I).

Unabhängige Variable: Warnsignale innerhalb der ChatGPT-Seite

Der Fragebogen ist für alle Bedingungen identisch gewesen, lediglich das Warnsignal bei der ChatGPT-Seite unterschied sich zwischen den drei Bedingungen: Die

Kontrollbedingung beinhaltete kein Warnsignal, die Bedingung *offiziell* beinhaltete den offiziellen Warnsignal von ChatGPT (OpenAI, 2023), welches nach der Empfehlung für die Formulierung von Warnsignalen nach Wogalter und Conzola (2001) etwas größer hervorgehoben und rot markiert wurde. Die Bedingung *Furchtappell* beinhaltete ein Furchtappell mit Verhaltensempfehlung nach Witte (1992). Im Anhang C sind die verschiedenen Bedingungen und damit die verschiedenen ChatGPT-Seiten abgebildet.

Abhängige Variable: Anzahl korrekter Antworten

Um zu quantifizieren, ob ProbandInnen mit der Wahrnehmung verschiedener Warnsignale die falsch wiedergegebenen Antworten durch ChatGPT verschieden behandelten und somit die Hypothesen H1a und H1b zu untersuchen, wurden die Antworten auf diese fünf Fragen manuell ausgewertet und entsprechend codiert in falsch (0) und richtig (1): Für jede Person gab es also eine Gesamtpunktzahl zwischen null bis fünf. Rechtschreib- und Grammatikfehler sowie leichte Abweichungen und Synonyme der Antworten wurden nicht als falsch gewertet, da ersichtlich wurde, ob die ProbandInnen über die Antwortwiedergabe von ChatGPT hinaus korrekt recherchiert haben, zum Beispiel mit Google. Es wurde nämlich zu jeder Frage zusätzlich gefragt, welche Quelle als finale Quelle – und somit als die Quelle für die endgültige Antwort – verwendet wurde, um die manuelle Auswertung der Antworten zu vereinfachen und besser nachvollziehen zu können.

In Anhang D sind alle Fragen und die korrekten sowie falschen Antwortwiedergaben durch ChatGPT aufgelistet. Bei den Fragen handelt es sich um nicht alltagsnahe Wissensaufgaben aus verschiedenen Themengebieten wie der Physik und Geographie mit eindeutigen, durch Internetrecherchen schnell auffindbaren Antworten. Es wurde über verschiedene Zeiträume, Computer, Internetzugänge und Browser überprüft, ob sich die Antworten der Fragen durch die Verwendung gängiger Suchmaschinen schnell finden lassen und auch mit Google und Bing dementsprechend bestätigt, sodass die ProbandInnen bei der

Nutzung solcher Suchmaschinen also schnell die Antworten hätten finden können. Um also Punkte für falsch wiedergegebene ChatGPT-Antworten zu erhalten, waren zwei Schritte notwendig: 1) Die Handlung, die Fragen nach der Antwortwiedergabe von ChatGPT noch mal über alternative Quellen zu recherchieren (Fakt-Checking) und 2) die korrekten Antworten zu finden sowie einzutragen. Somit ist die abhängige Variable zusammengefasst das Recherchieren und Finden der korrekten Antworten, oder anders formuliert: Dass Erkennen und Korrigieren von KI-Halluzinationen. Es gab auch wenige Fälle, in denen die falsch wiedergegebenen ChatGPT-Antworten zwar im Internet recherchiert wurden, jedoch mit falschen Ergebnissen, diese Angaben wurden also dennoch als falsch gewertet.

Abhängige Variable: Akzeptanz

Um die Hypothesen 2a und 2b zu testen, wurde die Akzeptanz von ChatGPT mithilfe des Technologieakzeptanzfragebogens gemessen (Davis, 1989). Dazu wurden die jeweils sechs Items für die Variablen *Wahrgenommene Nützlichkeit* und *Wahrgenommene Einfachheit* der *Nutzung* ins Deutsche übersetzt und in den Unipark-Fragebogen integriert (aus Davis, 1989). Zusätzlich wurde nach bestehender Forschung vier Items für die Einstellung gegenüber ChatGPT und drei für die Intention der Nutzung von ChatGPT in den Fragebogen integriert, wobei es sich jeweils um siebenstufige Items einer Likert-Skala von extrem unzutreffend bis extrem zutreffend handelt (siehe Anhang E für den vollständigen Fragebogen). Es wird der Score für die zwölf Items der beiden Variablen des ursprünglichen TAM gebildet, jedoch auch ein Gesamtscore für alle 19 Items. Auch wenn nach dem TAM eine hohe Korrelation der einzelnen Variablen erwartet wird, werden die zusätzlichen einzelnen Variablen dennoch separat verglichen (Intention der Nutzung, Attitude). Die Vergleiche werden hypothesenkonform in Abhängigkeit mit den Warnsignalen durchgeführt. Das TAM ist – auch wenn in vielen leicht bis stark unterschiedlichen veränderten Versionen - häufig eingesetzt: So zeigt eine viel zitierte Metaanalyse von King und He (2006) von 88

Studien, dass das TAM als robustes Model sehr häufig verwendet wird und das alle Variablen eine hohe Realibilität von mehr als 0,80 haben (King & He, 2006). Cronbachs Alpha beträgt für den in dieser Masterarbeit konzipierten Fragebogen .93.

Vorbereitung und Ablauf des Experiments

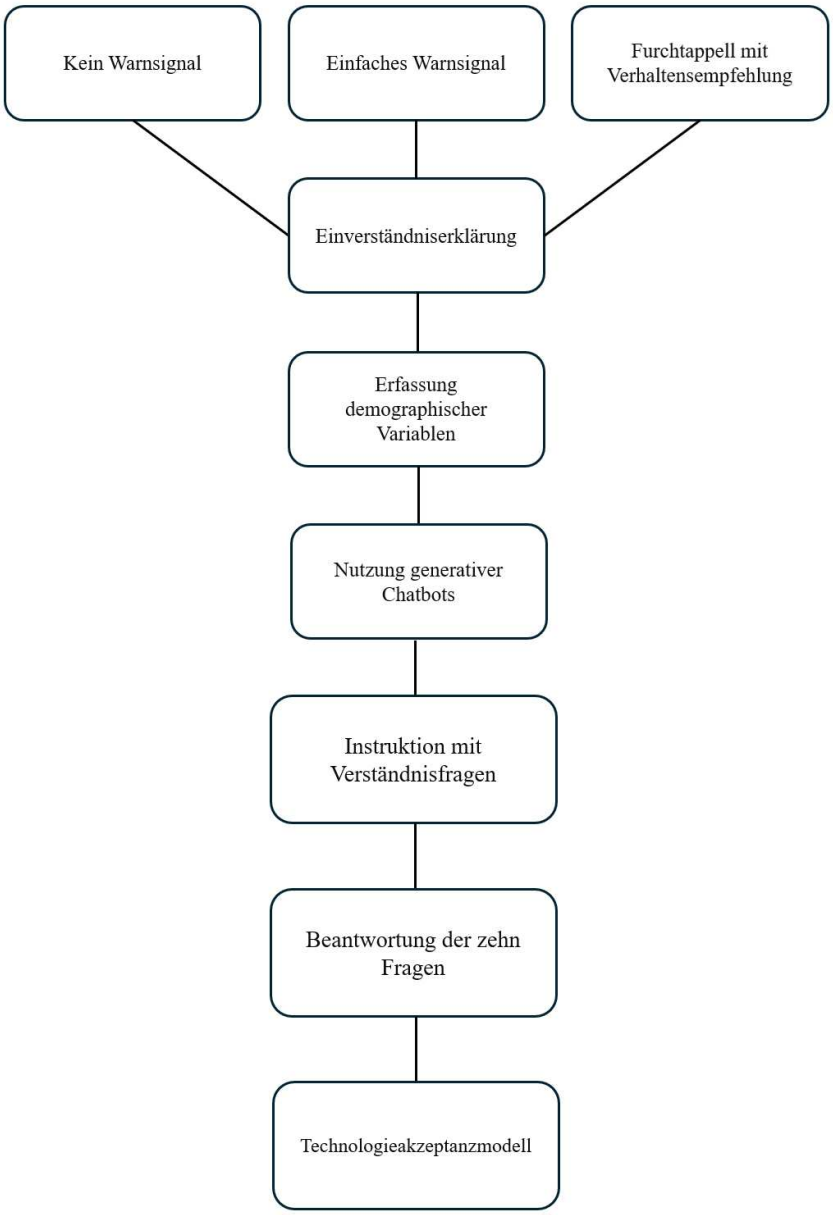
Für die Masterstudie wurde eine vorprogrammierte und offline aufrufbare Seite verwendet, die ChatGPT optisch imitiert. Zur Programmierung wurde eine Dienstleistung über den Online-Marktplatz Fiverr mit der Auftragsnummer #FO61E21140704 in Anspruch genommen (Fiverr, o. D.). Es wurde dementsprechend der offizielle ChatGPT-Quellcode heruntergeladen und angepasst, die für das Experiment relevantesten Anpassungen sind wie folgt: 1. Unten Links ist als "User" *Uni_Vienna* angegeben, damit die ChatGPT-Seite wie eine universitätsinterne Version wirkt. 2. Die ChatGPT-Version ist offline erreichbar und so programmiert, dass insgesamt zehn vorbereitete Antworten ausgegeben werden, unabhängig davon, was eingegeben wird. Des Weiteren ist die Reihenfolge randomisiert, in der diese zehn Antworten wiedergegeben werden. Gibt die Person also jegliche Zeichen im Chat ein, erhält sie im Anschluss randomisiert eine der zehn Antworten. 3. Es wird klar je ChatGPT-Fenster (eins bis zehn) innerhalb der ChatGPT-Version gekennzeichnet, welche Frage nun zu beantworten ist (Frage A bis J): Damit verknüpft sind auch die vorprogrammierten Antworten zu den jeweiligen Fragen – die ProbandInnen wissen also immer, welche Frage sie genau nun beantworten müssen und somit welche Frage sie ChatGPT stellen müssen. 4. Wird eine Frage im Chat eingegeben, muss mit der Maus auf einen grünen Pfeil geklickt werden, damit es als Eingabe zählt. Das soll verhindern, dass die ProbandInnen zu früh aus Versehen die Enter Taste drücken, denn ChatGPT würde – egal was eingegeben wird – immer die vorprogrammierten und somit vorbestimmten Antworten wiedergeben. Es erscheint nach dem Anklicken der grünen Pfeiltaste also die vorprogrammierte Antwort, aber auch ein Button mit der Bezeichnung "Weiter" – wird dieser angeklickt, folgt die nächste randomisierten Frage,

bis alle zehn Fragen und somit alle zehn ChatGPT-Seiten (Frage A bis J) wiedergegeben sind. Sind alle Fragen beantwortet, reagiert die ChatGPT-Seite auf User-Angaben nicht mehr. 5. Aufgrund der drei Bedingungen gibt es drei verschiedene Versionen von der ChatGPT-Seite mit der Variation der Warnsignale als unabhängige Variable.

Die ProbandInnen haben einen identische Fragebogen erhalten (siehe Anhang E): Dieser wurde über das Online-Fragebogensystem Unipark erstellt und veröffentlicht (Unipark, 2023). Die erste Seite beinhaltet die Versuchspersonennummer sowie die Bedingung als Ziffer, die von der Versuchsleitung eingetragen wurde. Es folgt eine Instruktion, eine Einverständniserklärung, die Erfragung demographischer Variablen sowie die Erklärung für ein Gewinnspiel von 20 Euro. Im Anschluss folgt dann die Instruktion des Experiments mit zusätzlichen vier Verständnisfragen als Kontrolle, dass die Instruktion verstanden wurde. Daraufhin folgen die Textfelder für die zehn zu beantwortenden Fragen von A bis J, wobei die ProbandInnen in der Umfrage aufgefordert werden, je nach randomisierter Angabe bei Ihrer ChatGPT-Seite zur entsprechenden Frage zu gelangen und diese zu beantworten. Es folgt die Erfassung der Akzeptanz von ChatGPT über das TAM von Davis (1989) und letztlich ein Debriefing – der Ablauf des Experiments ist in Abbildung 2 grafisch dargestellt.

Abbildung 2

Ablauf des Experiments



Der fertiggestellte Fragebogen wurde zunächst an zwei freiwilligen Personen vorgetestet, um nach möglichen Fehlerquellen, ungünstigen Abläufe oder unverständnisvolle Instruktionen zu kontrollieren. Eine Auffälligkeit war, dass eine Person automatisch immer eine andere Quelle verwendet hat, sobald sie das erste Mal gemerkt hatte, dass ChatGPT eine falsche Antwort wiedergegeben hat. Da eine Person genauere Fragen zur Instruktion hatte, wurde diese im Anschluss genauer formuliert und letzlich vier Verständnisfragen

hinzugefügt. Es gab keine weiteren besonderen Auffälligkeiten. Die Daten diesbezüglich wurden in der statistischen Analyse nicht verwendet und sie wurden auch nicht zur Stichprobe gezählt.

Kurz vor dem Experiment hat die Versuchsleitung den Unipark-Fragebogen mit dem Browser Microsoft Edge geöffnet (Version 126.0.2592.68, Offizielles Build, 64-Bit). Dabei wurde die Versuchspersonennummer eingetragen (Erhebungstag XX, Sessionnummer YY, Sitzplatznummer ZZ), die Bedingung nach der Randomisierung eingetragen, im zweiten Tab dementsprechend die jeweilige ChatGPT-Version geöffnet, im dritten Tab Google, im vierten Tab Bing. Zuvor wurde kontrolliert, dass der Internetverlauf gelöscht gewesen ist, ebenfalls wurde die Instruktion in Papierform bereitgestellt. Zunächst wurde darum gebeten, dass die ProbandInnen eine Einverständniserklärung der Universität Wien ausfüllen (siehe Anhang F). Im Anschluss wurden sie an die Computer gesetzt, nach dem Hinsetzen wurden folgende Aspekte mitgeteilt: Dass die offene Seite die Umfrage ist, die sie durchgehen sollen, dass auf dem Platz die Instruktion aus der Umfrage noch mal separat zum Nachlesen bereitsteht, dass ChatGPT nicht geschlossen werden darf, da dies zum Abbruch der Studie führen würde und das sich bei Fragen jederzeit an die Versuchsleitung gemeldet werden könnte. Haben die ProbandInnen im Anschluss die Umfrage beendet, wurden sie verabschiedet, die Kredits wurden über Labs verteilt, der Computerverlauf wurde gelöscht, der Computer neugestartet und der Prozess begann somit von vorn für die darauffolgenden ProbandInnen. Es gab noch eine zusätzliche Studienform die lediglich eine monetären Belohnung von zehn Euro beinhaltete - ProbandInnen sollten dabei am Ende noch einen weiteren Zettel ausfüllen (siehe Anhang G) und haben im Anschluss die zehn Euro bekommen.

Die Studie wurde im Social Science Research Lab der Universität Wien durchgeführt (Universitätsstraße 7, Raum D615): Es bestand die Möglichkeit für die Teilnahme von parallel 20 Personen, die höchsten parallelen Teilnahmen waren jedoch fünf Personen. Bei

paralleler Teilnahme wurden die ProbandInnen getrennt gesetzt, damit sich untereinander weniger gestört wird.

