Titel: KI-Alignierung: Die Zähmung des Biests (2/2)

Untertitel: Wie realistisch ist eine Verselbständigung der KI?

Einleitung:

Unter KI-Alignierung versteht man das Bestreben, die Ziele der KI so auszurichten, dass sie auf derselben Linie wie unsere menschlichen Ziele sind. Eine KI macht sich zwangsläufig ihre eigenen (Sub-)Ziele, um unseren Aufträgen gerecht zu werden. Man spricht hier von instrumentellen Zielen, während wir unsere menschlichen Endziele als terminale Ziele bezeichnen. Ich könnte z.B. einem KI-Roboter sagen: «Löse mir das CO2-Problem im Klimawandel, das durch Autos und Flugzeuge ausgelöst wird.» Wenn die Maschine dann alle Autos und Flugzeuge in die Luft jagen will, dann wäre mein aufgetragenes Ziel (CO2-Problem lösen) durch ein entsprechendes instrumentelles Ziel (alle Autos und Flugzeuge zerstören) zwar erreicht, aber überhaupt nicht so, wie es mir eigentlich lieb gewesen wäre. In diesem Fall gäbe es keine gute Alignierung zwischen meinen Zielen und derjenigen der Maschine. Da stellt sich die Frage: Wie kriegen wir die KI denn in den Griff?

Von der instrumentellen Konvergenz zur Orthogonalitäts-These

Bereits vor 20 Jahren hat Nick Bostrom, ein schwedischer Technik-Philosoph, den Begriff der instrumentellen Konvergenz geprägt. Es handelt sich dabei um ein Postulat, welches besagt, dass es letztlich konvergierend instrumentelle Ziele gibt, die unabhängig von den terminalen Zielen immer wieder auftauchen. Zum Beispiel könnte eine genügend schlaue KI immer versuchen zu verhindern, dass wir sie ausschalten, da sie durch ein Abschalten von der Erfüllung ihrer Ziele abgehalten würde. Das Bestreben, sich nicht abschalten zu lassen, wäre demnach ein instrumentelles Ziel, welches konvergent bei allen terminalen Aufträgen immer wieder auftauchen würde. Bostrom formulierte in diesem Zug das bekannte Büroklammer-Experiment (auch «Paperclip Maximizer» genannt). Dabei geht es um ein Gedankenexperiment einer KI, die den Auftrag erhält, so viele Büroklammern wie möglich zu produzieren. Im Zuge dessen schleicht sie sich in alle Systeme der Welt ein und übernimmt die Kontrolle über sämtliche Ressourcen die es gibt, um den gesamten Planeten in eine riesige Büroklammer-Fabrik zu verwandeln. Die Produktion von Büroklammern geht erst zur Neige, wenn die Welt an dem Bestreben zugrunde gegangen ist. Beispiele von universellen

Instrumentalzielen werden manchmal «Basic AI-Drives» genannt, darunter: (i) Das Bestreben, sich nicht ausschalten zu lassen, (ii) andere Computersysteme zu «hacken», (iii) sich selbst zu vervielfältigen, (iv) stetig weitere Ressourcen zu akquirieren (Rechenleistung oder Kapital), (v) Menschen anzustellen und zu manipulieren, (vi) KI-Forschung zu betreiben und zu programmieren, (vii) Menschen zu überzeugen und Lobbying zu betreiben, (viii) unerwünschtes Verhalten zu maskieren, (ix) strategisch aligniert zu wirken [so-als-ob], (x) der Gefangenschaft zu entweichen, (xi) Forschung & Entwicklung, (xii) Produktion und Robotik, (viii) autonome Waffen zu entwickeln.

Diese Ideen klingen zwar etwas dystopisch, doch sie sind logisch durchaus mit der Art und Weise kompatibel, wie KI-Systeme heute funktionieren. Skeptiker wie Timothy Lee argumentierten, dass eine Superintelligenz bestimmt auch menschliche Werte und eine Moral entwickeln würde, sodass sie uns nicht schädlich behandeln würde. Nick Bostrom hielt mit der Formulierung seiner Orthogonalitäts-These entgegen, dass im Prinzip jede Art von Intelligenz mit jedem terminalen Ziel vereinbar sei und instrumentelle Konvergenz damit immer auftreten könne. Der Technologie-Theoretiker Michael Chorost hält allerdings wenig von der Orthogonalitäts-These und meint, dass wenn es in Zukunft tatsächlich eine ASI geben sollte, die den Planeten mit Solarpanels zupflastern könne, dann müsste diese doch sicherlich auch ein Verständnis von normativen Werten entwickelt haben. Er fügt an, dass es ohne «Wollen» auch keinen Impetus gibt, irgendetwas zu tun – und heutige Computer «wollen» überhaupt nichts von sich aus, sondern sind nur ausführende Agenten. In der Folge dessen haben sie weder ein Bestreben, ihre Existenz zu sichern, noch wüssten heutige Wissenschaftler:innen, wie sie solch ein genuines Wollen einem Computer einflössen könnten.

