STIFTUNG FÜR EFFEKTIVEN ALTRUISMUS

Künstliche Intelligenz: Chancen und Risiken

Diskussionspapier der Stiftung für Effektiven Altruismus

Medienkonferenz

Donnerstag, 12. November 2015, 14:00

Technopark, Raum Newton 1010, Technoparkstrasse 1, 8005 Zürich

www.ea-stiftung.org/kuenstliche-intelligenz

Die Stiftung für Effektiven Altruismus

Die **Stiftung für Effektiven Altruismus (EAS)** ist eine unabhängige Denkfabrik und Projektschmiede für Effektiven Altruismus, gegründet von einem jungen, interdisziplinären Team. Sie ist bestrebt, ethische Fragen unserer Zeit wissenschaftlich fundiert anzugehen. Die Resultate macht sie unter anderem im Rahmen von Diskussionspapieren der Politik und Öffentlichkeit zugänglich. Der Hauptfokus der Stiftung richtet sich dabei auf die evidenzbasierte, kosteneffektive Armutsbekämpfung; auf die Reduktion des Tierleids; sowie auf die Chancen und Risiken von Zukunftstechnologien.

Kontakt

Adresse
Stiftung für Effektiven Altruismus
Efringerstrasse 25
CH-4057 Basel
www.ea-stiftung.org

Medienkontakt und Hauptautor KI-Diskussionspapier
Adriano Mannino
+41 78 858 22 70
adriano.mannino@ea-stiftung.org

Liste der Unterstützer/innen

Die Kerninhalte des Diskussionspapiers werden getragen von:

Prof. Dr. Fred Hamker, Professor für Künstliche Intelligenz, Technische Universität Chemnitz

Prof. Dr. Dirk Helbing, Professor für Computational Social Science, ETH Zürich

Prof. Dr. Malte Helmert, Professor für Künstliche Intelligenz, Universität Basel

Prof. Dr. Manfred Hild, Professor für Digitale Systeme, Beuth Hochschule für Technik (Berlin)

Prof. Dr. Dr. Eric Hilgendorf, Leiter Forschungsstelle *RobotRecht*, Universität Würzburg

Prof. Dr. Marius Kloft, Professor für Maschinelles Lernen, Humboldt Universität Berlin

Prof. Dr. Dr. Franz Josef Radermacher, Professor für Datenbanken und Künstliche Intelligenz, Universität Ulm

Einladung zur Medienkonferenz: Künstliche Intelligenz

BASEL/ZÜRICH, 5. November 2015. **Die Übernahme des KI-Unternehmens DeepMind durch Google für rund eine halbe Milliarde US-Dollar signalisierte vor einem Jahr, dass von der KI-Forschung vielversprechende Ergebnisse erwartet werden. Spätestens seit bekannte Wissenschaftler wie Stephen Hawking und Unternehmer wie Elon Musk oder Bill Gates davor warnen, dass künstliche Intelligenz eine Bedrohung für die Menschheit darstellt, schlägt das KI-Thema hohe Wellen. Die Stiftung für Effektiven Altruismus (vormals GBS Schweiz) präsentiert am 12. November ein umfassendes Diskussionspapier zu den Chancen und Risiken der künstlichen Intelligenz und geht darin auf aktuelle, mittel- und langfristige Herausforderungen im Bereich der KI-Entwicklung ein.**

Der Fokus liegt dabei u.a. auf den folgenden Fragen:

» Welche Probleme werden durch bereits existierende KI-Technologien aufgeworfen (beispielsweise selbstgesteuerte Fahrzeuge)?

» Wie wird sich die zunehmende Automatisierung/Computerisierung von Arbeitsbereichen gesellschaftlich (Arbeit, Bildung) auswirken und wie können wir negative Auswirkungen abfedern?

» Werden dereinst KIs mit (über)menschlicher Intelligenz entwickelt? Welche Bedeutung hätte dies für die Menschheit?

Zu jedem Punkt werden wir mögliche Massnahmen präsentieren. Die Diskussions- und Handlungsanstösse richten sich nicht nur an die Politik und die Forschung, sondern auch an eine breitere Öffentlichkeit.