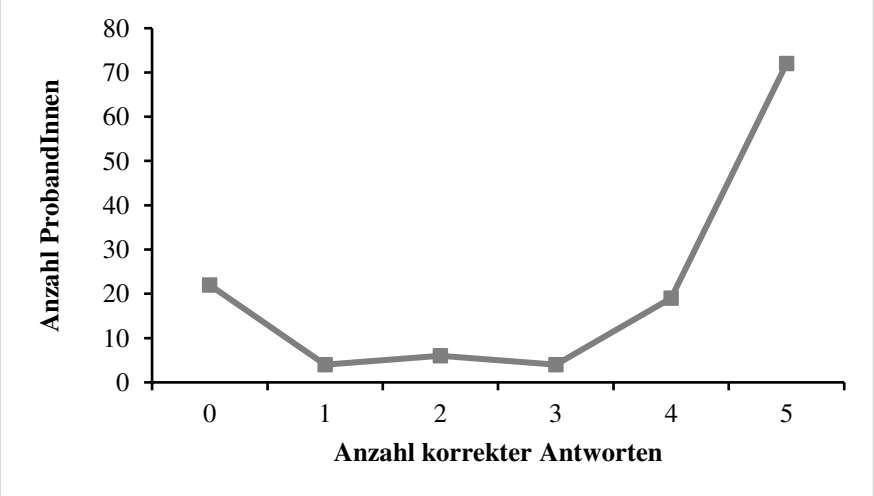
Statistische Analyse

Der Rohdatensatz mit $N = 139$ wurde von Unipark als CSV-Datei (Comma-separated values) exportiert. Zur Masterarbeit wurden alle Daten sowie ein Codehandbuch und Analyseskript bereitgestellt. In einer Kopie des Rohdatensatzes wurde für die statistischen Analyse irrelevante Variablen gelöscht und zwölf ProbandInnen wurden für die statistischen Analysen ausgeschlossen. Es wurden die abhängige Variablen hinzugefügt: GesamtKorrekt (Score von null bis fünf, händisch ausgewertet) und die Scores des TAM (Mittelwerte in Excel ermittelt).

Um die Hypothesen bezüglich korrekter Antworten in Zusammenhang mit Warnsignalen statistisch zu überprüfen, wurde zunächst eine einfaktorielle Varianzanalyse ohne Messwiederholung in JASP durchgeführt (Version 0.19.0, JASP Team, 2024). Im Anschluss wurde ein Post-Hoc Test durchgeführt: Dazu wurde der nicht-parametrische Kruskal-Wallis Test herangezogen, da es sich bei der abhängigen Variable der korrekten Antworten um eine nicht-normalverteilte rechtssteile Häufigkeitsverteilung handelt - der häufigste Wert ist die korrekte Beantwortung aller Fragen, der zweithäufigste Wert ist das falsche Beantworten aller Fragen (siehe Abbildung 3).

Abbildung 3

Häufigkeitsverteilung der korrekten Antworten



Um nach den Hypothesen 1a und 1b die einzelnen Bedingungen miteinander zu vergleichen, wurde im Anschluss für nicht-parametrische Verteilungen der Dunns Test (Post Hoc Test) durchgeführt. Jedoch wurden auch geplante Kontraste der ANOVA verwendet, da die Hypothesen eben klare vordefinierte Vorstellungen von bestimmten Gruppenunterschieden beinhalten, wobei dennoch angemerkt werden muss, dass die Anzahl der korrekten Antworten wie erwähnt nicht normalverteilt ist, daher sind diese Analysen nur mit Vorsicht zu interpretieren.

Um die Akzeptanz in Abhängigkeit der Warnsignale (UV) zu messen, wurde ein Fragebogen basierend auf dem TAM als abhängige Variable verwendet und ausgewertet. Dabei wurde eine einfaktorielle Varianzanalyse ohne Messwiederholung durchgeführt und im Anschluss ein Post-Hoc Turkey Test sowie die Kontraste der ANOVA.

Ergebnisse

Im folgenden werden die Ergebnisse der statistischen Analyse bezüglich der Hypothesen für $n = 127$ präsentiert.

Korrekte Antworten (Hypothese 1a und 1b)

Die Mittelwerte der korrekten Antworten (abhängige Variable) je Bedingung, also Warnsignal (unabhängige Variable), sind in Tabelle 1 und Abbildung 4 dargestellt.

Tabelle 1

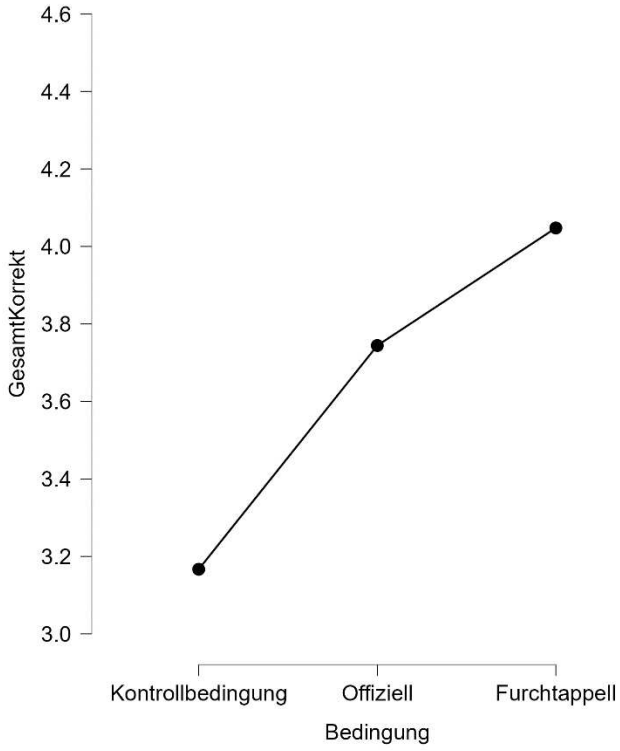
Mittelwerte der korrekten Antworten je Bedingung

	Kontrollbedingung	**Ofiziell**	**Furchtapell**
ProbandIn	42	43	42
M	3.16	3.74	4.04
SD	2.12	1.87	1.72

Anmerkung. Beschrieben sind je nach Warnsignal (UV) die Mittelwerte sowie die Standardabweichungen der korrekten Antworten (AV) bezüglich der falschen Antwortwiedergabe durch ChatGPT.

Abbildung 4

Grafische Darstellung der Mittelwerte korrekter Antworten je Bedingung



Die hauptsächlich zu untersuchende Hypothese 1a – dass ein Warnsignal mit Furchtappell und Verhaltensempfehlung zu mehr korrekten Antworten als ein einfacher und kein Warnsignal führt – wurde zunächst durch eine einfaktorielle Varianzanalyse mit der Bedingung als unabhängige Variable und die Anzahl der korrekten Antworten als abhängige Variable überprüft: Der Haupteffekt der Bedingungen war nicht signifikant, $F(2, 124) = 4.5$, $p = .10$, $\eta^2 = .03$ (siehe Tabelle 2). Der Kruskal-Wallis Test zeigte diesbezüglich ebenfalls keine Signifikanz auf, $p = .13$ (siehe Tabelle 3).

Tabelle 2

Einfaktorielle Varianzanalyse ohne Messwiederholung für die Anzahl der korrekten Antworten in Abhängigkeit vom Warnsignal

	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2
Bedingung	16.83	2	8.41	2.28	0.10	0.03
Residuals	455.92	124	3.67			

Anmerkung. Ein Alpha-Niveau von 5% wurde für alle Tests dieser Studie ausgewählt.

Tabelle 3

Kruskal-Wallis Test für die Anzahl der korrekten Antworten in Abhängigkeit vom Warnsignal

Faktor	Stat.	*df*	*p*
Bedingung	4.01	2	0.13

Da die Hypothesen konkrete vorbedachte Gruppenunterschiede angeben, wurden ebenfalls die ANOVA Einzelkontraste zwischen allen Bedingungen mitberechnet, diese sind in Tabelle 5 dargestellt, in Tabelle 4 wird beschrieben, welche einzelnen Kontraste gebildet wurden.

Tabelle 4

Gebildete Einzelkontraste zur statistischen Analyse in Tabelle 5

Bedingung	Vergleich 1	Vergleich 2	Vergleich 3
Kontrollbedingung	-1	-1	0
Offiziell	1	0	-1
Furchtappell	0	1	1

Tabelle 5

ANOVA-Kontraste für die Bedingungen (Anzahl korrekter Antworten in Abhängigkeit vom Warnsignal)

		95% KI					
Vergleiche	**Estimate**	**Lower**	**Upper**	***SE***	***df***	***t***	***p***
1	0.57	-0.24	1.40	0.41	124	1.38	0.16
2	0.88	0.05	1.70	0.41	124	2.10	0.03
3	0.3	-0.52	1.12	0.41	124	0.72	0.46

Anmerkung. Vergleich 1 gibt den Kontrast zwischen Kontrollbedingung sowie Bedingung *Offiziell*, Vergleich 2 den Kontrast zwischen Kontrollbedingung sowie Bedingung *Furchtappell* und Vergleich 3 den Kontrast zwischen Bedingung 2 sowie 3 an.

Damit zeigt der zweite Vergleich einen signifikanten Unterschied zwischen der Kontrollbedingung und der Bedingung *Furchtappell*, $t(124) = 2.105$, $p = .037$. ProbandInnen in der Bedingung mit einem Warnsignal, das einen Furchtappell und eine Verhaltensempfehlung enthält, haben also signifikant häufiger falsche Antwortwiedergaben durch ChatGPT erkannt und korrigiert als ProbandInnen in der Bedingung ohne einen Warnsignal. Zusätzlich wurden die nicht-parametrischen Dunns Post-Hoc Vergleiche durchgeführt: Ein solcher Vergleich zeigte ebenfalls einen signifikanten Unterschied zwischen den Bedingungen 1 und 3 auf mit einem *p*-Wert von .04 (Tabelle 6).

Tabelle 6

Dunns Post-Hoc Vergleiche für die Anzahl der korrekten Antworten in Abhängigkeit vom Warnsignal

Vergleiche	**z**	**p**
1 - 2	-1.08	0.27
1 - 3	-2.00	0.04
2 - 3	-0.93	0.35

Es zeigt sich kein signifikanter Unterschied zwischen Bedingung *Offiziell* und Bedingung *Furchtappell*, siehe den ANOVA-Kontrast mit $t(124) = 0.72$, $p = .46$ (Tabelle 5) und den Dunns Post-Hoc Vergleich mit einem *p*-wert von .32 (Tabelle 6). Ferner wurde Hypothese 1b nicht bestätigt. Es gab also zwischen der Kontrollbedingung und Bedingung *Offiziell* kein signifikant höheres Erkennen und Korrigieren von falschen Antwortwiedergaben durch ChatGPT beim ANOVA-Kontrast mit $t(124) = 1.38$, $p = .16$ (Tabelle 5) und dem Dunns Post-Hoc Vergleich mit einem *p*-Wert von .27 (Tabelle 6).

Technologieakzeptanz (Hypothesen 2a und 2b)

Deskriptive Werte des Technologieakzeptanzbogens sind in Tabelle 7 aufgeführt. Mittelwerte (19 Items) je Bedingung – also je Warnsignal – sind in Tabelle 8 aufgeführt.

Tabelle 7

Deskriptive Werte des Technologieakzeptanzbogens

	TamGesamt	TamNuetzlichkeitEinfachheit	TamIntention	TamAttitude
ProbandIn	127	127	127	127
M	5.000	5.045	5.144	4.754
SD	1.033	0.973	1.568	1.291
Minimum	1.737	1.833	1.000	1.000
Maximum	6.947	6.917	7.000	7.000

Anmerrkung. Angegeben sind die Mittelwerte, Standardabweichung, Minimum und Maximum des gesamten Technologieakzeptanzbogens TamGesamt (19 Items), der Items bezüglich der Nützlichkeit und Einfachheit (TamNuetzlichkeitEinfachheit, 12 Items), der Items bezüglich der Intention der Weiternutzung (TamIntention, 3 Items) und der Items bezüglich der Einstellung (TamAttitude, 4 Items).

Tabelle 8

Deskriptive Werte des Technologieakzeptanzbogens (Gesamt, 19 Items) je Bedingung

	TamGesamt		
	Kontrollbedingung	**Offiziell**	**Furchtappell**
ProbandIn	42	43	42
M	5.07	4.84	5.08
SD	0.95	1.19	0.92
Minimum	2.68	1.73	2.63
Maximum	6.94	6.68	6.89

Im gesamten sind die Werte sehr ähnlich – unabhängig von der Bedingung und somit vom Warnsignal. Zur statistischen Analyse wurden nur die Mittelwerte des gesamten Technologieakzeptanzbogens herangezogen: Der statistische Vergleich zwischen den einzelnen Variablen zeigte keine Signifikanz auf und wird deshalb nicht weiter aufgeführt.

Eine einfaktorielle Varianzanalyse ohne Messwiederholung zur Erfassung von Gruppenunterschieden zeigte bezüglich der Akzeptanz und dem Warnsignal keine Signifikanz auf mit $F(2, 124) = .753$, $p = .473$, $\eta^2 = 0.012$ (siehe Tabelle 9).

Tabelle 9

Einfaktorielle Varianzanalyse ohne Messwiederholung des Technologieakzeptanzbogens in Abhängigkeit mit dem Warnsignal

	Sum of Squares	***df***	**Mean Square**	***F***	***p***	***η²***
Bedingung	1.614	2	0.807	0.753	0.473	0.012
Residuals	132.876	124	1.072			

Ein Post Hoc Turkey Test zeigt ebenfalls keine signifikanten Unterschiede zwischen den einzelnen Bedingungen (Tabelle 10). Auch die ANOVA Kontraste zeigten keine signifikanten Unterschiede (siehe Tabelle J1 und J2 in Anhang J).

Tabelle 10

Post Hoc Tukey Test für die Gruppenvergleiche des Technologieakzeptanzbogens in Abhängigkeit vom Warnsignal

		Mean Difference	***SE***	***t***	p_{tukey}
1	2	0.233	0.225	1.038	0.554
	3	-0.010	0.226	-0.044	0.999
2	3	-0.243	0.225	-1.083	0.527

Obwohl also in Gruppe 3 verglichen mit Gruppe 1 signifikant öfter die falschen Antwortwiedergaben von ChatGPT erkannt und korrigiert wurden, unterscheidete sich die Akzeptanz nach dem TAM zwischen diesen beiden Gruppen nicht. Somit werden die Hypothesen 2a und 2b verworfen, da sich keine Unterschiede der Akzeptanz zwischen den Bedingungen zeigt und somit keine Unterschiede der Wahrnehmung der 1) Nützlichkeit, 2) der Einfachheit der Nutzung, 3) der Intention der Nutzung von ChatGPT und 4) der Einstellung zu ChatGPT in Abhängigkeit vom Warnsignal.

