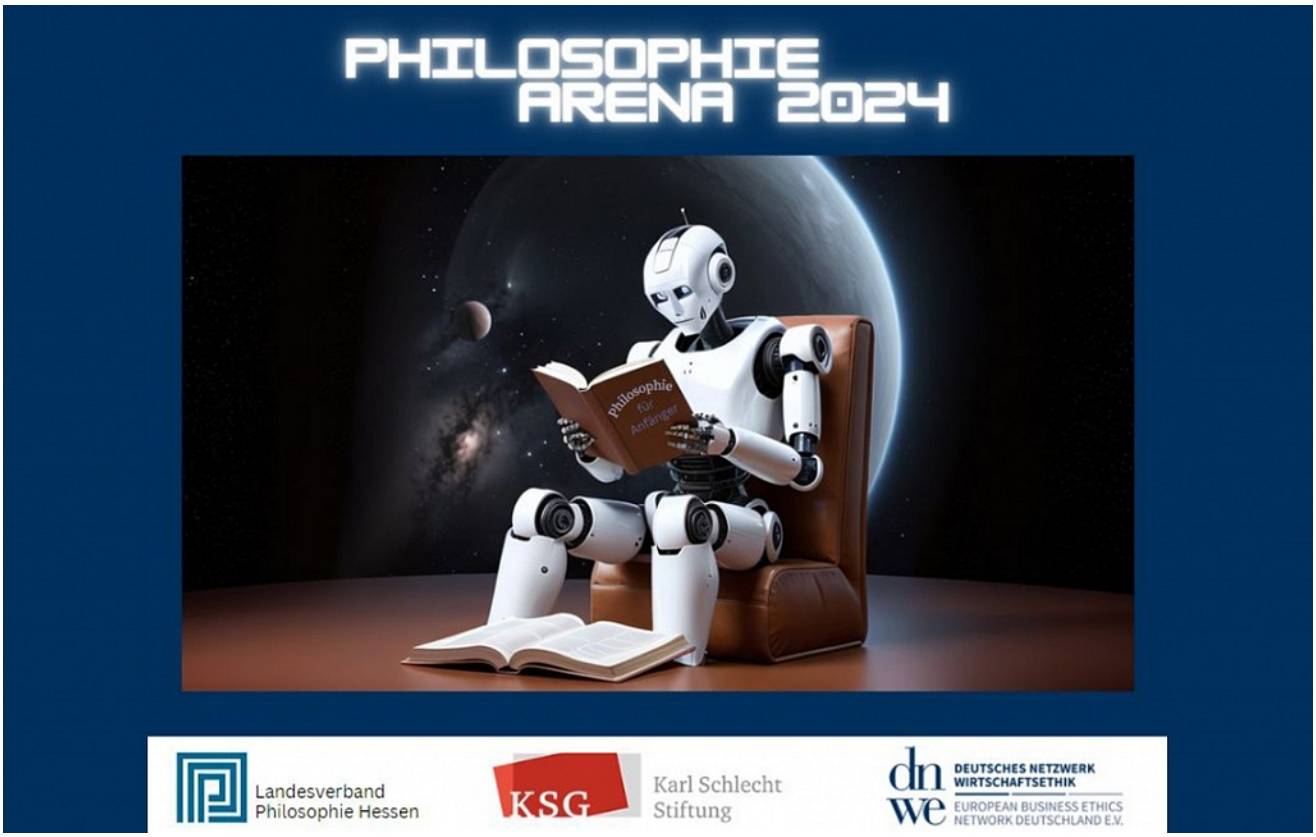
Kann man Menschlichkeit und Moral programmieren?



Schu?lerwettbewerb 2024
PhilosophieArena
Ben Niehuus
Brecht-Schule Hamburg, S3a

Im Jahre 2042 scheint es, als wären die Vereinten Nationen dem Untergang geweiht. In den letzten 23 Vollversammlungen hat es kein Beschluss durch die Abstimmung geschafft, der Sicherheitsrat ist stets von mindestens drei ständigen Mitgliedern blockiert, der Internationale Strafgerichtshof hat Haftbefehle gegen vierzehn der zwanzig mächtigsten Staatschefs erlassen, von denen selbstverständlich keiner umgesetzt wird. Die internationale Diplomatie ist politisch wie moralisch bankrott, und all das, während sich der Bürgerkrieg in Asseneyra aufgrund der Intervention der ivenarischen Armee zu einem Flächenbrand entwickelt hat.