Referenten an der Medienkonferenz

» Adriano Mannino, Co-Präsident der Stiftung für Effektiven Altruismus
» Dr. Jonathan Erhardt, Wissenschaftlicher Mitarbeiter, Stiftung für Effektiven Altruismus
» Prof. Thomas Metzinger, Professor für Philosophie, Universität Mainz

Weitere Co-Autoren des Diskussionspapiers

» David Althaus, Wissenschaftlicher Mitarbeiter, Stiftung für Effektiven Altruismus
» Dr. Adrian Hutter, Quantum Computing, Berater, Stiftung für Effektiven Altruismus
» Lukas Gloor, Wissenschaftlicher Mitarbeiter, Stiftung für Effektiven Altruismus

Herausforderungen und Massnahmen

Executive Summary des Diskussionspapiers

Künstliche Intelligenz (KI) und immer komplexer werdende Algorithmen beeinflussen unser Leben und unsere Zivilisation stärker denn je. Die KI-Anwendungsbereiche sind vielfältig und die Möglichkeiten weitreichend: Insbesondere aufgrund von Verbesserungen in der Computerhardware übertreffen gewisse KI-Algorithmen bereits heute die Leistungen menschlicher Experten/innen. Ihr Anwendungsgebiet wird künftig weiter wachsen und die KI-Leistungen werden sich verbessern. Konkret ist zu erwarten, dass sich die entsprechenden Algorithmen in immer stärkerem Ausmaß selbst optimieren – auf übermenschliches Niveau. Dieser technologische Fortschritt stellt uns wahrscheinlich vor historisch beispiellose ethische Herausforderungen. Nicht wenige Experten/innen sind der Meinung, dass von der KI neben globalen Chancen auch globale Risiken ausgehen, welche diejenigen etwa der Nukleartechnologie – die historisch ebenfalls lange unterschätzt wurde – übertreffen werden. Eine wissenschaftliche Risikoanalyse legt zudem nahe, dass hohe potenzielle Schadensausmaße auch dann sehr ernst zu nehmen sind, wenn die Eintretenswahrscheinlichkeiten tief wären.

Aktuell

In engen, gut erprobten Anwendungsbereichen (z.B. bei selbstfahrenden Autos und in Teilbereichen der medizinischen Diagnostik) konnte die Überlegenheit von KIs gegenüber Menschen bereits nachgewiesen werden. Ein vermehrter Einsatz dieser Technologien birgt großes Potenzial (z.B. bedeutend weniger Unfälle im Straßenverkehr und weniger Fehler bei der medizinischen Behandlung von Patienten/innen bzw. Erfindung vieler neuartiger Therapien). In komplexeren Systemen, wo mehrere Algorithmen mit hoher Geschwindigkeit interagieren (z.B. im Finanzmarkt oder bei absehbaren militärischen Anwendungen) besteht ein erhöhtes Risiko, dass die neuen KI-Technologien unerwartet systemisch fehlschlagen oder missbraucht werden. Es droht ein KI-Wettrüsten, das die Sicherheit der Technologieentwicklung ihrem Tempo opfert. In jedem Fall relevant ist die Frage, welche Ziele bzw. ethischen Werte einem KI-Algorithmus einprogrammiert werden sollen und wie technisch garantiert werden kann, dass die Ziele stabil bleiben und nicht manipuliert werden können. Bei selbstfahrenden Autos stellt sich etwa die Frage, wie der Algorithmus entscheiden soll, falls ein Zusammenstoß mit mehreren Fußgängern nur so vermieden werden kann, dass die eine Autoinsassin gefährdet wird – und wie sichergestellt werden kann, dass die Algorithmen der selbstfahrenden Autos nicht systemisch versagen.

- **Maßnahme 1:** Die Förderung eines sachlich-rationalen Diskurses zum KI-Thema ist vonnöten, damit Vorurteile abgebaut werden können und der Fokus auf die wichtigsten und dringendsten Sicherheitsfragen gelegt werden kann.

- **Maßnahme 2:** Die gesetzlichen Rahmenbedingungen sollen den neuen Technologien entsprechend angepasst werden. KI-Hersteller sind zu verpflichten, mehr in die Sicherheit und Verlässlichkeit der Technologien zu investieren und Prinzipien wie Vorhersagbarkeit, Transparenz und Nicht-Manipulierbarkeit zu beachten, damit das Risiko unerwarteter Katastrophenfälle minimiert werden kann.