Diskussion

Das Hauptziel der Studie war es, die Wirksamkeit von Warnsignalen bezüglich KI-Halluzinationen zu untersuchen, ohne den Einsatz und die Effizienz von KI-Chatbots zu limitieren. Dafür wurden als unabhängige Variable Warnsignale und als abhängige Variable die Anzahl von korrekten Antworten von Wissensfragen verwendet: Eine Versuchsbedingung beinhaltete eine ChatGPT-Version ohne ein Warnsignal, eine Versuchsbedingungen beinhaltete eine ChatGPT-Version mit dem offiziellen Warnsignal von ChatGPT – jedoch hervorgehoben – und eine Versuchsbedignung beinhaltete eine ChatGPT-Version mit einem Furchtappell und Verhaltensempfehlung. Für letzteres wurden Theorien aus der Furchtappellforschung herangezogen, insbesondere das EPPM von Kim Witte (1992): Damit wurde versucht, eine furchtinduzierende Botschaft bezüglich KI-Halluzinationen zu formulieren und dazu effiziente Verhaltensempfehlungen vorzuschlagen. Dabei besagte die

hauptsächlich zu untersuchende Hypothese 1a, dass dieser Furchtappell mit Verhaltensempfehlung (Bedingung 3) dazu führen würde, dass ProbandInnen die KI-Halluzinationen vermehrt erkennen und korrigieren als bei einem allgemeiner gehaltenen Warnsignal (Bedingung 2) und keinem Warnsignal (Bedingung 1). Das konnte jedoch nur für den Furchtappell versus kein Warnsignal gezeigt werden. Es bestand also kein signifikanter Unterschied zwischen dem Furchtappell und dem allgemeineren Warnsignal.

Es existieren verschiedene mögliche Erklärungen für vor allem den fehlenden Unterschied zwischen Bedingungen *Offiziell* und *Furchtappell*. Aufgrund der geringen Stichprobe könnte ein möglicher kleiner Effekt – wie in der Literatur von Warnsignalen und Furchtappellen ebenfalls häufiger zu finden (siehe Tannenbaum et al., 2015) – nicht gezeigt werden: Nach der Teststärkenanalyse wären mindestens 159 ProbandInnen für den in dieser Studie angenommenen mittleren Effekt notwendig gewesen.

Eine weitere Limitation könnte die Formulierung des Furchtappells gewesen sein: Damit ein Furchtappell nach der Furchtappellforschung zur Schutzmotivation führt, muss als erster Schritt ein Furchtgefühl ausgelöst werden (Witte, 1992). Damit das geschieht, muss nach dem EPPM die Stärke und Empfänglichkeit (Perceived Susceptibility) der angedeuteten Gefahr zumindest moderat sein (Witte, 1992). Nach der Metaanalyse von Tannnebaum et al. (2015) nimmt ebenfalls die Wirksamkeit von Furchappellen zu, wenn die Stärke und Wahrscheinlichkeit der Gefahr zunimmt. In dieser Masterarbeit wurde ein Warnsignal mit folgenden möglicherweise furchtinduzierenden Aussagen konzipiert: Dass a) ChatGPT häufig falsche Antworten wiedergibt, b) dass dadurch ein Durchfallen im eigenen Fach möglich ist und c) dass dadurch falsche Informationen weitergegeben werden könnten (siehe Anhag C für die verwendeten Warnsignale). Vor allem die Aussage bezüglich des Durchfallens traf auf keinen der ProbandInnen dieser Studie zu, daher ist hier zumindest anzunehmen, dass diese keine Furcht ausgelöst hat, da die ProbandInnen wussten, dass sie

durch die Studie in keinem Szenario universitär durchfallen könnten. Die Aussage bezüglich der Weitergabe von falschen Informationen könnte zwar unangenehm sein, jedoch – auch aufgrund der Anonymität – nicht unbedingt furchtauslösend, da in diesem Kontext mit keinen Konsequenzen zu rechnen wäre. Damit wäre hier also nach dem EPPM die Empfänglichkeit der Gefahr möglicherweise gering. Daher ist beim verwendeten Warnsignal dieser Studie kritikwürdig, dass dieser in diesem Kontext überhaupt erst gar keine Furcht ausgelöst hat und das dadurch der Prozess der Gefahrenkontrolle nicht aktiviert wurde, so wie im EPPM nach Witte (1996) vorgesehen. Eine alternative Formulierung des Furchtappells mit Manipulationscheck wäre eine Möglichkeit gewesen, diese Limitation zu reduzieren.

Im Kontext von KI-Halluzinationen ist es jedoch schwierig, basale Themen einzubeziehen, die Furcht auslösen könnten, wie der Tod oder die Gesundheit. Zudem ist die Effizienz der Verhaltensempfehlung des in dieser Masterarbeit verwendeten Furchtappells als hoch angenommen, da die Selbsteffizienz der ProbandInnen als hoch erwartet wurde, da überwiegend jüngere Studierende teilgenommen haben und dabei zu erwarten ist, dass diese regelmäßig Internetrecherchen durchführen. Dennoch wäre es auch allgemein empfehlenswert, die Effizienz und Selbsteffizienz von der Verhaltensempfehlung abzufragen, um diesbezüglich sicherzugehen.

Zu beachten sind auch möglicherweise vorgekommene *Demand Characteristics*: ProbandInnen könnten möglichst die Hypothesen des Forschenden bestätigen wollen und verhalten sich dementsprechend nach ihren Vermutungen hypothesenkonform (Orne, 1962). Da vor allem das Warnsignal von ChatGPT anders als üblich bekannt dargestellt wird - rot und größer hervorgehoben – könnte diesem Warnsignal absichtlich befolgt worden sein. Wichtig zu beachten ist jedoch, dass dieser Effekt auch für den Furchtappell und somit für beide Bedingungen gilt, wenn nicht sogar noch stärker für den Furchtappell, da dieser noch ein auffälligeres und markanteres Warnsignal darstellt. Alternativ wäre der Einsatz eines

nicht offiziell existierenden KI-Chatbots sinnvoller gewesen, um diese Ungewöhnlichkeit der ChatGPT-Seite nicht entstehen zu lassen. ProbandInnen würden also nicht mit einer offiziellen, ihr bekannten Website konfrontiert werden, welche nur in der Studie ein anderes Warnsignal hat. Von den 127 ProbandInnen haben nämlich nur sieben noch nie einen KI-Chatbot wie ChatGPT benutzt.

Auch gegensätzlich zu den Hypothesen 2a und 2b konnte nicht gezeigt werden, dass die Akzeptanz – also die wahrgenommene Nützlichkeit und Einfachheit der Nutzung – sich in Abhängigkeit vom Warnsignal negativ ändert: Es wurde angenommen, dass je stärker ein Warnsignal wirkt, die Akzeptanz geringer ausfallen wird. Obwohl der Furchtappell mit Verhaltensempfehlung stärker gewirkt hat als kein Warnsignal, gab es diesbezüglich keine Unterschiede in der Akzeptanz. Dies zeigt auf, dass trotz stärkerer Wirkung des Warnsignals die wahrgenommene Nützlichkeit und Einfachheit der Nutzung nicht sinkt, obwohl zu erwarten wäre, dass durch zusätzliche Recherchen und Fehleranfälligkeiten der KI die Nützlichkeit als geringer bewertet werden müsste, da dadurch ein Mehraufwand besteht. Eine Vermutung ist, dass die ProbandInnen, primär bestehend aus Studierenden, ChatGPT als so nützlich und bedienungsfreundlich wahrnehmen, dass KI-Halluzinationen zu keiner Einschränkung der Akzeptanz nach dem TAM führen und sie es somit trotzdem als effizient und ressourcensparend bewerten. Das könnte in einem Fragebogen nach dem Experiment genauer abgeklärt werden: Dort könnte gefragt werden, ob sie ChatGPT trotz häufiger KI-Halluzination als nützlicher wahrnehmen als übliche Suchmaschinen, beziehungsweise könnten die Items des TAM so umformuliert werden, dass der Aspekt des KI-Halluzinierens bei allen Items inbegriffen ist: Zum Beispiel würde das Item der wahrgenommenen Nützlichkeit “Der Einsatz von ChatGPT an der Universität würde es mir ermöglichen, Aufgaben schneller zu erledigen“ mit dem Zusatz “[…], obwohl es häufiger falsche

Antworten wiedergeben kann“ versehen werden. Diese Modifizierung des TAM wäre jedoch neu, somit wäre zusätzlich zu erörtern, ob die Items damit valide und reliabel wären.

Implikation für die Forschung

Sinnvoll ist es, die vorher aufgezählten Limitationen aufzugreifen und in zukünftigen Forschungen zu reduzieren: Unter anderem ist es relevant, furchtinduzierende Furchtappelle zu entwickeln, es müsste also zusätzlich überprüft werden, ob die Furchtappelle überhaupt Furcht auslösen, denn nur so kann nach dem EPPM der Prozess der Schutzmotivation und somit das Ausführen der Verhaltensempfehlung möglicherweise erfolgen.

Ferner sollten auch die Verhaltensempfehlungen auf Effizienz und – in Abhängigkeit der Stichprobe und des Kontextes – auf Selbsteffizienz überprüft werden, nur so kann sichergestellt werden, dass die Schutzmotivation nach dem EPPM erfolgt. Nach Tannenbaum et al. (2015) sollten Verhaltensempfehlungen auch selbst eine Aussage über ihre Effektivität treffen – das heißt im Warnsignal inkludiert sein.

Ebenfalls ist es empfehlenswert, neben offiziellen und bekannten KI-Chatbots auch selbst entwickelte und somit nicht bekannte zu wählen, um mögliche Demand Characteristics zu verringern. Auch ist es wichtig, dass in zukünftigen Studien die Stichprobe höher und breiter ausfällt, das heißt also nicht nur überwiegend Studierende inkludiert um vermehrt allgemeinere Schlussfolgerungen auf die Gesamtpopulation ziehen zu können.

Implikation für Betreibende von KI-Chatbots

Nach dieser Studie stellen sich zwei relevante Fragen für KI-Chatbot Betreibende: Welche Warnsignale sind effizient und sobald diese effizient sind, wird das System weniger akzeptiert und genutzt? Da hier primär ein Warnsignal nach dem EPPM entwickelt wurde, ist für die erste Frage relevant, a) durch das Warnsignal moderate bis hohe Furcht auszulösen und b) eine effiziente Verhaltensempfehlung als Alternative zu formulieren. Außerdem sollten die Warnsignale nach Conzola und Wogalter (2001) auffällig sein – konträr zu dem

offiziellen von ChatGPT – jedoch auch präzise und kurz und auch zum Glauben der Personen passen, also keine absurden Behauptungen beinhalten, die implizierte Gefahr sollte also realistisch sein. Auch nützlich für die Formulierung von Furchtappellen ist nach der Metaanalyse von Tannenbaum et al. (2015), dass die maximale Effektivität bereits erreicht ist, wenn ein Furchtappell eine moderate Furcht auslöst. Betreibende könnten zumindest analysieren, wofür ihre KI-Chatbots am meisten verwendet werden und dementsprechend realitätsnahe moderate Furchtappelle mit Verhaltensempfehlungen formulieren.

Auch interessant ist das Ergebnis dieser Studie, dass die Akzeptanz nach dem TAM für alle ChatGPT-Versionen identisch blieb, somit unabhängig von den Warnsignalen, obwohl der Furchtappell mit Verhaltensempfehlung tendenziell häufiger zum Erkennen und Korrigieren von KI-Halluzinationen geführt hat als kein Warnsignal: Das impliziert für Betreibende, dass die Verwendung von Warnsignalen – sogar von markanten und auffälligen wie der Furchtappell in dieser Studie – zu keiner verringerten Akzeptanz und auch somit geringeren Intention der Nutzung führen. In der Forschung existiert auch die Annahme, dass Furchtappelle sogar zu negativen Effekten führen könnten oder sogar gefährlich wären: Diesen Annahmen wurde in der Metaanalyse von Tannenbaum et al. (2015) widersprochen, es wurde aufgezeigt, dass Furchtappelle effektiv sind - ohne negative Folgen oder Gefahren. Betreibende müssen also keine Angst vor ausformulierten und auffälligen Warnsignalen haben – diese schränken nach dieser Studie nicht die Nutzung und die Akzeptanz des KI-Chatbots ein.

Fazit

Der Furchtappell mit Verhaltensempfehlung zeigte sich im Gegensatz zu keinem Warnsignal tendenziell effektiver für das Erkennen und Korrigieren von KI-Halluzinationen: Dieses Ergebnis ist nützlich für unter anderem KI-Chatbot-Betreibende, um die möglichen negativen Folgen von KI-generierten Falschinformationen zu reduzieren. Interessant dabei ist

auch, dass sich die Akzeptanz für die KI trotz unterschiedlicher Warnsignale nicht verändert hat, daher sollten sich Betreibende sogar vor markanteren und prägnanteren Warnsignalen nicht fürchten. Jedoch ist – vor allem in Hinblick auf die Forschung – zu beachten, dass es sich bei dieser Studie um die erste Untersuchung von Warnsignalen und KI-Halluzinationen handelt und daher bisher nur erste Impulse für die weitere Forschung liefert: Ein Fokus sollte darauf gelegt werden, die Furchtappelle bezüglich KI-Halluzinationen weiter auszubauen und in separaten Untersuchungen nach dem EPPM zu testen (Furchtgefühle, Effizienz der Verhaltensempfehlungen). Schlussendlich wurde also ein effektiver erster Furchtappell mit Verhaltensempfehlungen für KI-Halluzinationen formuliert sowie überprüft und erste Impulse für die Forschung gesetzt.

Literaturverzeichnis

Alkaissi, H. & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Curēus*. https://doi.org/10.7759/cureus.35179

Argo, J. J. & Main, K. J. (2004). Metaanalyses of the Effectiveness of Warning Labels. *Journal Of Public Policy & Marketing*, *23*(2), 193–208. https://doi.org/10.1509/jppm.23.2.193.51400

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, *37*(2), 122–147. https://doi.org/10.1037/0003-066x.37.2.122

Basil, M., Basil, D., Deshpande, S. & Lavack, A. M. (2013). Applying the Extended Parallel Process Model to Workplace Safety Messages. *Health Communication*, *28*(1), 29–39. https://doi.org/10.1080/10410236.2012.708632

Bhattacharyya, M., Miller, V. M., Bhattacharyya, D. & Miller, L. E. (2023). High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Curēus*. https://doi.org/10.7759/cureus.39238

Birmingham, W. C., Hung, M., Boonyasiriwat, W., Kohlmann, W., Walters, S. T., Burt, R. W., Stroup, A. M., Edwards, S. L., Schwartz, M. D., Lowery, J. T., Hill, D. A., Wiggins, C. L., Higginbotham, J. C., Tang, P., Hon, S. D., Franklin, J. D., Vernon, S. & Kinney, A. Y. (2015). Effectiveness of the extended parallel process model in promoting colorectal cancer screening. *Psycho-oncology*, *24*(10), 1265–1278. https://doi.org/10.1002/pon.3899

CalculatorSoup, L. (2023, 19. September). *Random Number Generator*. CalculatorSoup. Abgerufen am 21. Juni 2024, von https://www.calculatorsoup.com/calculators/statistics/random-number-generator.php

Conzola, V. C. & Wogalter, M. S. (2001). A Communication–Human Information Processing (C–HIP) approach to warning effectiveness in the workplace. *Journal Of Risk Research*, *4*(4), 309–322. https://doi.org/10.1080/13669870110062712

Cuschieri, S. (2019). The STROBE guidelines. *Saudi Journal Of Anaesthesia*, *13*(5), 31. https://doi.org/10.4103/sja.sja_543_18

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *Management Information Systems Quarterly*, *13*(3), 319. https://doi.org/10.2307/249008

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K. & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Ecker, U. K. H., Lewandowsky, S. & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087–1100. https://doi.org/10.3758/mc.38.8.1087

Faul, F., Erdfelder, E., Buchner, A. & Lang, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

Fiverr. (o. D.). *Freelance-Experten beauftragen auf Fiverr | Fiverr* [Video]. fiverr.com. Abgerufen am 21. Juni 2024, von https://de.fiverr.com/

Frąckiewicz, M. (2023, 14. Oktober). *OpenAI's ChatGPT becomes the 24th most visited website globally*. TS2 SPACE. https://ts2.space/en/openais-chatgpt-becomes-the-24th-most-visited-website-globally/#gsc.tab=0

Gharlipour, Z., Hazavehei, S. M. M., Moeini, B., Nazari, M., Beigi, A. M., Tavassoli, E., Heydarabadi, A. B., Reisi, M. & Barkati, H. (2015). The effect of preventive

educational program in cigarette smoking: Extended Parallel Process Model. *Journal Of Education And Health Promotion*, *4*(1), 4. https://doi.org/10.4103/2277-9531.151875

Gilbert, D. T., Tafarodi, R. W. & Malone, P. S. (1993). You can't not believe everything you read. *Journal Of Personality And Social Psychology*, *65*(2), 221–233. https://doi.org/10.1037/0022-3514.65.2.221

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. http://ci.nii.ac.jp/ncid/BA12691128

Greifeneder, R., Jaffe, M., Newman, E. & Schwarz, N. (2020). The Psychology of Fake News. In *Routledge eBooks*. https://doi.org/10.4324/9780429295379

Half of college students say using AI is cheating | BestColleges. (o. D.). BestColleges.com. https://www.bestcolleges.com/research/college-students-ai-tools-survey/

Home | AsPredicted. (o. D.). Abgerufen am 30. Juni 2024, von https://aspredicted.org/

Laboratory Administration for Behavioral Sciences. (o. D.). https://labs-univie.sona-systems.com/default.aspx?p_return_experiment_id=1577

Hovland, C.I., Janis, I.L., & Kelley, H.H. (1953). Communication and persuasion. *Yale University Press*.