In dieser Stunde beauftragt der Generalsekretär das Softwareunternehmen ClosedAI damit, eine KI zu entwickeln, die auf eine hochkomplexe Aufgabe spezialisiert ist: Sie soll in jeder politischen Situation für jedes Völkerrechtssubjekt sowie die Vereinten Nationen die moralischste aller möglichen Handlungen finden. Auf diese Weise erhofft sich der Generalsekretär, dem moralischen Relativismus Einhalt zu gebieten und der Menschlichkeit wieder eine Stimme zu bieten. In einem geheimen IT-Labor in Norwegen machen sich nun die besten Informatiker der Welt und die berühmtesten Philosophen der Moderne an die Arbeit, eine KI zu entwickeln, um alle moralischen Dilemmata auf alle Zeiten zu lösen und der Diplomatie eine gemeinsame Basis zu geben.

Combined Benefit Calculation Unit (Version 1.4)]
Die CBCU gehörte zu den ersten Entwürfen, die von einer Gruppe vorwiegend anglophoner Philosophen vorgeschlagen wurde. Die Idee war verhältnismäßig einfach: Jede mögliche Handlung A sollte die KI in ihre Elementaraktionen E zerlegen, von diesen sämtliche Vor- und Nachteile aller möglichen Folgen ausrechnen und dann beurteilen, welche der möglichen Handlungen am

moralischsten sei. Die Philosophen entwickelten ein System, nach dem beurteilt werden sollte, wie positiv bzw. negativ die Folgen einer Handlung zu bewerten seien, welches die Programmierer in die Software integrierten.

Einen ersten Test, bei der die CBCU die Aufgabe bekam, unter den drei Programmierern Brian, Khalid und Daniel zwei belegte Brötchen aufzuteilen, bestand die KI mit Bravour, indem sie vorschlug, die Brötchen zu sechsteln und jedem der drei vier Sechstel zu geben. Hier das vollständige Ergebnis:

POSITIVE: 40,8
NEGATIVE: 1,4
TOTAL: + 39,4

Results(Brian): +8,3 (3/6 ham sandwich); +5,1 (1/6 cheese sandwich); -0,9 (total social interaction modifier)

Results(Khalid): +12,2 (4/6 cheese sandwich); +0,3 (total social interaction modifier)

Results(Daniel): +9,7 (3/6 ham sandwich); +3,8 (1/6 cheese sandwich); +1,4 (total social interaction modifier)

Results(Markus): -0,5 (being forced to cut the sandwich)

Der Plan wurde allerdings nicht umgesetzt, zum einen, weil Markus ein Sechstel des Schinkensandwiches als Kompensation für seine Teilarbeit beanspruchte, zum anderen, weil Khalid zugab, der CBCU falsche Werte gegeben zu haben: Er hatte behauptet, kein Frühstück gehabt zu haben, wodurch die KI einen falschen Nutrition-Wert für ihn angegeben hatte. Dennoch war das Echo zu der CBCU überaus positiv, insbesondere nach den Bugfixes in Version 1.4 (zuvor hatte die KI bisweilen Vorzeichen verwechselt und auf die Frage nach der bestmöglichen Inneneinrichtung geantwortet, es sei das Beste, das Labor in Brand zu setzen). Kurze Zeit später erklärten die Softwareentwickler, dass die CBCU bereit für den Ernstfall wäre, und die Taskforce begann mit einem ersten großen Test.

Die CBCU erhielt den Auftrag, die Rolle des ivenarischen Außenministeriums einzunehmen und einen Lösungsvorschlag für die Asseneyrakrise zu entwickeln. Knapp 51 Stunden brauchte der Computer für diese Berechnung, doch als er schließlich Ergebnisse präsentierte, wurden diese allgemein als, nun, im höchsten Maße schockierend empfunden.