Mittelfristig

Die Fortschritte in der KI-Forschung ermöglichen es, mehr und mehr menschliche Arbeit von Maschinen erledigen zu lassen. Viele Ökonomen/innen gehen davon aus, dass die zunehmende Automatisierung bereits innerhalb der nächsten 10-20 Jahre zu einer massiven Erhöhung der Arbeitslosigkeit führen könnte. (Sie tun dies durchaus im Bewusstsein, dass sich ähnliche Prognosen in der Vergangenheit nicht bewahrheitet haben, denn die aktuellen Entwicklungen sind von neuartiger Qualität und es wäre unverantwortlich, die Augen vor der Möglichkeit zu verschließen, dass die Prognosen irgendwann zutreffen: Selbst tiefe Wahrscheinlichkeiten auf ein sehr hohes Schadensausmaß sind im Rahmen einer wissenschaftlichen Risikoanalyse zu berücksichtigen und für unser Handeln hochrelevant.) Durch die fortschreitende Automatisierung wird der Lebensstandard im statistischen Durchschnitt steigen. Es ist jedoch nicht garantiert, dass alle Menschen – oder auch nur eine Mehrheit der Menschen – davon profitieren werden.

- **Maßnahme 3:** Können wir gesellschaftlich sinnvoll mit den Folgen der KI-Automatisierung umgehen? Sind die aktuellen Sozialsysteme dafür geeignet? Diese Fragen sind ausführlich zu klären. Gegebenenfalls sind neuartige Maßnahmen zu ergreifen, um die negativen Entwicklungen abzufedern bzw. positiv zu wenden. Modelle eines bedingungslosen Grundeinkommens oder einer negativen Einkommenssteuer etwa sind zur gerechteren Verteilung der Produktivitätsgewinne prüfenswert.

Langfristig

Viele KI-Experten/innen halten es für plausibel, dass noch in diesem Jahrhundert KIs erschaffen werden, deren Intelligenz der menschlichen in allen Bereichen weit überlegen ist. Die Ziele solcher KIs, welche prinzipiell alles Mögliche zum Gegenstand haben können (menschliche, ethisch geprägte Ziele stellen eine winzige Teilmenge aller möglichen Ziele dar), würden die Zukunft unseres Planeten maßgeblich beeinflussen – was für die Menschheit ein existenzielles Risiko darstellen könnte. Unsere Spezies hat deshalb eine dominante Stellung inne, weil sie (aktuell) über die am höchsten entwickelte Intelligenz verfügt. Es ist aber wahrscheinlich, dass bis zum Ende des Jahrhunderts KIs entwickelt werden, deren Intelligenz sich zu der unseren so verhält wie die unsere zu derjenigen etwa der Schimpansen. Zudem ist die Möglichkeit nicht auszuschließen, dass KIs in Zukunft auch phänomenale Zustände entwickeln, d.h. (Selbst-)Bewusstsein und besonders auch subjektive Präferenzen und Leidensfähigkeit, was uns mit neuartigen ethischen Herausforderungen

konfrontiert. Angesichts der unmittelbaren Relevanz der Thematik und dessen, was längerfristig auf dem Spiel steht, sind Überlegungen zur KI-Sicherheit sowohl in der Politik als auch in der Forschung aktuell stark unterrepräsentiert.

- **Maßnahme 4:** Es gilt, institutionell sicherheitsfördernde Maßnahmen auszuarbeiten, beispielsweise die Vergabe von Forschungsgeldern für Projekte, die sich auf die Analyse und Prävention von Risiken der KI-Entwicklungen konzentrieren. Die Politik muss insgesamt mehr Ressourcen für die kritische, wissenschaftlich-ethische Begleitung folgenschwerer Technologieentwicklungen bereitstellen.
- **Maßnahme 5:** Bestrebungen zur internationalen Forschungskollaboration (analog etwa zum CERN in der Teilchenphysik) sind voranzutreiben. Internationale Koordination ist im KI-Bereich besonders deshalb essenziell, weil sie das Risiko eines technologischen Wettrüstens minimiert. Ein Verbot jeder risikobehafteten KI-Forschung wäre nicht praktikabel und würde zu einer schnellen und gefährlichen Verlagerung der Forschung in Länder mit tieferen Sicherheitsstandards führen.
- **Maßnahme 6:** Forschungsprojekte, die selbstoptimierende neuromorphe, d.h. gehirnanaloge KI-Architekturen entwickeln oder testen, die mit hoher Wahrscheinlichkeit über Leidensfähigkeit verfügen werden, sollten unter die Aufsicht von Ethikkommissionen gestellt werden (in Analogie zu den Tierversuchskommissionen).