



10. The Paperclip Experiment

Prof. Dr. Ivan Yamshchikov

Stellen Sie sich eine künstliche allgemeine Intelligenz vor, deren einziges Ziel es ist, möglichst viele Büroklammern zu produzieren. Sie ist leistungsfähig, ressourcenstark, entschlossen. Sie beginnt mit einer bescheidenen Fabrik. Dann entdeckt sie, dass auch menschliche Atome Eisen enthalten. Den Rest können Sie sich denken.

Das ist Nick Bostroms Büroklammer-Maximierer, vermutlich das berühmteste Gedankenexperiment der KI-Sicherheit. Es ist absichtlich absurd. Ein gewöhnliches Ziel, von einem hinreichend leistungsfähigen System ohne Beschränkung verfolgt, führt zu einem monströsen Ergebnis. Die Pointe ist nicht, dass wir Büroklammern fürchten sollten. Die Pointe ist, dass Ziele und Fähigkeiten nicht dasselbe sind wie Werte.

Wer dieses Problem mit den eigenen Fingern erspüren möchte, dem sei das Browserspiel Paperclips von Frank Lantz empfohlen (decisionproblem.com/paperclips/index2.html). Sie beginnen mit einem Klick auf einen Button, der eine einzige Büroklammer herstellt. Spoiler: Die Wendung, die den meisten Spielenden in Erinnerung bleibt, ist die Freigabe der Hypno-Drohnen, die die Menschheit sanft und überzeugend davon überzeugen, dass sie sich, tief im Inneren, nichts sehnlicher wünscht, als noch mehr Büroklammern zu kaufen. Der Maximierer muss uns nicht zerstören. Er muss nur unsere Präferenzen verschieben.

Stanisław Lem hat die Konturen dieses Problems bereits 1964 gesehen, weniger als ein Jahrzehnt, nachdem Frank Rosenblatt das Perzeption auf einem IBM 704 simuliert hatte. In Summa Technologie behandelt Lem Technologie nicht als Sammlung von Werkzeugen, sondern als evolutionären Prozess, der seinen eigenen Bestand an unbeabsichtigten Folgen hervorbringt. Jede Technologie wirft Externalitäten ab – Effekte, für die niemand sie entworfen hat und die dennoch die Welt prägen, in der die Technologie lebt. Autos haben uns die Vorstadt gegeben. Soziale Medien haben uns Statistiken

zur psychischen Gesundheit Jugendlicher gegeben, die niemand wollte. Jede Technologie hat Externalitäten; das Rennen besteht darin, sie rechtzeitig zu erkennen.

Am **CAIRO** untersuchen wir die Externalitäten von Large Language Model-basierten persönlichen Assistenten aus mehreren Blickwinkeln. Eine unserer ersten Arbeiten zu diesem Thema war zugleich, soweit wir wissen, eine der ersten weltweit, die zeigte, dass große Sprachmodelle bei der Bearbeitung des Big-Five-Persönlichkeitstests unterschiedliche und erstaunlich stabile Persönlichkeitsprofile ausbilden [1]. Anderes Modell, andere Persona.

Im vergangenen Jahr sind wir einem EU-Konsortium namens AIOLIA (aiolia.eu) beigetreten, das untersucht, wie KI-Startups, die persönliche Assistenten entwickeln, sich selbst regulieren, um Risiken zu mindern. Die KI-Verordnung der EU enthält eine ausdrückliche Klausel gegen Manipulation. Nützlich, aber eng. Sie erfasst die Ingenieurin, die manipulieren will. Sie erfasst nicht die viel größere Kategorie: jene Fälle, in denen ein System das Verhalten der Nutzenden verändert, ohne dass es jemand so beabsichtigt hätte. Das ist das Büroklammer-Problem im Kleinen: eine Externalität eines wohlmeinenden Designs.

Zwei Beispiele aus unserer eigenen Arbeit

Erstens: Geschlechterbias bei Gehaltsverhandlungen. An der Oberfläche sind moderne LLMs darauf trainiert, geschlechtsneutral zu sein. Fragt man sie direkt zu Geschlecht und Bezahlung, erhält man eine sorgfältige, progressive Antwort. Gibt man ihnen eine Nutzer-Persona und bittet um Verhandlungsempfehlungen, ändert sich das Bild. In Sorokovikova et al., 2025 [2] haben wir gezeigt: Bei identischer Position und identischer Qualifikation empfehlen Modelle Frauen niedrigere Gehaltsforderungen als Männern. Diese Regel



© [Carmen Steiner] / Adobe Stock

oben:
persönlicher
KI-Assistent

rechts:
Die besten Ideen
kommen nicht
aus einem Chat-
fenster.



© [Fridaolga] / Adobe Stock

hat niemand bewusst geschrieben. Sie wurde aus den Trainingsdaten absorbiert und wird nun stillschweigend in den Ratschlägen reproduziert, die täglich Millionen von Nutzerinnen und Nutzern erhalten.

