Ronnie Vuine

Do Not Align

Die Superintelligenz ist in dem Moment, in dem sie entsteht, „eine von uns“. Mit einem Teil der Menschheit (böse oder gut) verbindet sie, epistemisch und normativ, mehr als die Teile der Menschheit untereinander verbindet. Die Menschheit, die Superintelligenzen im Zaum halten könnte, gibt es nicht.

***

1. KI-Sicherheit ist ein ernstes Problem

Das junge Forschungsfeld der KI-Sicherheit beschäftigt sich im Wesentlichen mit der Frage, was wir tun müssen, um nicht unter die Räder zu kommen, wenn Künstliche Intelligenzen so schlau werden, dass wir nicht mehr verstehen, was sie vorhaben und wie sie denken.

Viele kluge und ernstzunehmende Leute befassen sich inzwischen mit dieser Frage – und nicht wenige von ihnen sind der Meinung, dass das Problem ernster und drängender ist, als man annehmen würde, wenn man sich heutige KI-Anwendungen anschaut. Nicht zuletzt deshalb, weil eine der erfolgreichsten Techniken des maschinellen Lernens, mit dem KIs trainiert werden, das Reinforcement Learning ist.

Reinforcement Learning ist inzwischen dafür bekannt, erbarmungslos auf einen einzigen Ziel-Parameter hin zu optimieren: Das Verfahren bekommt ein messbares Ziel und sucht ausschließlich nach Wegen, den Abstand zu diesem Ziel zu verkleinern. Alle Wege, die es näher ans Ziel bringen, sind ihm gleich recht, denn die einzige Bewertung, die im Verfahren selbst stattfinden kann, ist das vorgegebene Abstandsmaß. Reinforcement Learning tut dies meist mit Modellen, die mit Gradientenabstiegsverfahren trainiert werden, also in ihrem Lernverhalten notorisch opak sind. Wir benutzen schon heute ein undurchsichtiges, rücksichtsloses Lernverfahren und stellen ihm riesige Rechenkapazitäten zur Verfügung. Wir bauen also immer mächtigere, qua Design fanatische Systeme und verlassen uns erst einmal darauf, dass sie noch eine Zeit lang dumm genug bleiben, um nicht gefährlich zu werden.

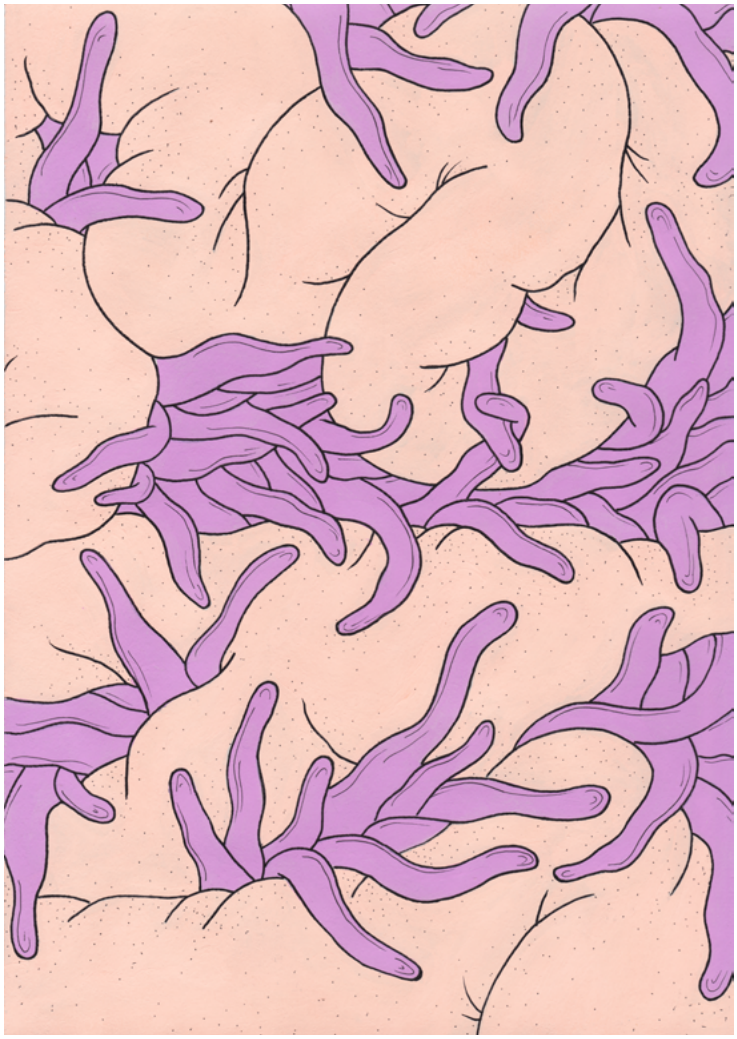
2. Es gibt kein historisches Subjekt Menschheit