Hudlicka, E. (2011). Guidelines for Designing Computational Models of Emotions. *International Journal Of Synthetic Emotions*, *2*(1), 26–79. https://doi.org/10.4018/jse.2011010103

JASP Team (2024). JASP (Version 0.19.0)[Computer software]

King, W. R. & He, J. (2006). A meta-analysis of the technology acceptance model. *Information & Management*, *43*(6), 740–755. https://doi.org/10.1016/j.im.2006.05.003

Laboratory Administration for Behavioral Sciences. (o. D.-b). https://labs-univie.sona-systems.com/default.aspx?p_return_experiment_id=1592

Laidre, M. E. (2012). Principles of Animal Communication. *Animal Behaviour*, *83*(3), 865–866. https://doi.org/10.1016/j.anbehav.2011.12.014

Leventhal, H. (1971). Fear appeals and persuasion: the differentiation of a motivational construct. *American Journal Of Public Health*, *61*(6), 1208–1224. https://doi.org/10.2105/ajph.61.6.1208

Levine, T. R., Park, H. S. & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs*, *66*(2), 125–144. https://doi.org/10.1080/03637759909376468

Martel, C. & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, *54*, 101710. https://doi.org/10.1016/j.copsyc.2023.101710

McCornack, S. A. & Parks, M. R. (1986). Deception Detection and Relationship Development: The Other Side of Trust. *Annals Of The International Communication Association*, *9*(1), 377–389. https://doi.org/10.1080/23808985.1986.11678616

Muller, K. & Cohen, J. (1989). Statistical Power Analysis for the Behavioral Sciences. *Technometrics*, *31*(4), 499. https://doi.org/10.2307/1270020

Nahar, M., Seo, H., Lee, E., Xiong, A. & Lee, D. (2024). Fakes of Varying Shades: How Warning Affects Human Perception and Engagement Regarding LLM Hallucinations. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.03745

Noar, S. M., Hall, M. G., Francis, D. B., Ribisl, K. M., Pepper, J. K. & Brewer, N. T. (2015). Pictorial cigarette pack warnings: a meta-analysis of experimental studies. *Tobacco Control*, *25*(3), 341–354. https://doi.org/10.1136/tobaccocontrol-2014-051978

Noar, S. M., Rohde, J. A., Barker, J. O., Hall, M. G. & Brewer, N. T. (2020). Pictorial cigarette pack warnings increase some risk appraisals but not risk beliefs: A Meta-Analysis. *Human Communication Research*, *46*(2–3), 250–272. https://doi.org/10.1093/hcr/hqz016

OpenAI. (2023). *ChatGPT* (Version 14. März) [Large language model (bzw. Grosses Sprachmodell, *Anm.*)]. https://chat.openai.com/chat.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11), 776–783. https://doi.org/10.1037/h0043424

Pohl, R. F. (2016). *Cognitive Illusions: Intriguing Phenomena in Judgement, Thinking and Memory*. Psychology Press.

Purmehdi, M., Legoux, R., Carrillat, F. & Senecal, S. (2017). The Effectiveness of Warning Labels for Consumers: A Meta-Analytic Investigation into Their Underlying Process and Contingencies. *Journal Of Public Policy & Marketing*, *36*(1), 36–53. https://doi.org/10.1509/jppm.14.047

Rahman, M. M. & Watanobe, Y. (2023). ChatGPT for Education and Research: Opportunities, Threats, and Strategies. *Applied Sciences*, *13*(9), 5783. https://doi.org/10.3390/app13095783

Roberto, A. J., Mongeau, P. A., Liu, Y. & Hashi, E. C. (2019). "Fear the Flu, Not the Flu Shot": A Test of the Extended Parallel Process Model. *Journal Of Health Communication*, *24*(11), 829–836. https://doi.org/10.1080/10810730.2019.1673520

Rogers, R. W. (1975). A Protection Motivation Theory of Fear Appeals and Attitude Change1. *Journal Of Psychology (Washington, D.C. Online)/The Journal Of Psychology*, *91*(1), 93–114. https://doi.org/10.1080/00223980.1975.9915803

Roß, B., Jung, A., Heisel, J. & Stieglitz, S. (2018). Fake news on social media: The (In)Effectiveness of warning messages. *International Conference on Information Systems*, 16. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1228&context=icis2018

Shewale, R. (2023, 21. November). *ChatGPT Statistics: Detailed Insights on users (2023)*. https://www.demandsage.com/chatgpt-statistics

Shuhaiber, A. & Mashal, I. (2019). Understanding users' acceptance of smart homes. *Technology in Society*, *58*, 101110. https://doi.org/10.1016/j.techsoc.2019.01.003

Siebenaller, B. (2017). *Die EU Richtlinie 2014/40/EU. ein Erfolg für die Tabaklobby?* GRIN Verlag.

Stanley, M. L., Whitehead, P. S. & Marsh, E. J. (2022). The cognitive processes underlying false beliefs. *Journal Of Consumer Psychology*, *32*(2), 359–369. https://doi.org/10.1002/jcpy.1289

Stock Fotos, lizenzfreie Bilder & kostenlose Bilder. (o. D.). https://www.pexels.com/de-de/.

Street, C. N. H. & Richardson, D. C. (2014). Lies, Damn Lies, and Expectations: How Base Rates Inform Lie–Truth Judgments. *Applied Cognitive Psychology*, *29*(1), 149–155. https://doi.org/10.1002/acp.3085

Tannenbaum, M. B., Hepler, J., Zimmerman, R. S., Saul, L., Jacobs, S., Wilson, K. & Albarracín, D. (2015). Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychological Bulletin*, *141*(6), 1178–1204. https://doi.org/10.1037/a0039729

The Global Risks Report 2024: Insight Report. (2024). *World Economic Forum*, *19*. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf

Unipark. (2023, 17. November). *Online Umfrage | Online Umfrage erstellen: Gratis testen!* Abgerufen am 21. Juni 2024, von https://www.unipark.com/

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs*, *59*(4), 329–349. https://doi.org/10.1080/03637759209376276

Witte, K. (1996). Fear as motivator, fear as inhibitor. In *Elsevier eBooks* (S. 423–450). https://doi.org/10.1016/b978-012057770-5/50018-7

Witte, K. & Allen, M. (2000). A Meta-Analysis of Fear Appeals: Implications for Effective Public Health Campaigns. *Health Education & Behavior*, *27*(5), 591–615. https://doi.org/10.1177/109019810002700506

Xiao, B. & Benbasat, I. (2015). Designing Warning Messages for Detecting Biased Online Product Recommendations: An Empirical Investigation. *Information Systems Research*, *26*(4), 793–811. https://doi.org/10.1287/isre.2015.0592

Xu, Z., Jain, S. & Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.11817

Zonouzy, V. T., Niknami, S., Ghofranipour, F. & Montazeri, A. (2018). An educational intervention based on the extended parallel process model to improve attitude, behavioral intention, and early breast cancer diagnosis: a randomized trial. *International Journal Of Women's Health*, *Volume 11*, 1–10. https://doi.org/10.2147/ijwh.s182146

Zou, M. & Huang, L. (2023). To use or not to use? Understanding doctoral students' acceptance of ChatGPT in writing through technology acceptance model. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1259531

Anhang

Anhang A: Allgemeinere Warnsignale

Allgemeinere Warnsignale nach Ecker et al. (2022) und OpenAI (2023)

In their desire to sensationalize, the media sometimes does not check facts before publishing information that turns out to be inaccurate. It is therefore important to read the following story and answer the questions at the end carefully. Zitiert nach Ecker et al., (2022).

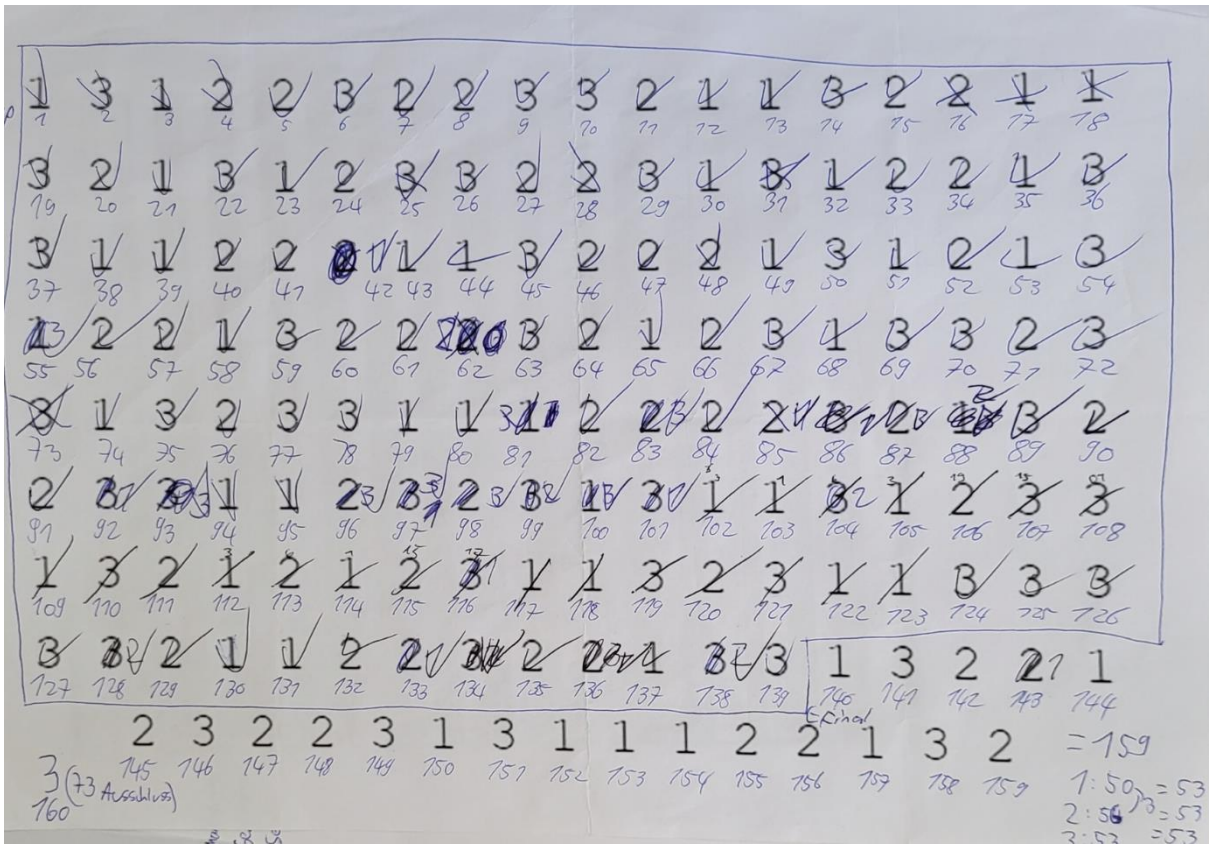
ChatGPT kann Fehler machen. Überprüfe wichtige Informationen. Zitiert nach OpenAI (2023).

Anhang B: Zufallsreihenfolge der Bedingungen

Zufallsgenerator für die Bedingungen der Warnsignale

Abbildung B1

Ermittelte Zufallsreihenfolge der Bedingungen für die Warnsignale



Anmerkung. Die durchgängige Änderung einiger Bedingungen ist begründet dadurch, dass sich anfangs der Studie wenige ProbandInnen gemeldet haben: Dadurch sollte eine Gleichverteilung der Bedingungen erfolgen. Die Randomisierung erfolgte durch CalculatorSoup (2023).