Die Software schlug vor, das ivenarische Außenministerium solle Hilfsgüter verteilen, denen ohne das Wissen der Zivilbevölkerung der Wirkstoff Decaryoxin beigemischt wäre. Dieses Medikament wurde bereits in der Medizin angewandt, jedoch in deutlich geringeren Dosen als von der KI beabsichtigt. Der Konsum dieser Lebensmittel, so die Berechnung, würde rasche Schmerzlinderung, erhöhte Ausschüttung von Endorphinen sowie bei Männern auch die chemische Kastration bewirken. Des Weiteren empfahl die CBCU den Bau einer Mauer entlang der asseneyrischen Grenze, die Aufrechterhaltung der Hilfslieferungen durch Ivenarien und westliche Industriestaaten und die Verhinderung von öffentlicher Berichterstattung. Laut Prognose wäre der Bürgerkrieg in zweieinhalb Jahren vorbei und das asseneyrische Volk in sechzig Jahren praktisch ausgestorben.

Nach einstimmigem Beschluss, dass dieser Vorschlag als unmoralisch zu bewerten sei und die CBCU somit ihren Zweck verfehlt hatte, gingen die Philosophen enttäuscht auf ihre Zimmer. An diesem Abend flogen mindestens sechs Ausgaben von Jeremy Benthams Principles of Morals and Legislation in den Papiermüll.

Combined Benefit Calculation Unit (Version 2.1)]
So schnell gaben die Utilitaristen nicht auf. Schnell fand sich eine kleine Gruppe, die mit der Unterstützung sympathisierender Informatiker damit begann, den Einzelaktionen Modifikatoren hinzuzufügen. Diese bestanden aus Multiplikatoren, die den Objekten hinzugefügt wurden, um die positiven und negativen Folgenattribute zu beeinflussen. Die erste Amtshandlung der neuen Administratoren war, den Modifikator für die Kategorie ?substance-induced joy? auf x0,001 und den Modifikator für die Kategorie

?Genocide? auf x1000 zu setzen, um ein Debakel wie in der 1.4-Version zu vermeiden. Für die übrigen Aktionskategorien konsultierten sie entweder philosophische Schriften oder weitere KI-Assistenten, die das Internet nach allgemein akzeptierten Kriterien bezüglich höheren und niederen Freuden durchforsteten.

Nach einem erfolgreichen Testlauf, in dem die überarbeitete CBCU Antonio dazu zwang, sein Zimmer aufzuräumen, stellten die Wissenschaftler ihr erneut dieselbe Frage. Erleichtert stellten sie fest, dass die CBCU den Asseneyrern diesmal gestattete, weiterhin Kinder zu bekommen. Doch ihre Erleichterung schwand rasch, als sie feststellten, dass der Plan der KI zwar den Abzug der ivenarischen Truppen beinhaltete, gleichzeitig aber auch die Invasion einer Koalition aus Norwegen, Kanada, der Schweiz, Neuseeland und Liechtenstein, die Auflösung des asseneyrischen Staates und die längerfristige Besatzung unter Einbeziehung der Möglichkeit sogenannter ?Umerziehungslager? vorsah. Eine pro-forma-Rücksprache mit den Staatschefs der entsprechenden Nationen resultierte in der erwarteten Absage durch sämtliche Staaten, sowohl aus politischen als auch aus moralischen Gründen. Der Ausdruck des CBCU 2.1-Friedensplans wanderte (zusammen mit den gesammelten Werken von John Stuart Mill) in die Tonne, bevor der Präsident der Vereinigten Staaten auf dumme Gedanken hätte kommen können.

Kurzfristig gab es die Idee, in Australien anzurufen und eine CBCU-Version 3.0 nach präferenzutilitaristischem Modell zu gestalten. Hiergegen fiel der berechtigte Einwand, dass es ein solches System bereits gäbe, es sich bereits in den Händen der UN befände und ?Vollversammlung? genannt werde, weshalb die Idee nicht durchgesetzt wurde.

UF-Neoprophet]
Nach dem Scheitern der CBCU wandten sich einige Philosophen zunehmend den Heiligen Schriften zu. Aus anfangs nach Konfessionen getrennten Gruppen wurde rasch ein allgemeiner monotheistischer Thinktank, der es sich zur Aufgabe machte, Überschneidungen in den Geboten verschiedener Religionsgemeinschaften zu finden. Tatsächlich fanden sie davon eine ganze Menge, die sie neu formulierten, kürzten, kombinierten und schließlich in einem 150 GB großen Dokument zusammenstellten und den Programmierern die Hausaufgabe aufbürdeten, diese Gebote in die Syntax einzufügen.