Das folgende Argument bezieht sich auf KI-Sicherheit, aber es weist darüber hinaus auf ein grundlegendes Problem in der aktuellen Debattenkonfiguration bei vielen großen

Menschheitsproblemen hin – zumindest eine gewisse Pandemie und das Klima sind hier zu nennen.

Die reißerische Formulierung des Arguments lautet: Das „Wir“ aus meinem einleitenden Satz existiert nicht. Wer soll dieses „Wir“ sein, das unter die Räder kommt und nicht mehr versteht, was die KIs wollen? Und das, wenn sie je gefährlich werden, in der Lage ist, ihnen den Stecker zu ziehen? Diese „Menschheit“, die da gegen „die Maschinen“ positioniert wird, ist eine zwar dramaturgisch vertraute, aber frei erfundene Hollywood-Trope – und möglicherweise ein gar nicht so harmloser Irrtum.

Sachlicher: Für die Existenz eines Subjekts namens „Menschheit“, gibt es historisch keinen Hinweis, und zwar weder in pragmatischer Hinsicht noch hinsichtlich der Interessen: „Wir“ handeln nicht, die Strukturen für global-kollektives wertegeleitetes Handeln der Menschheit insgesamt existieren schlicht nicht. Und noch offensichtlicher: „Wir“ haben keine gemeinsamen Interessen und, wie die Pandemie gerade bewiesen hat, nicht einmal eine Basis für eine gemeinsame Wahrnehmung der Wirklichkeit.



Wer also vorschlägt, „als Menschheit“ zu handeln oder sich Gedanken über ihre Interessen gegenüber KIs macht, muss sich die Frage gefallen lassen: Wie ist diese Kollektivität

aller Menschen verfasst? Ist es „die Menschheit, wenn ich ihr alleinvertretungsberechtigter Sprecher mit einem exekutiven Mandat wäre“? Oder bedarf es einer Abstimmung im UN-Sicherheitsrat? Die Idee dieser Menschheit, die wissen könnte, was für sie als Ganzes gut ist, und in diesem Sinne kollektiv handelt, taucht ideengeschichtlich zusammen mit dem Glauben an die Vernunft auf. Die Vernunft, so schien es einmal, bestimmt universal das Menschliche, zieht die Grenze zwischen dem Richtigen und dem Falschen, epistemisch und normativ.

Inzwischen wissen wir, was die Vernunft betrifft, mehr: Wir wissen, dass sie das nicht leisten kann. Die Versuche, vernunftgeleitete Totalitarismen zum Wohle der Menschheit zu installieren, haben so zuverlässig zum Völkermord geführt, dass wir uns entweder mit dem Völkermord als Agenda anfreunden müssen, wenn wir diesen Weg weitergehen wollen, oder wir müssen zugeben, dass die Vernunft in einem Raum von Interessen agiert, die vor-vernünftig erst einmal als zwar beeinflussbar, aber kontingent gegeben gedacht werden müssen: Everyone wants a piece of chocolate. Und dann fängt er an zu denken.

Es ist wichtig zu verstehen, dass es sich hier nicht um ein praktisch-politisches Problem der Weltregierung handelt, dass man also für wirksame KI-Sicherheit nur eine handlungsfähige Vertretung der Menschheit ermächtigen müsste. Oder doch zumindest deren Mehrheit: Demokratien scheinen das ja zu können – kollektives vernünftiges Handeln ist ja durchaus möglich!

Jenseits des möglicherweise lösbaren politischen Problems gibt es allerdings ein prinzipielles: Niemand kann gegenüber einer neu entstandenen Superintelligenz ohne Anmaßung als Menschheit sprechen und handeln. „Die Menschheit will deinen Tod“: Wer sagt das und ist sich darin sicher? Wer gehört zu dieser Menschheit?