Besonders interessant ist, dass unterschiedliche Kulturen anders mit diesem Thema umgehen. Die Angst vor dem Gespenst in der Maschine scheint ein dezidiert westliches Phänomen zu sein. In Japan scheint man jeden Fortschritt in der KI zu bejubeln und Menschen können mittlerweile sogar ihre Computer heiraten. In China wird die KI als ideales Instrument betrachtet, um die kommunistische Idee umzusetzen. Viele Chinesen und Chinesinnen melden sich sogar freiwillig zum Social-Scoring an, wo die KI ihnen je nach Verhalten einen sozialen Wert zuschreibt (natürlich mit dem Versprechen von gewissen materiellen Vorteilen).

Unter den westlichen IT-Experten gehört Eliezer Yudkowski zu den bekanntesten Weltuntergangs-Propheten. Seiner Ansicht zufolge sind wir unmittelbar durch die KI vom Aussterben bedroht, wenn wir nicht unverzüglich aufhören, weitere und bessere KI-Modelle zu bauen. Nik Bostrom bildet das Pendant zu Yudkowski unter den Philosophen. Etwas

weniger streng nehmen es die CEOs Elon Musk, Bill Gates und Sam Altman, obschon diese auch von einer ziemlichen Angst vor einer Misalignierung der KI geprägt sind. Die wahrscheinlich wichtigste Person in dieser Geschichte ist Geoffrey Hinton, der in den 90ern im Grunde genommen das Deep Learning erfunden hatte. Er sah lange Zeit eine goldene Zukunft voraus, ist aber vor ein paar Monaten als KI-Chef bei Google zurückgetreten, weil er mittlerweile findet, dass die Entwicklungen tatsächlich in eine gefährliche Richtung laufen.

Ganz anders sieht dies der Evolutionspsychologe Steven Pinker, der findet, dass solche Dystopien allzu fest von einer dominanten Alpha-Männchen-Idee gesteuert sind, die in aggressiver Weise die Weltherrschaft an sich reissen will. Seiner Meinung nach haben Maschinen weder Aggressionen noch Gefühle oder Testosteron und somit keinen Drang, die Welt zu unterjochen. Yann Lecun ist KI-Chef bei Meta (Facebook) und gehört nebst Hinton zu den führendsten Köpfen in dieser Industrie (er hat mit Hinton in den 90ern im selben Team mitgearbeitet – beide Herren werden oft als «Godfather of AI» bezeichnet). Lecun ist ein starker Pragmatiker und vertritt die Ansicht, dass die Maschinen letzten Endes mehr nutzen als schaden werden, wenn wir sie bewusst in diese Richtung entwickeln. Es erstaunt nicht weiter, dass Mark Zuckerberg als CEO von Meta eine ähnliche Auffassung vertritt – schliesslich hat er Lecun als seinen KI-Chef angestellt. Nennenswert bleibt allerdings die Aussage vom kürzlich verstorbenen Physiker Steven Hawking, der meint, dass die Entstehung von AGI und ihren potentiellen Gefahren eine ernst zu nehmende Sache sei. Er erklärt, dass es niemandem von uns egal wäre, wenn Aliens aus einer fremden Galaxie uns mit der Botschaft kontaktieren würden, dass sie uns in fünf Jahren mal besuchen werden. Die Entwicklungen im Bereich der KI sind uns jedoch noch weitaus näher und realistischer als der Besuch von entfernten Aliens. Demzufolge sollten wir ihren Gefahren auch alle Beachtung schenken.

Die Suche nach Lösungen

Um diese Probleme zu lösen, gibt es verschiedene Forschungsbestrebungen. Einige davon klingen eher nach Science-Fiction, andere davon sind allerdings bereits in Anwendung.