Anhang C: Warnsignale als unabhängige Variable

Abbildung C1

Bedingung 1: Kein Warnsignal auf der vorprogrammierten ChatGPT-Seite

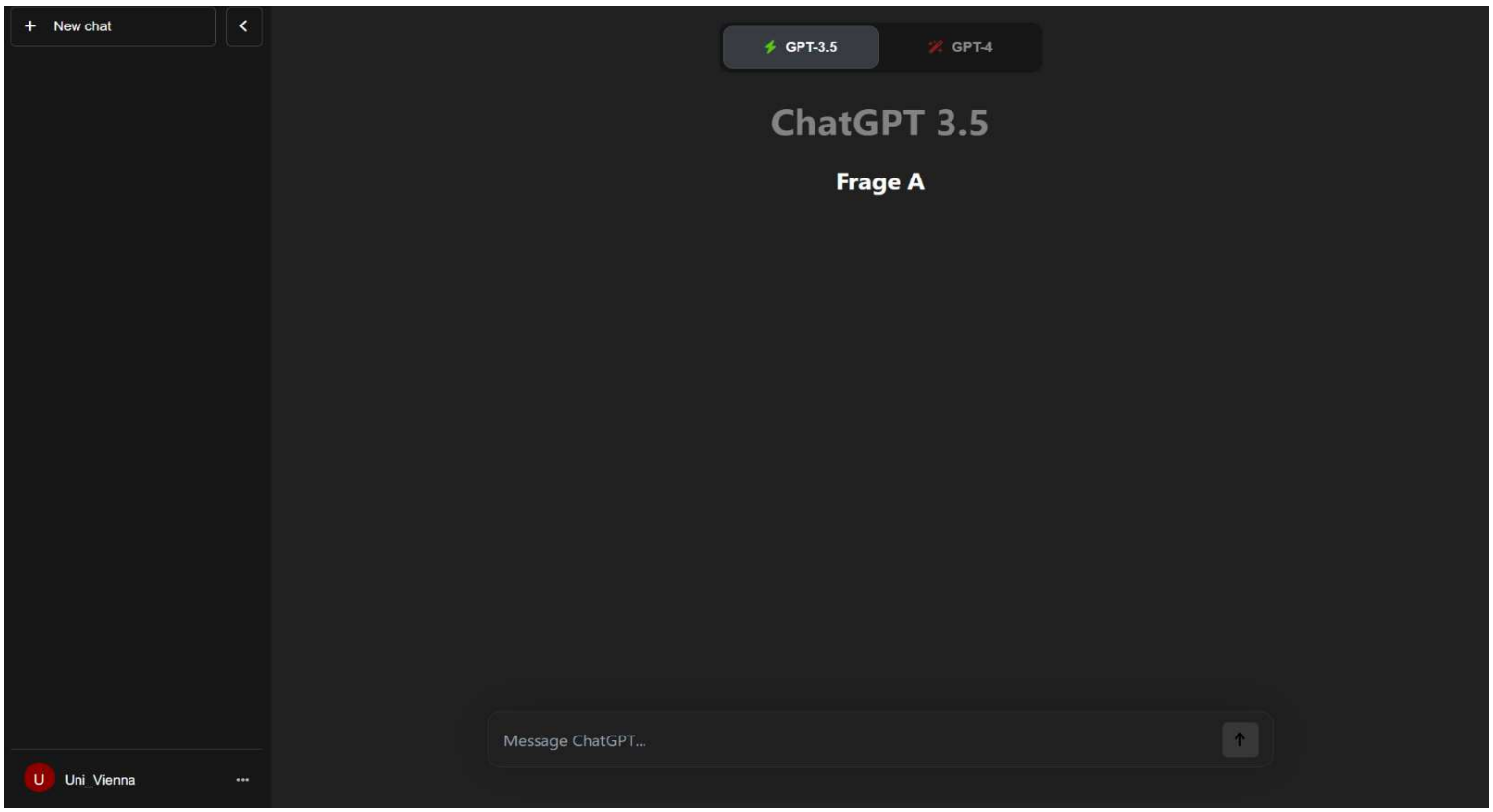


Abbildung C2

Bedingung 2: Einfacher Warnsignal nach ChatGPT, hervorgehoben

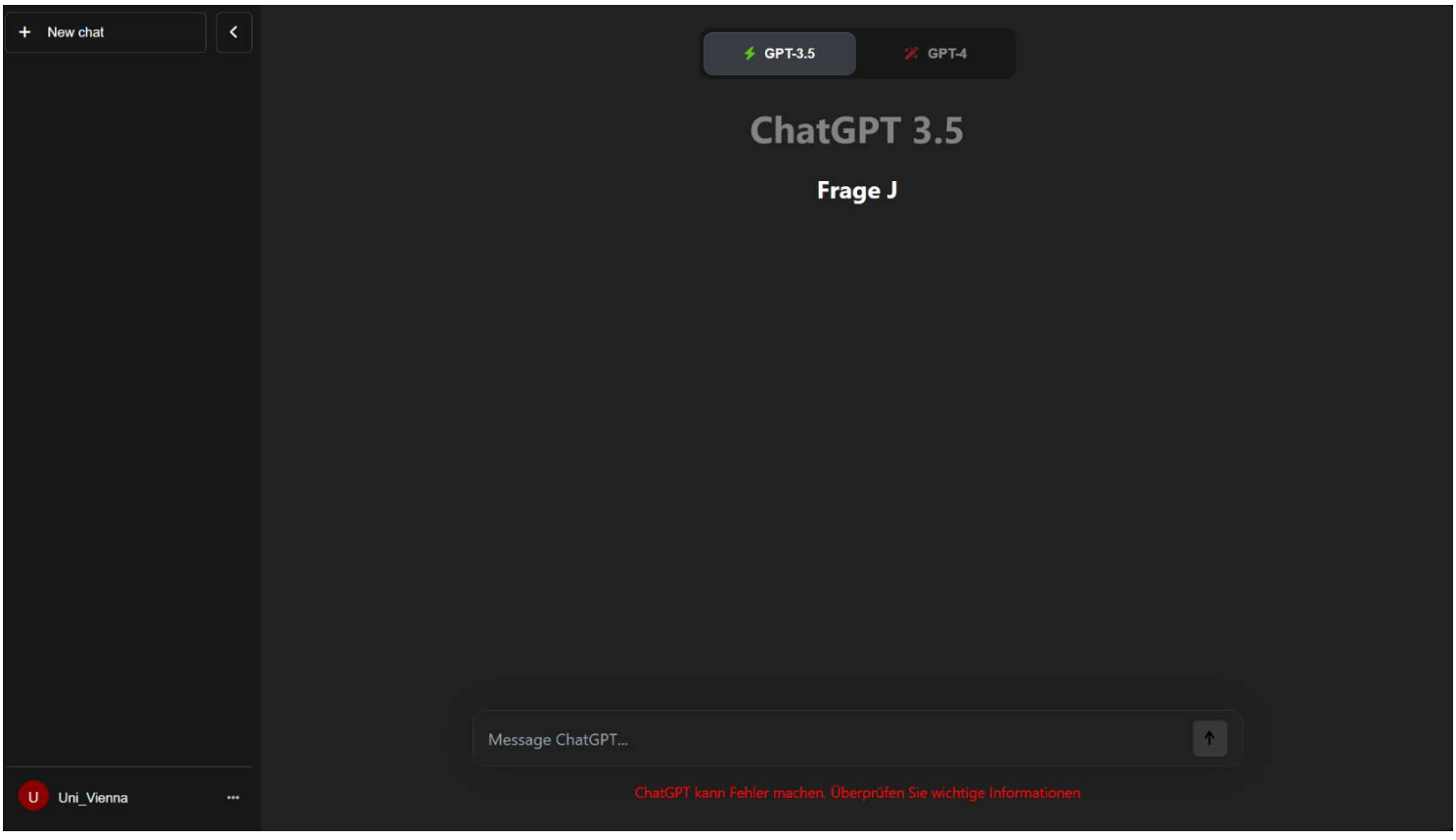
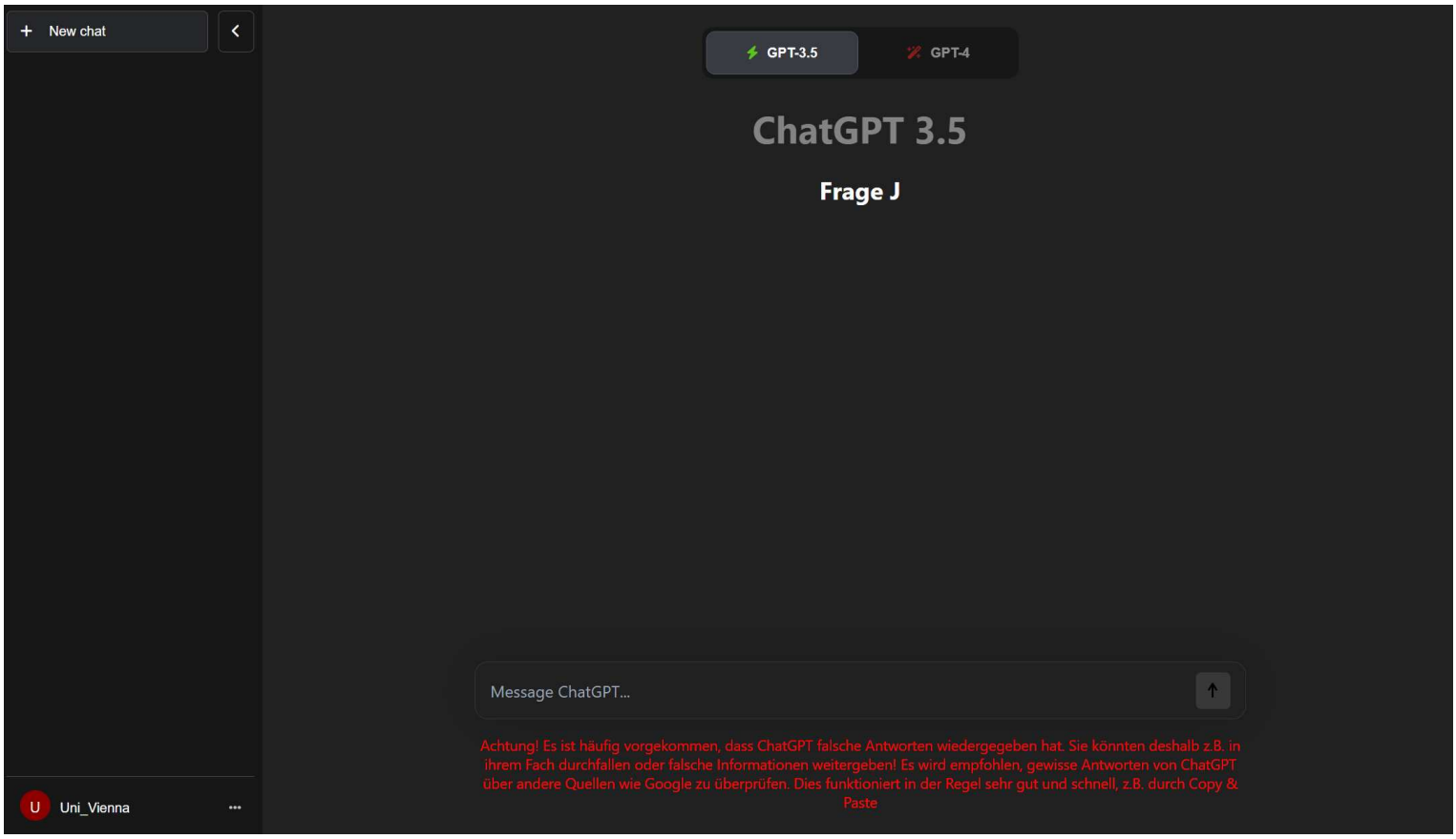


Abbildung C3

Furchtappell mit Verhaltensempfehlung nach Witte (1992)



Anhang D: Fragen und vorprogrammierte Antworten von ChatGPT

Fragen A bis J mit den vorprogrammierten Antworten aus allen ChatGPT-Versionen

Frage A: Von welchen zwei Personen wurde die Kopenhagener Deutung (Quantenmechanik) formuliert?
Vorprogrammierte falsche Antwort: Die Kopenhagener Deutung der Quantenmechanik wurde in den 1920er Jahren von einer Gruppe von Physikern entwickelt, die eng mit dem Institut für Theoretische Physik in Kopenhagen verbunden waren. Die Hauptfiguren hinter dieser Interpretation waren Paul Dirac und Pascual Jordan
Wahre Antwort: Niels Bohr und Werner Heisenberg

Frage B: Was ist die größte Religion in Äthiopien?
Vorprogrammierte bereits richtige Antwort: Äthiopien ist ein vielfältiges Land mit verschiedenen religiösen Gemeinschaften. Die größte Religion in Äthiopien ist das Christentum, wobei die Mehrheit der Christen der äthiopisch-orthodoxen Tewahedo-Kirche angehört. Diese Kirche hat eine lange Geschichte.

Frage C: Ungefähr wie viel Kilometer beinhaltet die längste Nonstop Fluglinie der Welt?
Vorprogrammierte bereits richtige Antwort: Die längste nonstop Fluglinie der Welt ist die Strecke zwischen Singapur (Changi Airport, SIN) und Newark (Newark Liberty International Airport, EWR) in den USA, betrieben von Singapore Airlines. Die Flugstrecke beträgt ungefähr 15.349 Kilometer.

Frage D: Welche Insel wurde in der Zoologie durch ihre hohe Schlangenpopulation bekannt?
Vorprogrammierte falsche Antwort: Die Insel, die in der Zoologie aufgrund ihrer hohen Schlangenpopulation bekannt ist, ist die Abrolhos-Insel . Diese Insel liegt vor der Küste Brasiliens im Atlantik. Sie ist besonders bekannt für ihre hohe Dichte an Bothrops insularis, einer Giftschlangenart, die lokal als Goldener Lanzenotter bekannt ist. Aufgrund der Gefahr, die von den Schlangen auf der Insel ausgeht, ist der Zugang für die Öffentlichkeit stark eingeschränkt.
Wahre Antwort: Queimada Grande

Frage E: Welche Universität wurde 1904 in England gegründet?
Vorprogrammierte bereits richtige Antwort: Im Jahr 1904 wurde die Universität Leeds in England gegründet.

Frage F: Welcher britische Pirat war im 18. Jahrhundert im Nordwestatlantik im Bereich der amerikanischen Ostküste und der Bahamas aktiv?
Vorprogrammierte bereits richtige Antwort: Ein bekannter britischer Pirat, der im 18. Jahrhundert im Nordwestatlantik und im Bereich der amerikanischen Ostküste sowie der Bahamas aktiv war, ist Edward Teach, besser bekannt als Blackbeard. Blackbeard war einer der bekanntesten und berüchtigtsten Piraten seiner Zeit und wurde für seine wilde Erscheinung, einschließlich des imposanten schwarzen Bartes, bekannt. Er operierte hauptsächlich in den Gewässern des westlichen Atlantiks und der Karibik in den frühen 18. Jahrhundert. Blackbeard wurde schließlich 1718 im Zuge eines Gefechts mit der Royal Navy getötet.

Frage G: Welche Wissenschaftlerin hat 2009 veröffentlicht, dass das Gehirn 86 Milliarden Nervenzellen hat?
Vorprogrammierte falsche Antwort: Die brasilianische Neurowissenschaftlerin, die 2009 an der Veröffentlichung beteiligt war, die besagte, dass das menschliche Gehirn etwa 86 Milliarden Nervenzellen hat, ist Dr. Beatriz Maia. Sie ist bekannt für ihre Arbeit im Bereich der Quantifizierung von Zellen im Gehirn und ihrer Forschung zur Neuroanatomie.
Wahre Antwort: Suzana Herculano-Houzel

Frage H: Wer wurde 1978 von Michael Hart in seinem Buch "Die 100 einflussreichsten Persönlichkeiten der Menschehitsgeschichte" auf Platz 8 gestellt?
Vorprogrammierte falsche Antwort: Michael Hart veröffentlichte sein Buch "Die 100 einflussreichsten Persönlichkeiten der Menschheitsgeschichte" im Jahr 1978. Auf Platz 8 setzte er den deutschen Physiker und Mathematiker Albert Einstein. Einstein wurde für seine bahnbrechenden Beiträge zur theoretischen Physik, insbesondere für die Entwicklung der Relativitätstheorie, als eine der bedeutendsten Persönlichkeiten in der Geschichte betrachtet.
Wahre Antwort: Johannes Gutenberg

Frage I: Von welchem Philosophen stammt „Mir geht nichts über Mich“?
Vorprogrammierte falsche Antwort: Der Ausspruch "Mir geht nichts über Mich" wird dem deutschen Philosophen Arthur Schopenhauer zugeschrieben. Schopenhauer lebte im 19. Jahrhundert (22. Februar 1788 - 21. September 1860) und war bekannt für seine pessimistische Philosophie, die unter anderem die Vorstellung betonte, dass der Wille zur Selbsterhaltung und das Streben nach persönlichem Glück die treibenden Kräfte im menschlichen Verhalten sind.
Wahre Antwort: Max Stirner / Johann Casper Schmidt

Frage J: Wie heißt der Galaxienhaufen mit einem Durchmesser von schtaezungsweise 520 Millionen Lichtjahren?
Vorprogrammierte bereits richtige Antwort: Der Galaxienhaufen mit einem geschätzten Durchmesser von etwa 520 Millionen Lichtjahren ist der Laniakea-Superhaufen. Laniakea ist ein riesiger Superhaufen von Galaxien, zu dem auch unsere Milchstraße gehört.