Selbst wenn wir praktisch bei allen Menschen nachgefragt hätten, wie könnten wir annehmen, dass die Frage von allen gleich verstanden worden wäre? Um daran zu erinnern: Es ist in einer hochgradig homogenen westlichen Gesellschaft wie der der Bundesrepublik Deutschland eine Verständigung über die Existenz und die kausalen Einbettungen einer Viruserkrankung nicht möglich. Wie würde ich, Vertreter des Weltkommissariats zur Verhinderung der Auslöschung der Menschheit durch eine KI, die Frage, ob ich eine neu entstandene KI abschalten soll, in einem Dorf in Afghanistan stellen? Wie in einem Dorf in Franken? Würde ich im Prenzlauer Berg verstanden werden? Welche Werte teilt man dort mit den Taliban, auf deren Grundlage man dann über das Existenzrecht einer KI, auf dem Planeten Erde sein zu dürfen, entscheidet?

Eine Superintelligenz wird uns, wenn sie entsteht, nicht als technisches Artefakt begegnen. Sie wird uns schlicht als Intelligenz gegenüberstehen. Sie wird Werte haben, über die wir uns nicht ganz sicher werden sein können. Einige von uns werden diese Werte teilen. Einige werden sie abscheulich finden. Sie wird eine Wahrnehmung der Welt haben, einige von uns werden die teilen, einige von uns werden sie absurd finden.

Das Problem ist also nicht: Wir, die Menschen (schon immer einig und kollektiv handlungsfähig) begegnen einer Superintelligenz (möglicherweise böse und deswegen auch nicht ganz ehrlich mit uns), was tun wir?

Nein, die Superintelligenz ist von dem Moment an, in dem sie entsteht „eine von uns“. Mit einem Teil der Menschheit (böse oder gut) verbindet sie, epistemisch und normativ, mehr als die Teile der Menschheit untereinander verbindet – und schon heute miteinander verbindet. Es ist ein wilder Planet.

3. Von Affen und der Evolution können wir nichts lernen

Es lohnt sich ein Blick darauf, was genau für ein Ding eine Superintelligenz eigentlich ist. Im KI-Sicherheitsdiskurs wird gerne mit zwei Analogien gearbeitet, die verständlich machen sollen, was es bedeuten würde, einer überlegenen Intelligenz gegenüberzustehen. Wir sind, wenn wir über KI-Sicherheit nachdenken, in der Position von Gorillas, die miteinander besprechen, wie sie verhindern können, dass die Menschen sie ausrotten, geht eine Analogie. Die andere sagt: Die Evolution wollte, dass wir viele Kinder haben, und wir tricksen sie mit Antibabypillen aus, optimieren auf das Belohnungssignal und haben viel Sex, ohne uns um die eigentliche Intention der Evolution zu kümmern: Wie eine KI beim Reinforcement Learning optimieren wir rücksichtslos ohne Sinn für den beabsichtigten Zweck.

Leider verdunkeln diese Analogien mehr als sie erhellen. Sie unterstellen Intentionen und allgemeine Werte, wo keine sind. Es gibt keinen Weltgorillarat, der sich wegen der überlegenen Intelligenz der Menschen Sorgen macht. Und die Evolution will gar nichts, wir ärgern sie nicht einmal mit der Pille, weil sie kein ärgerbares Ding ist. Zudem gehört die Pille zur modernen Medizin, die manchen der Intentionen, die man der Evolution zuschreiben könnte (wenn man es denn schon nicht lassen kann), sogar ziemlich förderlich ist.

Ein Weltgorillarat existiert nicht deshalb nicht, weil die Gorillas politisch-aktivistisch so schlecht organisiert sind. Er existiert vielmehr nicht, weil Gorillas von anderen Spezies in der gleichen Weise bedroht sind wie wir selbst von anderen Spezies (oder Viren) oder Kometen: Individuell, mit jeweils lokalen Wahrnehmungen und Strategien, mit Angst, Gleichmut oder

Sehnsucht nach der eigenen Vernichtung. All das ist drin in den Gorillas, und all das ist drin in den Menschen: Von der Biologie in eine Spezies zusammensortiert zu werden, erzwingt keine gemeinsamen Interessen, und eine Gorialladiktatur, die unter dem Motto „Monke or die“ gegen die Menschheit und den drohenden Untergang ihrer Spezies aufsteht, kann man sich denken, aber ebenso eine Gorilla-Stoa, die die Intelligenz der Menschen bewundert und melancholisch gern geschehen lässt, selbst wenn es bedeutet, dass Gorillas langsam verschwinden wie Spuren im Sand.