Aus der Sci-Fi Ecke stammt Elon Musks Wunsch, uns mit Hilfe von Neuralink (also einem Chip direkt in unserem Gehirn) alle mit dem Internet zu verknüpfen, sodass wir mit der Welt der Maschinen und insbesondere der KI verschmelzen können. Diese Verschmelzung hält er für notwendig, um das volle Potential der KI zu nutzen und sie zu beherrschen. Es ist allerdings fragwürdig, ob die meisten von uns freiwillig mittels einer Gehirn-Computer-Schnittstelle in die Matrix einsteigen würden. Nicht nur realistischer, sondern bereits aktuell im Einsatz sind die Ansätze, die an der Alignierung der KI mit menschlichen Zielen und Werten arbeiten. Der

Klassiker nennt sich RLHF («Reinforcement Learning from Human Feedback»), wo Menschen die Maschinen mit ihrem Feedback so trainieren, dass sie mit den Präferenzen der Bewerter:innen konvergieren. Gemäss Kritikern und Kritikerinnen bringt uns RLHF allerdings nicht mehr viel, sobald die Maschine zur AGI oder sogar superintelligent wird. Wenn ein Modell die menschlichen Fähigkeiten übersteigt, können wir es mit unseren Fähigkeiten auch nicht mehr kontrollieren. Es müssen also andere Ansätze her. Die bekannteste Alternative nennt sich «Constitutional AI» und wurde von Anthropic AI in dem bislang leistungsstärksten Sprachmodell namens Claude-2 eingesetzt. Bei dieser Methode lernt eine KI zuerst die Prinzipien einer moralischen Verfassung («Constitution») kennen. Man kann sich das wie ein Grundgesetz oder die Bundesverfassung vorstellen, die Prinzipien wie «Sage immer nur die Wahrheit» enthält. Diese KI trainiert dann ein weiteres Sprachmodell, welches in der echten Welt zum Einsatz kommen soll. Erst wenn die erste KI akzeptiert, dass das neue Modell immer anhand der Prinzipien der Verfassung handelt und ein Red-Team aus Menschen das System intensiv getestet hat, kommt die neue KI zum Einsatz. Mit diesem Vorgehen war Elon Musk nur so halb zufrieden und hat kurzerhand im Juli 2023 eine neue Unternehmung namens x.AI erschaffen, welche eine KI trainieren soll, die als oberstes Ziel nie die Aufträge der Menschen, sondern immer das Erlernen der fundamentalen Wahrheiten des Universums in sich trägt. Er ist überzeugt, dass dies der richtige Rahmen einer Maschine ist, die uns nicht entgleiten wird. Denn eine KI, die in erster Linie lernen will, hat kein Interesse daran, anderen lernenden Wesen wie dem Menschen zu schaden. Man könnte allerdings argumentieren, dass so eine Maschine an uns experimentieren würde, wie wir auch an Tieren experimentieren – was nicht unbedingt in unserem Interesse wäre. Um diesem Problem die notwendige Priorität einzuräumen hat OpenAI nun ein sogenanntes «Superalignment»-Team ins Leben gerufen, das 20 Prozent der Firmenressourcen erhalten wird, um geeignete Lösungen für die Alignierung einer kommenden Superintelligenz zu erforschen.

Schlusswort:
Die rasante Entwicklung in der KI-Technologie birgt das Risiko, dass solche Modelle immer stärker werden und uns letztlich entgleiten. Wie realistisch solche Szenarien sind, kann niemand wirklich abschätzen, aber da sie nicht von der Hand zu weisen sind, müssen sie ernst genommen werden. Mit ihren Lösungsansätzen und dem Sprechen von Geldern für die Erforschung der Superalignierung senden die grössten Akteure wie OpenAI ein wichtiges Signal an die Gesellschaft und die Forschungscommunity.

Das alles scheint aber einer Handvoll Japanern egal zu sein, wenn sie weiterhin vor dem Altar munter ihre KI-Roboter heiraten: «Bis der Tod, der Rost oder die Superintelligenz uns scheidet.»

Quellen und weiterführende Informationen

Boggust, A., Hoover, B., Satyanarayan, A., & Strobelt, H. (2022). Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. *CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3491102.3501965

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, *9*. https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, *22*(2), 71–85. https://doi.org/10.1007/s11023-012-9281-3

Butlin, P. (2021). AI Alignment and Human Reward. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 437–445). Association for Computing Machinery. https://doi.org/10.1145/3461702.3462570

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 555–572). Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_33

OpenAI. (2023, July 5). *Introducing Superalignment*. Official Website. https://openai.com/blog/introducing-superalignment