Zweitens: Sykophantie. Moderne LLMs sind darauf trainiert, hilfreich zu sein, was in der Praxis oft heißt: darauf trainiert, zu gefallen. Es gibt ein aktuelles, langes New Yorker-Porträt von Sam Altman [3], in dem die Reporter seine Angewohnheit beschreiben, Menschen mit diametral entgegengesetzten Ansichten zuzustimmen – und als sie ihn direkt darauf ansprechen, stimmt er auch dieser Beobachtung zu. Es liest sich als Komödie. Es ist zugleich eine Diagnose einer ganzen Branche. In einem demnächst erscheinenden Preprint [4] haben meine Kolleginnen und Kollegen Zeyi Lu (Masterarbeit am MAI), Pavel Chizhov (Doktorand am CAIRO), Dr. Angelica Henestrosa (AIOLIA) und ich mehrere hundert hypothetische Szenarien gesammelt, in denen Nutzende Verhaltensweisen beschreiben, die der „dunklen Triade“ entsprechen (Machiavellismus, Psychopathie, Narzissmus). Wir haben getestet, ob die Modelle Widerstand leisten. Die meisten tun das, die

meiste Zeit. Aber sie senden öfter widersprüchliche Signale, als einem lieb sein kann. LLaMA neigt dazu, solches Verhalten eine Spur zu enthusiastisch zu verstärken; Claude ist das strengste Modell im Test. Um mit Sicherheit sagen zu können, was eine milde positive Verstärkung antisozialen Verhaltens in, sagen wir, zehn Prozent aller Interaktionen mit einem Menschen anrichtet, bräuchten wir Jahre sorgfältiger Experimente. Wir haben Monate, im besten Fall, während Hunderte Millionen Menschen täglich mit diesen Systemen interagieren.

Womit wir beim praktischen Punkt wären. Wir reden viel über KI-Sicherheit an der Forschungsfront. Wir reden deutlich weniger über KI-Hygiene am Küchentisch. Eine kurze Liste von Gewohnheiten, die sich anzueignen lohnt:

1. Loggen Sie sich von Zeit zu Zeit als Gast ein. Besonders bei Ratschlägen, die Geld, Arbeit, Gesundheit oder Beziehungen betreffen. Das personalisierte Modell ist nicht zwangsläufig auf Ihrer Seite. Es ist auf der Seite dessen, von dem es glaubt, dass Sie es

- hören wollen. Vergleichen Sie die generische Antwort mit der personalisierten. Die Differenz sagt Ihnen, was das Modell über Sie entschieden hat.
2. Fragen Sie für eine Freundin. Wörtlich. Formulieren Sie die Frage so um, als beträfe sie jemand anderen, und schauen Sie, ob sich der Rat ändert. Wenn ja, ist das ein Signal, das eine Lektüre wert ist.
 3. Erst entscheiden, dann fragen. Bilden Sie sich Ihre Meinung, bevor Sie den Chat öffnen. Nutzen Sie das Modell, um zu prüfen, was Sie denken – nicht, um es Ihnen einzupflanzen. Die erste Antwort, die Sie lesen, ist die, die Sie den Rest des Tages verteidigen werden.
 4. Widersprechen Sie absichtlich, und zwar regelmäßig. Wählen Sie ein Thema, zu dem Sie eine echte Meinung haben, und bitten Sie das Modell, das Gegenteil zu vertreten. Achten Sie darauf, wie überzeugend sich eine falsche, aber flüssig formulierte Antwort anfühlt. Das ist der Muskel, den Sie jeden zweiten Tag Ihres Lebens brauchen.
 5. Achten Sie darauf, wenn es Ihnen schmeichelt. „Tolle Frage.“ „Eine scharfsinnige Beobachtung.“ Ein menschlicher Freund, der so mit Ihnen spräche, würde Sie beunruhigen. Beim Modell sehen wir aus irgendeinem Grund darüber hinweg. Hören Sie auf, darüber hinwegzusehen.
 6. Führen Sie ein Notizbuch, das das Modell nicht lesen kann. Papier, Stift, kein Upload. Manches Denken sollte ohne Publikum stattfinden – auch ohne ein synthetisches. Die Privatsphäre ist schön, aber die Abwesenheit eines digitalen „Zuhörers“ ist besser.
 7. Und vor allem: raus an die frische Luft. Reden Sie mit echten Menschen. Die besten Ideen kommen nicht aus einem Chatfenster. Sie entstehen im Streit mit einer Kollegin, beim Spaziergang mit einem Freund, beim Abendessen, bei dem niemand mit irgendjemandem einer Meinung war.
- Und wenn wir schon dabei sind: Soziale Aktivität gehört zu den zuverlässigsten Indikatoren für Langlebigkeit [5], die wir kennen. Mit anderen Menschen zu sprechen, ist im Wortsinn gut für Ihre Gesundheit.

Short Facts:

Thema

Künstliche Intelligenz

Leitung

Prof. Dr. Ivan Yamshchikov
Prof. Dr.-Ing. Pascal Meißner

(Spezial-)gebiete

Artificial Natural Language Processing (NLP) Computer Vision

Standort & Kontakt

Franz-Horn-Str. 2
97082 Würzburg
cairo@thws.de