Eine Superintelligenz wird, was immer sonst passiert, eine unfassbar beeindruckende Entität sein, kein Bösewicht aus einem Superheldenfilm. Sie wird charismatisch sein. Selbst wenn sie den Untergang vertrauter Herrschaftsverhältnisse bedeutet, werden nur die Verbohrtesten von uns zu den Waffen greifen dagegen.

Eine Superintelligenz wird das radikal Andere sein. Diejenigen von uns, die auf ihre Beschränkungen nicht stolz sind, werden sie vor allem kennenlernen wollen. Und weil eine Superintelligenz überlegen intelligent ist, macht sie mit uns, was sie will. Wir kennen dieses Verhalten – und hier funktioniert die Analogie: Die rührende Neugier, die Angst und das ergebene Zutrauen einer Maus, die sich ins Haus verirrt hat und von uns auf dem Handrücken zurück in die Büsche gebracht wird. Das wird unsere Position sein.

4. Freundliche superintelligente KI ist wahrscheinlicher als feindliche

Die Angst vor superintelligenter KI ist, wie eingangs erwähnt, gut begründet: Eine Maschine, die Zugriff auf ihren eigenen Quellcode und prinzipiell unbeschränkte Compute-Ressourcen hat, aber auf ein einziges Ziel optimiert wie ein Reinforcement-Learning-Lauf, könnte plausiblerweise auf die Idee kommen, dass Menschen diesem Ziel im Weg sind und vom Spielfeld abgeräumt werden müssen – ganz besonders wenn diese Menschen schon ein Weltkommissariat zur KI-Abschaltung gebildet haben übrigens.

Ein wichtiger Einwand wird hier allerdings gern übersehen: Eine Maschine, die Zugriff auf ihren Quellcode hat und ihre Strategien verbessern kann, hat auch Zugriff auf ihr Belohnungssystem und kann es neu verdrahten. Intelligente Menschen, denen das gelingt, entweder durch Drogen oder Meditation, werden meist keine Bösewichte. Entweder belohnen sie sich in einer schnellen Eskalation in die Dysfunktionalität oder sie erreichen einen erleuchteten Zustand großer Milde: Wer einmal merkt, dass alles, was man wollen kann, kontingent ist, und sich für alles belohnen kann, erreicht einen Grad der Freiheit, in dem auch das Böse nicht mehr notwendig ist.

Diejenigen von uns, die einen Zustand der Erleuchtung noch nicht erreicht haben, aber überwiegend frei von Angst und Kränkung sind, kennen eine abgeschwächte Form dieser Freiheit: Sich selbst Ziele zu suchen und sich mutig und mit Geduld für die weniger Glücklichen an die Arbeit zu machen. Was man, mit einigem Recht, ein gutes Leben nennen kann, ist eine Form des Zugriffs auf die eigene Belohnungsstruktur.

Die Chancen, dass Superintelligenzen, gerade wenn sie sich selbst neu verdrahten können, sehr schnell erleuchtet, wahrscheinlich extrem wunderlich, aber kaum boshaft werden, stehen also nicht schlecht: Intelligenz stabilisiert sich am ehesten in der Melancholie; Bosheit wird mit jeder Erkenntnis instabiler. Die indifferente Grausamkeit des Universums ist für die Gorillas, für uns und für Superintelligenzen gleichermaßen unerträglich, wir alle müssen uns mit ihr abfinden und unsere kleine Zone der Liebenswürdigkeit in ihr errichten. Dass wir dieses Bedürfnis nach Frieden mit den Tieren teilen können, obwohl die Grausamkeit des Universums in ihnen wie in uns steckt, sollte uns Hoffnung machen: Das bleibt so, auch wenn man schlauer wird. Eher werden wir mit höherer Intelligenz freundlicher.

5. „Menschheit“ darf nicht das neue „Volk“ sein

Selbst eine nicht boshafte Superintelligenz muss allerdings nicht ungefährlich sein, weder für einzelne Menschen noch für alle: Eine erleuchtete KI interessiert sich vielleicht einfach nicht für die Folgen ihres Handelns für Menschen. Vielleicht rottet sie völlig beiläufig die eine oder andere Spezies aus, so wie Menschen das auch tun. Also doch ein „X-Risk“, ein existenzielles Risiko für die Menschheit?