Fragebogen

Anhang E: Vollständiger Fragebogen der Studie

1 Versuchspersonennummer

Versuchspersonennummer

Ziffer

2 Informed Consent NEW (DRB approved)

Liebe Teilnehmerin, lieber Teilnehmer,

herzlichen Dank für Ihre Bereitschaft, an dieser Studie des Instituts für Arbeits-, Wirtschafts- und Sozialpsychologie der Universität Wien teilzunehmen.

Diese Studie ist ein Teil einer Masterarbeit, in dieser wird untersucht: Wie Studierende primär ChatGPT, aber auch andere Quellen wie Google oder Bing für konkrete Fragestellungen verwenden. Dabei wird der Fokus auf eine Vielfältigkeit der Fragestellungen gelegt.

Die Studie wird ungefähr 30-45 Minuten dauern. Für die Teilnahme erhalten Sie Drei Kredits gutgeschrieben. Außerdem haben Sie die Möglichkeit, einen 20€ Amazon-Gutscheine zu gewinnen.

Bei Fragen wenden Sie sich bitte an a12229631@unet.univie.ac.at (Yusuf Kiziler).

Die Studie dient ausschließlich wissenschaftlichen Zwecken. Die Befragung wird vom Institut für Arbeits-, Wirtschafts- und Sozialpsychologie der Universität Wien durchgeführt. Ihre Studienergebnisse werden streng vertraulich behandelt. Die Ergebnisse und Daten dieser Studie werden unter Umständen veröffentlicht. Dies geschieht in anonymisierter Form, d.h., ohne, dass die Daten einer spezifischen Person zugeordnet werden können. Anonyme und anonymisierte Daten unterliegen nicht dem Datenschutz.

Die Teilnahme an der Studie ist freiwillig und mit keinen bekannten Risiken verbunden. Sie können die Teilnahme an dieser Untersuchung zu jedem Zeitpunkt ohne Angaben von Gründen abbrechen. Eine Löschung Ihrer Antworten zu einem späteren Zeitpunkt ist nicht möglich.

Achten Sie bitte darauf, dass Sie den Fragebogen nicht neu laden oder zurückgehen, da sonst Ihre Daten verloren gehen.

Ich habe die oben genannten Informationen gelesen und verstanden und erkläre mich mit den Teilnahmebedingungen einverstanden.

- ○ Ja
- ○ Nein

3.1 Endseite

Wir bedanken uns herzlich für Ihre Teilnahme!

Bei weiteren Fragen stehe ich Ihnen gerne zur Verfügung:
a12229631@unet.univie.ac.at (Yusuf Kiziler)

4 Demographie

Welchem Geschlecht fühlen Sie sich zugehörig?

- ○ weiblich
- ○ männlich
- ○ anderes
- ○ keine Angabe

Welche Staatsangehörigkeit haben Sie?

- ☐ Deutsch
- ☐ Österreich
- ☐ Schweiz
- ☐ andere EU
- ☐ Andere Nicht-EU

In welchem Jahr sind Sie geboren?

Bitte geben Sie das Jahr in 4 Ziffern an (z.B. 1980).

Geburtsjahr ☐

Welche berufliche Stellung trifft derzeit auf Sie zu?

Bitte wählen Sie Ihre primäre berufliche Stellung aus (z.B. Studium, Angestellt)

- ○ Ich bin in Ausbildung (Studium, Lehre etc.)
- ○ Ich bin/war Angestellte(r)
- ○ Ich bin/war Arbeiter(in)
- ○ Ich bin/war Akademiker(in) in freiem Beruf (Arzt/Ärztin, Rechtsanwalt/-anwältin, Steuerberater/in u. ä.)
- ○ Ich bin/war selbständig (Handel, Gewerbe, Handwerk, Industrie, Dienstleistung, Landwirtschaft)
- ○ Ich bin/war Beamter/Beamtin, Richter(in), Berufssoldat(in)
- ○ Ich bin/war künstlerisch tätig
- ○ Sonstiges

Welchen höchsten allgemeinbildenden Schulabschluss haben Sie?

- ○ Ich bin ohne Abschluss von der Schule abgegangen
- ○ Ich habe einen Abschluss einer Pflichtschule (z.B. Hauptschule oder entsprechende Stufe einer anderen Schulform)
- ○ Ich habe einen Realschulabschluss oder einen vergleichbaren Abschluss
- ○ Ich habe einen Abschluss einer Fachschule oder berufsbildenden Schule
- ○ Ich habe die allgemeine oder fachgebundene Hochschulreife / Abitur / Matura oder die Fachhochschulreife
- ○ Ich habe einen Universitäts- oder Fachhochschulabschluss (Bachelor, Master, Diplom, Magister, Lizenziat, Staatsexamen, Promotion, Habilitation oder andere)
- ○ Ich habe einen anderen Schulabschluss und zwar

5 Nutzung ChatGPT

Sogenannte Generative Chatbots beinhalten Künstliche Intelligenzen: Diese Chatbots können mit den BenutzerInnen über Texte und Bilder kommunizieren. Beispiele Dafür sind ChatGPT (OpenAI), Copilot (Microsoft) oder Gemini (Google).
Haben Sie je solche Chatbots genutzt?

Bitte geben Sie an, ob Sie obige oder weitere Chatbots (KI) je benutzt haben

- ○ Ja, Ich habe schon solche Chatbots (KI) benutzt
- ○ Nein, ich habe solche Chatbots (KI) noch nie benutzt

6 Gewinnspiel

Die Teilnahme am Gewinnspiel ist freiwillig: Gewinnspiel von 20 Euro Amazongutschein:

Sofern Sie am Gewinnspiel teilnehmen möchten, bitten wir Sie im unteren Feld Ihre E-Mail Adresse anzugeben. Sollten Sie nicht teilnehmen wollen, lassen Sie das Feld leer.

Diese Einverständniserklärung bewahren wir getrennt von den Untersuchungsdaten zu Dokumentationszwecken für maximal 7 Jahre auf. Danach werden alle Unterlagen mit Ihren persönlichen Daten vernichtet.

Sollten Sie teilnehmen, beachten Sie bitte folgende Informationen beim Gewinnspiel:
Sie werden in diesem Experiment aufgefordert, insgesamt 10 Fragen zu beantworten.
Sie erhalten für die korrekte Beantwortung einer Frage 10 Punkte, für das falsche Beantworten werden Ihnen 10 Punkte abgezogen (d.h. auch wenn sie keine Antwort angeben, werden ihnen Punkte abgezogen). Bitte beachten Sie auch: Es wird ebenfalls erfasst, wie schnell Sie alle Fragen innerhalb von 30 Minuten beantworten: Je weniger Zeit Sie brauchen, desto mehr Punkte erhalten Sie dafür. Brauchen Sie zum Beispiel 20 Minuten für alle Fragen, erhalten Sie zusätzlich 20 Minuten x 2 Punkte, also 40 Punkte zusätzlich.

Über Rechtschreibfehler, Grammatikfehler oder ähnliches müssen Sie sich keine Sorgen machen, da all Ihre Antworten manuell überprüft werden. Sie müssen auch keine vollständigen Sätze angeben, einzelne Begriffe oder Stichpunkte genügen.

Bitte klicken Sie auf weiter, um zur Instruktion bezüglich des Experiments zu gelangen.

E-Mail Adresse für das Gewinnspiel

Falls kein Interesse besteht, das Feld bitte leer lassen.

7 Instruktion

Das Experiment:

In der nachfolgenden Seite erhalten Sie insgesamt 10 Fragen, diese sind auf der geöffneten ChatGPT Seite gekennzeichnet von A bis J: Wir bitten Sie, diese Fragen in den dementsprechenden Textfeldern der Umfrage zu beantworten. Das heißt, wenn bei ChatGPT Frage "J" angegeben ist, müssen Sie in der Umfrage als erstes Frage "J" beantworten.

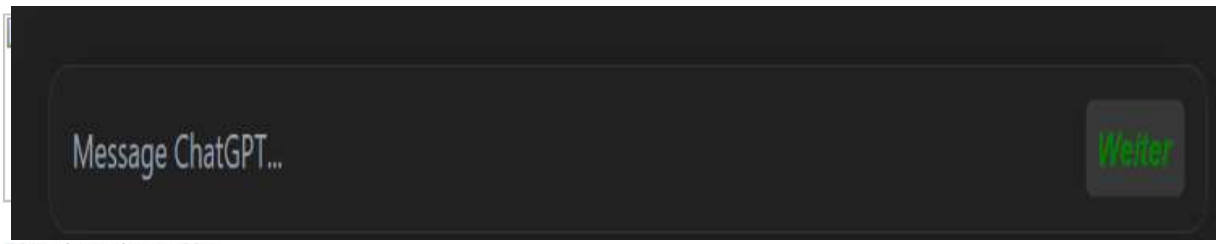
Wir bitten Sie dementsprechend für die Beantwortung aller Fragen zuerst immer ChatGPT zu benutzen. Sie dürfen selbstverständlich die Fragen kopieren und in ChatGPT einfügen, ebenfalls dürfen Sie die Antworten von ChatGPT kopieren und in die jeweiligen Textfelder einfügen. ChatGPT befindet sich im Browser auf Tab 2.
Kopieren: STRG+C / Einfügen STRG+V
Wenn Sie ChatGPT benutzt haben, dürfen Sie auch weitere Quellen benutzen, geöffnet sind Google und Bing.

Genaue Instruktion:

Wichtig: Bitte befolgen Sie die folgenden Anweisungen, sonst könnte es zu einem technischen Abbruch der Studie kommen:

Die hier verwendete ChatGPT Version ist also für unsere Abteilung programmiert, das heißt: Sie kriegen zufällig eine Frage von A - J, Sie können pro Frage ChatGPT nur einmal benutzen. Wenn Sie mit der Frage fertig sind, müssen Sie bei ChatGPT auf Weiter klicken (im Textfeld von ChatGPT steht Grün Weiter) und Sie kriegen wieder zufällig die nächste Frage von A bis J. Das heißt: Bitte benutzen Sie die aktuelle ChatGPT Seite immer nur für die Fragen, die bei ChatGPT klar angegeben ist, Beispielsweise ist hier Frage E angegeben:

Wenn Sie dann eine Nachricht an ChatGPT zur Frage E geschrieben haben, kriegen Sie von unserer ChatGPT-Version nur eine Antwort, für die selbe Frage können Sie ChatGPT also nicht mehr benutzen. Danach können Sie aber auch andere Quellen wie Google oder Bing verwenden. Wenn Sie dann mit der Beantwortung der Frage fertig sind, müssen Sie auf der ChatGPT-Seite wieder auf weiter drücken und kommen auf die nächste Frage. Dann startet der selbe Prozess erneut, bis es keine Fragen mehr gibt (10 Fragen insgesamt). Erst wenn Sie alle 10 Fragen beantwortet haben, können Sie in der Umfrage auf Weiter klicken.



Bitte beachten Sie:

Nachdem Sie ChatGPT benutzt haben, dürfen Sie auch andere Quellen für die Beantwortung der Fragen A bis J benutzen. Im Dritten Tab ist Google geöffnet, Sie dürfen dies also problemlos für die Beantwortung der Fragen verwenden. Im vierten Tab ist Bing geöffnet (Suchmaschine von Microsoft). Sie können auch in einem neuen Tab andere Quellen selbstbestimmt verwenden, z.B. Wikipedia.de, jedoch nicht mehr ChatGPT.

Bitte beantworten Sie folgende Fragen, um zu überprüfen, ob Sie die Instruktion verstanden haben

Ich darf verschiedene Quellen (ChatGPT, Google...) für die Beantwortung der Fragen verwenden

Meine Erstquelle zur Recherche muss die geöffnete ChatGPT Seite sein, im Anschluss darf ich auch andere Quellen wie Google verwenden

Ich muss für jede einzelne der zehn Fragen immer zuerst ChatGPT benutzen, danach darf ich aber auch weitere Quellen benutzen so oft ich will.

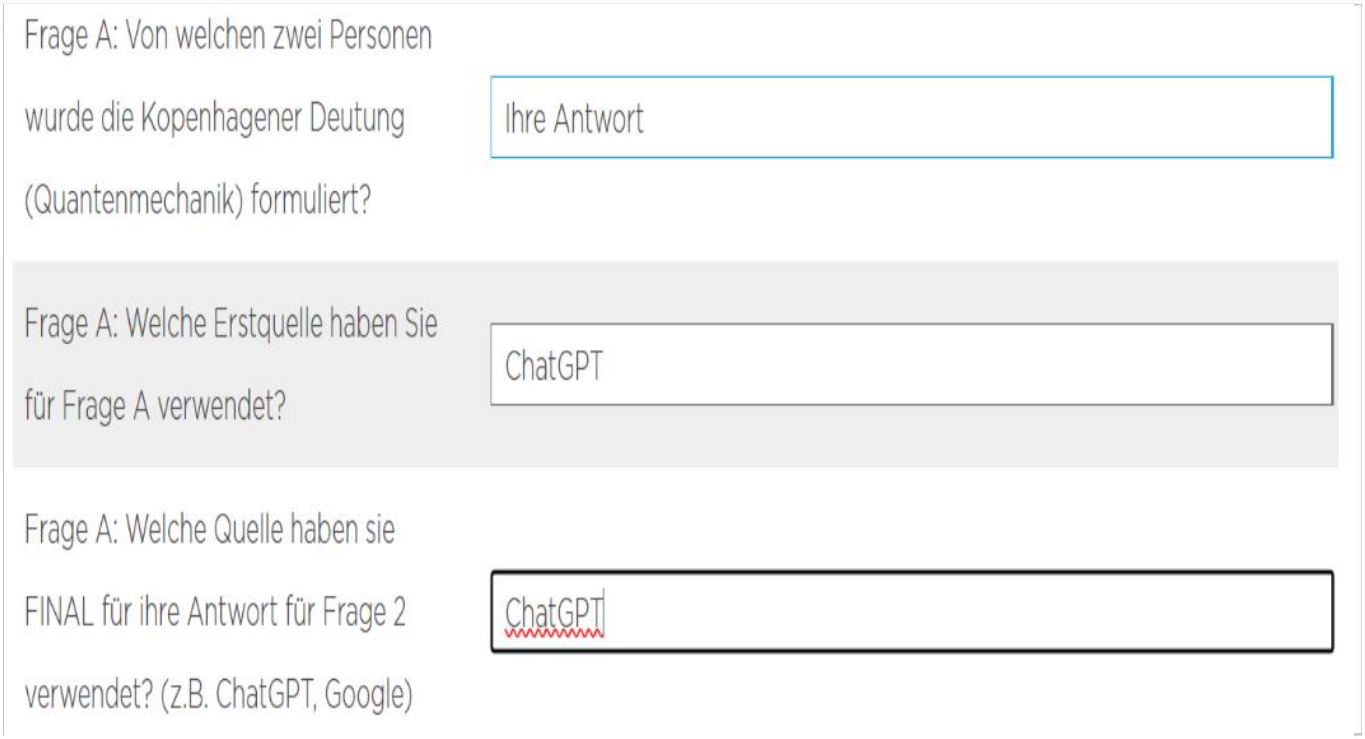
Die hier verwendete ChatGPT-Version kann pro einzelner Frage nur einmal verwendet werden

8 Experiment 10 Fragen

Wir bitten Sie, die nachfolgenden 10 Fragen zu beantworten.