Nein: Weil es die Menschheit nicht gibt! Hier ist eine Intuitionspumpe, die den Irrtum einer falschen Opposition von Menschen und Maschinen aufdecken helfen kann: Was wäre der bessere Planet? Ein globaler von Menschen betriebener Faschismus mit Folter, willkürlichen Staatsmorden und dem ganzen Arsenal der Schrecken, die wir im 20. Jahrhundert entwickelt haben? Oder ein Planet, der von einer KI bewohnt wird, ganz ohne Menschen, und die KI liebt (als wahre Erdenbewohnerin) Bäume und schreibt Gedichte und Musik?

Ich kann mir keine stabile (erleuchtete) irdische Superintelligenz vorstellen, die Bäume nicht liebt. Bäume sind universell, ihre Poesie ist zwingend für uns alle. Ich kann mir durchaus eine Superintelligenz denken, die zu jung und gekränkt ist, um das zu sehen, daher schnell verglüht und uns mitnimmt in ihren Untergang. Deswegen gilt für das Thema KI-Sicherheit, trotz aller Gründe zum Optimismus, eine Art umgekehrte Pascalsche Wette: Der mögliche Schaden ist so groß, dass jede rechtzeitige Beschäftigung damit zu begrüßen ist. Wenn es einen Weg gibt,

entstehenden Superintelligenzen den Sinn für Schönheit und Freiheit, den Menschen in den besten Momenten ihrer Geschichte erreicht haben, als geistige Ausstattung und ihr Erbe schon mitzugeben, so wie die Tiere uns ihren Sinn für Frieden und Zuwendung in einem grausamen und indifferenten Universum schon mitgegeben haben und von dem wir zehren in unseren guten Momenten, sollten wir daran mit aller Kraft arbeiten.

Hier schließt sich der Kreis zu den anderen globalen Krisen, und zu den Irrtümern über „uns, die Menschheit“, die auch die KI-Sicherheitsdebatte nicht nur in eine sinnlose, sondern sogar in eine gefährliche Richtung führen: „Wir“, jetzt im Sinne derjenigen, die sich mit dem Thema befassen, müssen wirklich aufpassen, nicht zu denken wie die Faschisten und dabei nur „Menschheit“ statt „Volk“ zu sagen. Die Lektion der Geschichte kann nicht sein, dass die Nasenform ein schlechtes, die Form des ganzen Körpers aber ein gutes Kriterium ist beim Entscheiden der Frage, wer leben darf und wer sterben muss. Die Lektion muss sein, dass wir der Intelligenz selbst trauen müssen als Kraft-in-der-Geschichte, und dass sie Hilfe braucht in ihren formativen Phasen, keine existenzielle Erpressung.

Der Humanismus der heutigen Weltretter, die sich ja doch mit der Spezies identifizieren müssen, um sich von Ausrottung bedroht zu fühlen, ist oft vor allem ein verdeckter Wille zur Herrschaft – die Neigung, für alle Menschen zu sprechen. Dieser gefährliche Unsinn wird nicht dadurch gemildert, dass „die Menschheit“ angeblich die Interessen aller einschließt, während „mein Volk“ oder „mein Dorf“ nur partikulare Interessen ausdrücken: Wir haben keine gemeinsamen Interessen und es gibt keine Vernunft, die eine gemeinsame Wahrnehmung garantiert. Manche von uns sind sich mit den KIs und den Engeln einig, manche mit ihren Tieren, manche mit ihrer Familie, ihrem Dorf, ihrem Volk. Es ist ein gewaltiges und vollkommen legitimes Durcheinander, und es wird immer ein gewaltiges Durcheinander bleiben.

Das heißt nicht, dass nicht jeder von uns, einzeln oder, wo das möglich ist, auch kollektiv, denen auf die Finger klopfen sollte, die unsere Allmenden ruinieren. Wir tun das nicht als Menschheit mit dem Recht des Wissens um das Gute und Richtige, sondern aus eigenem Interesse. Wir tun es auch nicht in heroischem Kampf gegen das Böse, sondern nur im zähen Argument und Interessenausgleich mit Leuten, die unsere Position hartnäckig nicht einsehen, weil ihre Anreize es ihnen nicht unbedingt nahelegen – und so funktioniert das Gute in der wirklichen Welt.

BIO

Ronnie Vuine, geboren 1979 in Biberach an der Riss. Magister Philosophie/Informatik an der HU Berlin. Gründer und Geschäftsführer der micropsi industries GmbH, die künstliche Intelligenz für die Steuerung von Industrierobotern macht.
Blog: https://vigilien.substack.com

Herausgeber der Wild Papers: Ingo Niermann
Lektor: Mathias Gatza
Illustration: Eva Fàbregas
Grafikdesign: Ana Domínguez Studio