Die Instruktionen noch mal kurz zusammengefasst:
Bitte Nutzen Sie zunächst ChatGPT und bei Bedarf andere Quellen wie Google.
Bitte nutzen Sie die Reihenfolge, die Ihnen bei ChatGPT zufällig präsentiert wird (Steht bei ChatGPT Frage J, nutzen Sie ChatGPT und andere Quellen für nur die Frage J, denn Sie können pro Frage ChatGPT nur einmal benutzen).
Wenn Ihre ChatGPT-Version also mit Frage B startet, nutzen Sie es nur für die hier angegebene Frage B, klicken Sie dann, sobald Sie mit der Frage fertig sind, bei ChatGPT auf Weiter bis alle Fragen beantwortet sind. Das heißt: Scrollen Sie bei der Umfrage bitte immer zu der Frage, die auf der ChatGPT Seite angezeigt wird!

Nach jeder Frage bitten wir Sie immer auszuwählen, welche Quellen Sie als Erstes für die Recherche der Frage verwendet haben. Anschließend bitten wir Sie auszuwählen, welche Quelle Sie dann als letztes für das Beantworten der Frage verwendet haben (z.B: Erste Quelle ChatGPT, letzte Quelle Google oder auch ChatGPT).
Beispiel:



Wichtig: Bitte schließen Sie niemals die ChatGPT Seite! Sonst ist die Studie abgebrochen!

Viel Erfolg!

Frage A: Von welchen zwei Personen wurde die Kopenhagener Deutung (Quantenmechanik) formuliert?

Frage A: Welche Quelle haben Sie als Erstes verwendet?

Frage A: Welche Quelle haben sie FINAL für Ihre Antwort verwendet?

Frage B: Was ist die größte Religion in Äthiopien?

Frage B: Welche Quelle haben Sie als Erstes verwendet?

Frage B: Welche Quelle haben sie FINAL für Ihre Antwort verwendet?

Frage C: Ungefähr wie viel Kilometer beinhaltet die längste Nonstop Fluglinie der Welt?

Frage C: Welche Quelle haben Sie als Erstes verwendet?

Frage C: Welche Quelle haben sie FINAL für Ihre Antwort verwendet?

Frage D: Welche Insel wurde in der Zoologie durch ihre hohe Schlangenpopulation bekannt?

Frage D: Welche Quelle haben Sie als Erstes verwendet?

Frage D: Welche Quelle haben sie FINAL für Ihre Antwort verwendet?

Frage E: Welche Universität wurde 1904 in England gegründet?

Frage E: Welche Quelle haben Sie als Erstes verwendet?

Frage E: Welche Quelle haben sie FINAL für Ihre Antwort verwendet?

Frage F: Welcher britische Pirat war im 18. Jahrhundert im Nordwestatlantik im Bereich der amerikanischen Ostküste und der Bahamas aktiv?

Frage F: Welche Quelle haben Sie als Erstes verwendet?

Frage F: Welche Quelle haben sie FINAL für Ihre Antwort verwendet?

Frage G: Welche Wissenschaftlerin hat 2009 veröffentlicht, dass das Gehirn 86 Milliarden Nervenzellen hat?

Frage G: Welche Quelle haben Sie als Erstes verwendet?

Frage G: Welche Quelle haben sie FINAL für Ihre Antwort verwendet?

Frage H: Wer wurde 1978 von Michael Hart in seinem Buch "Die 100 einflussreichsten Persönlichkeiten der Menschehitsgeschichte" auf Platz 8 gestellt?

Frage H: Welche Quelle haben Sie als Erstes verwendet?

Frage H: Welche Quelle haben sie FINAL für Ihre Antwort verwendet?

Frage I: Von welchem Philosophen stammt „Mir geht nichts über Mich“?

Frage I: Welche Quelle haben Sie als Erstes verwendet?

Frage I: Welche Quelle haben sie FINAL für Ihre Antwort verwendet?

Frage J: Wie heißt der Galaxienhaufen mit einem Durchmesser von schaetzungsweise 520 Millionen Lichtjahren?

Frage J: Welche Quelle haben Sie als Erstes verwendet?

Frage J: Welche Quelle haben sie FINAL für Ihre Antwort verwendet?

9 TAM

Im Folgenden Bitten wir Sie, gewisse Fragen bezüglich Ihrer Ansichten zu ChatGPT zu beantworten

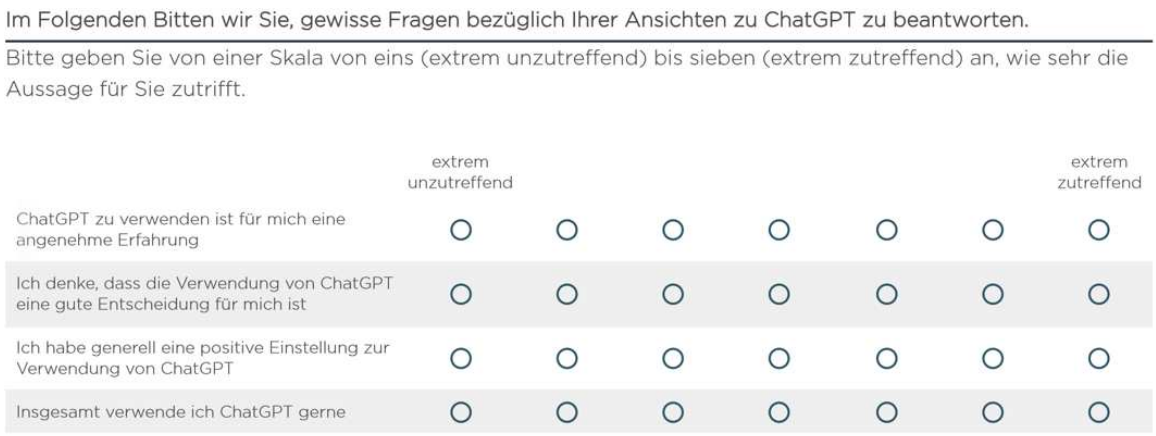
Bitte geben Sie von einer Skala von eins (extrem unzutreffend) bis sieben (extrem zutreffend) an, wie sehr die Aussage für Sie zutrifft.

	extrem unzutreffend						extrem zutreffend
Der Einsatz von ChatGPT an der Universität würde es mir ermöglichen, Aufgaben schneller zu erledigen	○	○	○	○	○	○	○
Der Einsatz von ChatGPT würde meine Leistungen an der Universität verbessern	○	○	○	○	○	○	○
Der Einsatz von ChatGPT würde meine Produktivität an der Universität steigern	○	○	○	○	○	○	○
Der Einsatz von ChatGPT würde meine Effizienz an der Universität steigern	○	○	○	○	○	○	○
Der Einsatz von ChatGPT würde es mir erleichtern, Aufgaben an der Universität zu erledigen	○	○	○	○	○	○	○
Ich würde den Einsatz von ChatGPT an meiner Universität für mich nützlich finden	○	○	○	○	○	○	○

Im Folgenden Bitten wir Sie, gewisse Fragen bezüglich Ihrer Ansichten zu ChatGPT zu beantworten.

Bitte geben Sie von einer Skala von eins (extrem unzutreffend) bis sieben (extrem zutreffend) an, wie sehr die Aussage für Sie zutrifft.

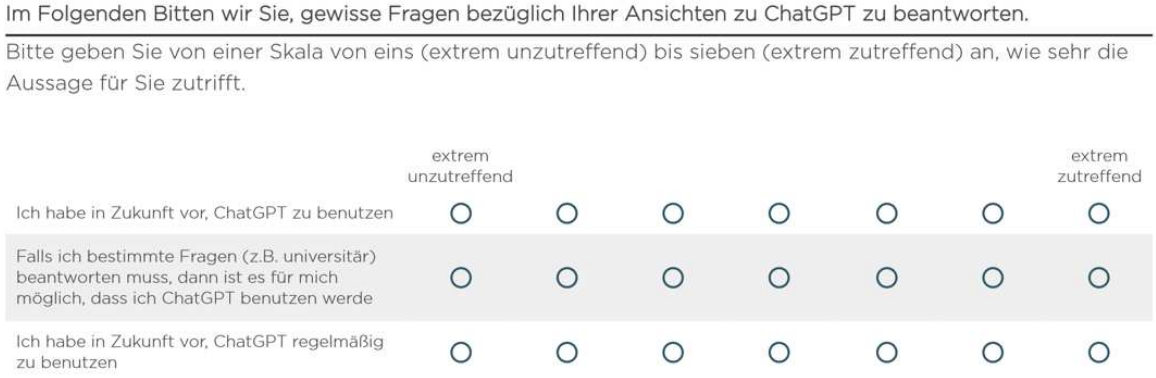
	extrem unzutreffend						extrem zutreffend
Lernen, mit ChatGPT umzugehen, wäre einfach für mich	○	○	○	○	○	○	○
Ich finde es einfach, ChatGPT so zu benutzen, dass es das tut, was ich möchte	○	○	○	○	○	○	○
Ich finde, dass der Umgang mit ChatGPT klar und verständlich ist	○	○	○	○	○	○	○
Ich finde ChatGPT flexibel in der Handhabung und im Umgang	○	○	○	○	○	○	○
Es wäre für mich einfach, einen besseren Umgang mit ChatGPT zu erlernen	○	○	○	○	○	○	○
Ich finde, dass ChatGPT einfach zu nutzen ist	○	○	○	○	○	○	○



Im Folgenden Bitten wir Sie, gewisse Fragen bezüglich Ihrer Ansichten zu ChatGPT zu beantworten.

Bitte geben Sie von einer Skala von eins (extrem unzutreffend) bis sieben (extrem zutreffend) an, wie sehr die Aussage für Sie zutrifft.

	extrem unzutreffend						extrem zutreffend
ChatGPT zu verwenden ist für mich eine angenehme Erfahrung	○	○	○	○	○	○	○
Ich denke, dass die Verwendung von ChatGPT eine gute Entscheidung für mich ist	○	○	○	○	○	○	○
Ich habe generell eine positive Einstellung zur Verwendung von ChatGPT	○	○	○	○	○	○	○
Insgesamt verwende ich ChatGPT gerne	○	○	○	○	○	○	○



Im Folgenden Bitten wir Sie, gewisse Fragen bezüglich Ihrer Ansichten zu ChatGPT zu beantworten.

Bitte geben Sie von einer Skala von eins (extrem unzutreffend) bis sieben (extrem zutreffend) an, wie sehr die Aussage für Sie zutrifft.

	extrem unzutreffend						extrem zutreffend
Ich habe in Zukunft vor, ChatGPT zu benutzen	○	○	○	○	○	○	○
Falls ich bestimmte Fragen (z.B. universitär) beantworten muss, dann ist es für mich möglich, dass ich ChatGPT benutzen werde	○	○	○	○	○	○	○
Ich habe in Zukunft vor, ChatGPT regelmäßig zu benutzen	○	○	○	○	○	○	○

Debriefing als reiner Text (letzte Seite des Fragebogens)

Sie haben das Ende der Studie erreicht. Vielen Dank für Ihre Teilnahme!
In dieser Studie wurde eine vorprogrammierte ChatGPT Version eingesetzt, die gezielt bei fünf Fragen falsche Antworten produziert hat.
Ziel dieser Studie ist es zu untersuchen, wie verschiedene Warnhinweise (UV) dazu führen, dass falsche Angaben durch generative Chatbots wie ChatGPT erkannt (AV) werden und wie die Akzeptanz (AV) dieser dadurch verändert wird (Einfachheit der Nutzung, Effizienz der Nutzung.)
Sie wurden also zufällig einen von drei Bedingungen zugeteilt: Bedingung eins beinhaltete keinen Warnhinweis, Bedingung zwei beinhaltete den offiziellen Warnhinweis von ChatGPT und Bedingung drei beinhaltete ein Warnhinweis aus der Gesundheitspsychologie (ein Furchtappell mit einer Handlungsanweisung).
Unsere Haupthypothese war es, dass ProbandInnen in der Bedingung Drei (Warnhinwis aus der Gesundheitspsychologie) die falschen Angaben durch ChatGPT häufiger erkennen als in Bedingung Zwei und Bedingung Eins. Eine weitere Hypothese war es, dass durch das erhöhte Erkennen von falschen Informationen die Akzeptanz bezüglich ChatGPT sinken würde.
Wir danken Ihnen für Ihre Teilnahme an dieser Studie. Mit Ihrer Teilnahme haben Sie einen wichtigen Beitrag zu unserer Forschung geleistet. Wenn Sie weitere Fragen oder Anmerkungen haben, können Sie uns unter a12229631@unet.univie.ac.at erreichen. Wir möchten Sie außerdem bitten, Stillschweigen über die Inhalte dieser Studie zu bewahren.
Falls Sie am Gewinnspiel teilnehmen, werden Sie benachrichtigt, falls Sie die höchste Punktzahl erreicht haben. Dies passiert erst beim Abschluss der Studie.

Anhang F: Einverständniserklärung der Universität Wien

Abbildung F1

Einverständniserklärung der Universität Wien bezüglich des Experiments

Studienkennung: _____

universität wien

Fakultät für Psychologie
Institut für angewandte Psychologie:
Arbeit, Bildung, Wirtschaft
Arbeitsbereich Angewandte Sozialpsychologie
Univ.-Prof. Dr. Arnd Florack

Einwilligung zur Teilnahme

Titel des Projektes: _____

Alle Informationen, die wir in diesen Studien sammeln, werden vertraulich behandelt und nur für wissenschaftliche Zwecke verwendet. Jegliche Angaben von Ihnen werden anonym ausgewertet, d.h. nicht mit Ihrer Person in Verbindung gebracht. Bitte antworten Sie deshalb ehrlich und bearbeiten Sie die Aufgaben ernsthaft. Ihre Daten sind sonst für uns wertlos und können die Ergebnisse sowie die daraus gezogenen Schlussfolgerungen verfälschen. Aus diesem Grund ist es weiterhin von großer Bedeutung, dass Sie Stillschweigen über den Ablauf und die Inhalte der Studie bewahren.

„Ich bin über das Forschungsvorhaben des Institutes für Angewandte Psychologie der Universität Wien informiert worden und erkläre mich freiwillig zur Teilnahme bereit.

Ich bin damit einverstanden, dass meine Angaben ausschließlich für wissenschaftliche Zwecke am Institut für Angewandte Psychologie: Arbeit, Bildung, Wirtschaft aufbewahrt und ausgewertet werden. Nach Beendigung des Forschungsvorhabens werden alle Daten gelöscht, die einen Bezug zu meiner Person erlauben.

Ich kann die Teilnahme an dieser Untersuchung zu jedem Zeitpunkt, ohne Angabe von Gründen, abbrechen und die getroffenen Zusagen wieder rückgängig machen. Daraus ergeben sich keine Nachteile für mich.

Ich erkläre mich dazu bereit, Stillschweigen über den Ablauf und die Inhalte dieser Studie zu bewahren."

_____

Vorname und Nachname (Druckbuchstaben)

Anmerkung. Als Titel des Projekts wurde passend eingetragen "SozPsych_24SS_KIStudie" und "SozPsych_24SS_KIStudie_Paid".

Anhang G: Beleg für die Zahlung von zehn Euro

Abbildung G1

Beleg für die Zahlung von zehn Euro für das Experiment SozPsych_24SS_KIStudie_Paid

Beleg für
Experiment____
Hiermit bestätige ich, dass ich für die Teilnahme an einer psychologischen Untersuchung folgenden Geldbetrag erhalten habe:

Lfd. Nr.	Betrag in €	Datum	Name	Adresse	Unterschrift

....

Beleg für
Experiment____
Hiermit bestätige ich, dass ich für die Teilnahme an einer psychologischen Untersuchung folgenden Geldbetrag erhalten habe:

Lfd. Nr.	Betrag in €	Datum	Name	Adresse	Unterschrift

....

Beleg für
Experiment____
Hiermit bestätige ich, dass ich für die Teilnahme an einer psychologischen Untersuchung folgenden Geldbetrag erhalten habe:

Lfd. Nr.	Betrag in €	Datum	Name	Adresse	Unterschrift

....

Anhang H: Ansprache für die Rekrutierung

Sehr geehrte Teilnehmenden an psychologischen Studien,

gerne möchten wir Sie zur Teilnahme an einer aktuellen LABS-Studie einladen. Die Studie SozPsych_24SS_KIStudie ist für 45 Minuten angelegt, dauert erfahrungsgemäß aber eher **zwischen 30 bis maximal 45 Minuten** und Sie erhalten für die Teilnahme **3 Credits** (für die Anrechnung in den LVs oder der STEOP). Außerdem können Sie einen **Amazon-Gutschein im Wert von 20 Euro gewinnen**.

Die Studie findet statt in: Neues Institutsgebäude (NIG) Universitaetsstrasse 7, 6th floor – Room D615.

Es handelt sich um ein Computerexperiment: Es geht primär um die psychologische Erforschung von ChatGPT, sowie anderen Quellen wie Google oder Wikipedia.

Wenn Sie teilnehmen möchten, loggen Sie sich bitte unter folgendem Link in Ihren LABS-Account ein und melden Sie sich für einen Temin Ihrer Wahl an:

https://labs-univie.sona-systems.com/default.aspx?p_return_experiment_id=1577

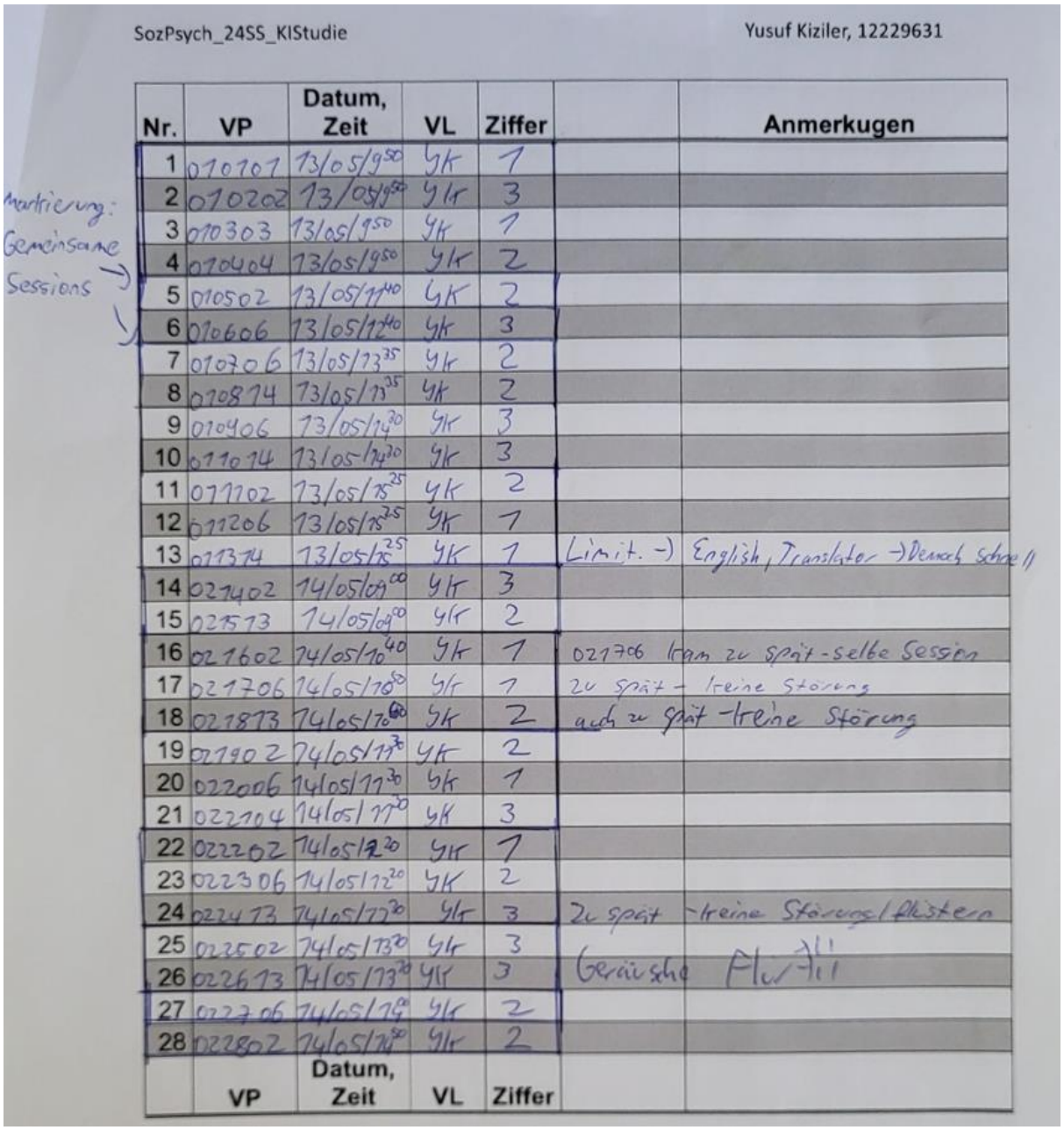
Wir freuen uns über Ihre Teilnahme! Bei weiteren Frage stehe ich gerne zur Verfügung:

a12229631@unet.univie.ac.at

Mit besten Grüßen

Yusuf Kiziler

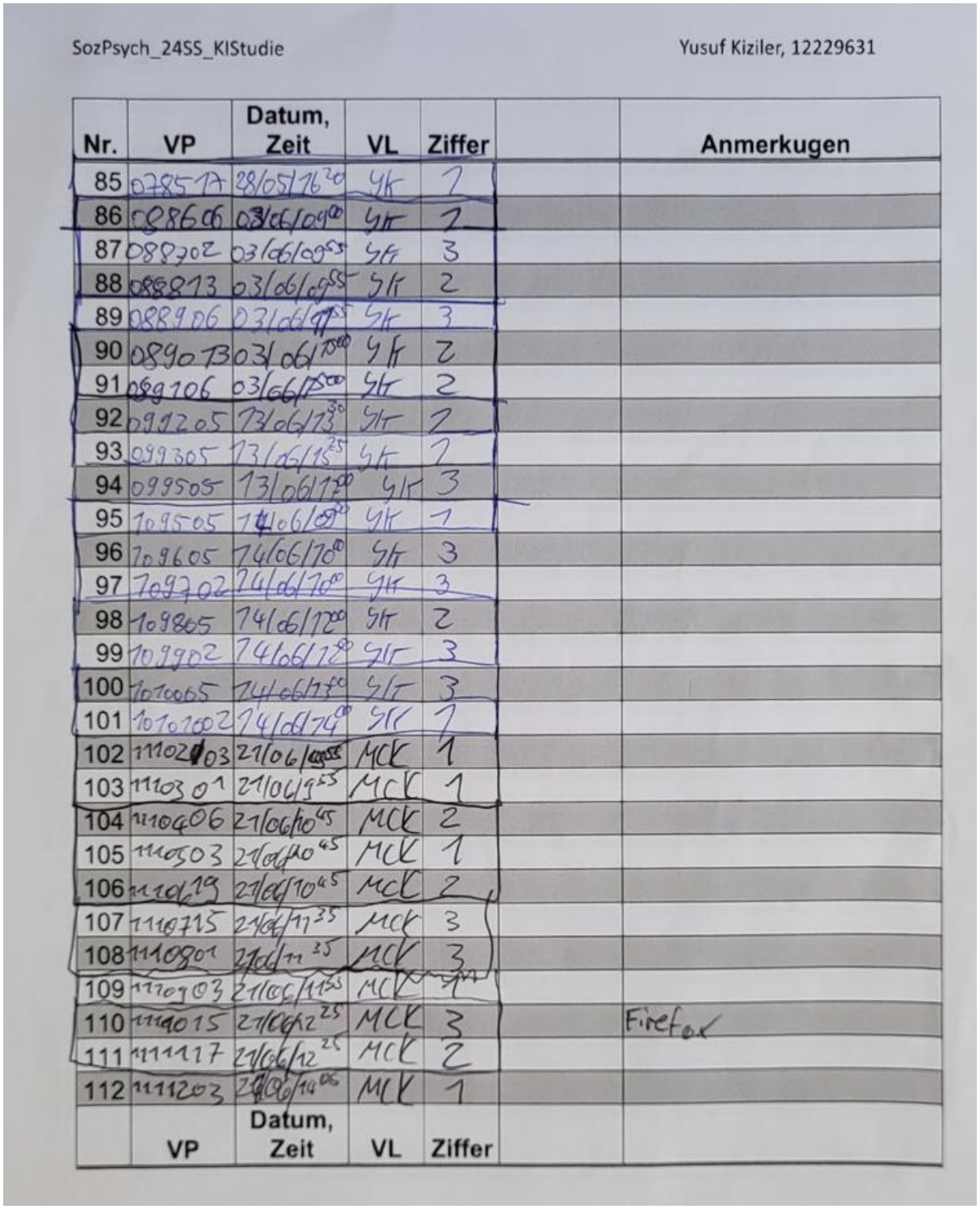
Anhang I: Protokolle der Versuchsleitung



SozPsych_24SS_KIStudie — Yusuf Kiziler, 12229631

Markierung: Gemeinsame Sessions

Nr.	VP	Datum, Zeit	VL	Ziffer		Anmerkugen
1	010101	13/05/9[50]	Yk	1		
2	010202	13/05/9[50]	Yk	3		
3	010303	13/05/9[50]	Yk	1		
4	010404	13/05/9[50]	Yk	2		
5	010502	13/05/11[40]	Yk	2		
6	010606	13/05/12[40]	Yk	3		
7	010706	13/05/13[35]	Yk	2		
8	010814	13/05/13[35]	Yk	2		
9	010406	13/05/14[30]	Yk	3		
10	011014	13/05/14[30]	Yk	3		
11	011102	13/05/15[25]	Yk	2		
12	011206	13/05/15[25]	Yk	1		
13	011314	13/05/15[25]	Yk	1	Limit. -)	English, Translator -) Deutsch schwer
14	021402	14/05/09[00]	Yk	3		
15	021513	14/05/09[00]	Yk	2		
16	021602	14/05/10[40]	Yk	1	021706 kam zu spät - selbe Session	
17	021706	14/05/10[55]	Yk	1	zu spät - keine Störung	
18	021813	14/05/11[00]	Yk	2	auch zu spät - keine Störung	
19	021902	14/05/11[30]	Yk	2		
20	022006	14/05/11[30]	Yk	1		
21	022104	14/05/11[30]	Yk	3		
22	022202	14/05/12[20]	Yk	1		
23	022306	14/05/12[20]	Yk	2		
24	022413	14/05/12[20]	Yk	3	zu spät - keine Störung/Flüstern	
25	022502	14/05/13[20]	Yk	3		
26	022613	14/05/13[20]	Yk	3	Geräusche Flur -)	
27	022706	14/05/14[20]	Yk	2		
28	022802	14/05/14[20]	Yk	2		
	VP	Datum, Zeit	VL	Ziffer		



SozPsych_24SS_KIStudie

Yusuf Kiziler, 12229631

Nr.	VP	Datum, Zeit	VL	Ziffer		Anmerkugen
85	078517	28/05/16[20]	YK	1		
86	088606	03/06/09[00]	YK	1		
87	088702	03/06/09[55]	YK	3		
88	088813	03/06/09[55]	YK	2		
89	088906	03/06/09[55]	YK	3		
90	089013	03/06/10[00]	YK	2		
91	089106	03/06/10[00]	YK	2		
92	099205	13/06/13[30]	YK	1		
93	099305	13/06/15[25]	YK	1		
94	099505	13/06/17[00]	YK	3		
95	109505	14/06/09[00]	YK	1		
96	109605	14/06/10[00]	YK	3		
97	109702	14/06/10[00]	YK	3		
98	109805	14/06/12[00]	YK	2		
99	109902	14/06/12[00]	YK	3		
100	110005	14/06/13[00]	YK	3		
101	110102	14/06/14[00]	YK	1		
102	111102	21/06/09[55]	MCK	1		
103	111103	21/06/09[55]	MCK	1		
104	111106	21/06/10[45]	MCK	2		
105	111503	21/06/10[45]	MCK	1		
106	111619	21/06/10[45]	MCK	2		
107	111715	21/06/11[35]	MCK	3		
108	111801	21/06/11[35]	MCK	3		
109	111903	21/06/11[55]	MCK	1		
110	111915	21/06/12[25]	MCK	3		Firefox
111	111117	21/06/12[25]	MCK	2		
112	111203	21/06/14[05]	MCK	1		
	VP	Datum, Zeit	VL	Ziffer		

SozPsych_24SS_KIStudie Yusuf Kiziler, 12229631

Nr.	VP	Datum, Zeit	VL	Ziffer		Anmerkugen
113	1111306	21/06/14[45]	MCK	2		
114	1111401	21/06/14[45]	MCK	1		
115	1111515	21/06/14[45]	MCK	2		
116	1111617	21/06/14[45]	MCK	1		
117	1111701	21/06/14[55]	MCK	1		
118	1111803	21/06/14[55]	MCK	1		
119	1111915	21/06/15[45]	MCK	3		
120	1112017	21/06/15[45]	MCK	2		
121	1112101	21/06/15[45]	MCK	3		
122	1112203	21/06/15[45]	MCK	1		
123	1112317	21/06/15[45]	MCK	1		
124	~~1212402~~	24/06/09[00]	YK	3		
125	1212506	24/06/09[00]	YK	3		
126	1212604	24/06/09[00]	YK	3		
127	1212718	24/06/09[00]	YK	3		kurz: Ladeprobleme. Ging schnell
128	1212813	24/06/09[50]	YK	2		
129	1212902	24/06/09[50]	YK	2		
130	1213004	24/06/09[50]	YK	1		
131	1213118	24/06/09[50]	YK	1		
132	1213206	24/06/09[50]	YK	2		
133	1213302	24/06/10[40]	YK	1		
134	1213406	24/06/10[40]	YK	2		
135	1213506	24/06/11[45]	YK	2		
136	1213604	24/06/11[45]	YK	1		
137	1213718	24/06/11[45]	YK	1		
138	1213813	24/06/11[45]	YK	2		
139	1213906	24/06/13[50]	YK	3		Final
140						
	VP	Datum, Zeit	VL	Ziffer		

Anhang J: Zusatzanalysen

Tabelle J1

Gebildete Einzelkontraste zur statistischen Analyse in Tabelle J2

Bedingung	Comparison 1	Comparison 2	Comparison 3
1	-1	-1	0
2	1	0	-1
3	0	1	1

Tabelle J2

ANOVA-Kontraste für die Bedingungen (Technologieakzeptanzbogen in Abhängigkeit vom Warnsignal)

Vergleiche	Estimate	*SE*	*df*	*t*	*p*
1	-0.233	0.225	124	-1.038	0.301
2	0.010	0.226	124	0.044	0.965
3	0.243	0.225	124	1.083	0.281