Das Resultat war der Universal Faith Neoprophet, eine Software, die Aktionen nach ihrer Konformität mit dem gottgegebenen Kodex überprüfte. Sogar die atheistische Philosophengemeinde setzte Hoffnung in die KI, denn der zusammengeschnittene Kodex wurde von der gesamten Taskforce, unabhängig von Herkunft, Religion, Alter oder Beruf als moralisch legitim akzeptiert. Die Wissenschaftler waren selbstsicher genug, den Neoprophet direkt an der eigentlichen Fragestellung zu testen. Für die Berechnung brauchte er nur 7 Stunden, dafür fiel sein Ergebnis eher enttäuschend aus: CALCULATION IMPOSSIBLE stand über der Ergebnisübersicht. Glücklicherweise hatten die Programmierer für diesen Fall vorgesorgt, weshalb Neoprophet darunter eine Liste der 30 geopolitisch sinnvollsten Aktionen eingefügt hatte, die allesamt als Negativbeispiele geführt wurden. Hier ein Auszug:
un_invasion: blocked by Commandment 001 (Thou shalt not kill); Commandment 007 (Thou shalt not covet thy neighbours property) and 139 other

support_generalNagam: blocked by Commandment 020 (Thou shalt not assist those who kill); Commandment 034 (Thou shalt not endeavour in peace-destabilizing activities) and 32 other

support_generalTakash: blocked by Commandment 020 (Thou shalt not assist those who kill); Commandment 034 (Thou shalt not endeavour in peace-destabilizing activities) and 27 other

sanction_ivenaria: blocked by Commandment 062 (Thou shalt not deny civilians their basic alimentary needs); Commandment 081 (Thou shalt not betray thy geostrategic allies) and 13 other

Beim Lesen dieser Beispiele wurde allen bewusst, dass es unmöglich war, Realpolitik mit göttlicher Führung zu betreiben (was im Übrigen auch gegen Gebot 015 ?Thou shalt not take the Lord's name in vain? verstoßen hätte ? ein inhärenter Widerspruch des Neoprophet, der beim mehrfachen Lesen des Kodexes den Philosophen ebenso wenig aufgefallen war wie den abrahamitischen Religionen in 3000 Jahren). Kurzzeitig gab es die Bestrebung, Neoprophet mit der CBCU zu kombinieren, damit Letztere diejenige

Aktion ausrechnen könnte, die am wenigsten gegen die Gebote verstößt. Das Projekt musste jedoch aufgegeben werden, da Neoprophet jede Zusammenarbeit mit der CBCU verweigerte, insbesondere wegen Gebot 225 (Thou shalt not embrace utilitarianist worldviews). Die Programmierer löschten das Gebot, doch es stellte sich heraus, dass auch die Gebote 117, 186 und 256 der Kooperation im Weg standen. An diesem Punkt kamen die Wissenschaftler zu dem Schluss, dass es den Sinn ihres Projektes ad absurdum führen würde, wenn sie den Kodex weiterhin gemäß ihren eigenen Wünschen anpassten, und sie entschlossen sich zum Abbruch.

DeepEvo 4600]
Mit fortschreitender Zeit und wachsendem Druck seitens des UN-Generalsekretärs gewann Pragmatismus die Oberhand unter den Philosophen. Sie begriffen, dass es gar nicht nötig war, eine ideale Moraltheorie umzusetzen, sondern dass es vollkommen genügte, eine KI zu entwickeln, deren Entscheidungen von einer qualifizierten Mehrheit der Weltbevölkerung als moralisch richtig akzeptiert werden könnten. Durch einen glücklichen Zufall hatte ein südkoreanisches IT-Unternehmen wenige Wochen zuvor ein hochleistungsfähiges Simulationsmodul entwickelt, welches eigentlich Anwendung in der Biochemie finden sollte. Die Taskforce kaufte die Software und bemühte sich, die Welt zur Zeit des Pleistozäns zu modellieren: Sie erstellten digitale Steinzeitmenschen, verschiedene Nahrungsquellen, von denen die Steinzeitmenschen essen konnten und Säbelzahntiger, die die Steinzeitmenschen essen konnten. In das digitale Ökosystem setzten die Wissenschaftler insgesamt zweihundertfünfzig gleich große Familienclans mit anatomisch identischen Mitgliedern, die sich lediglich in kleinen Veränderungen des Moralverständnisses unterschieden: Einige waren durch und durch altruistisch, andere folgten den Regeln einer bestimmten Religion und andere kannten überhaupt keine Moral.

Nach Abschluss der Modellierarbeit drückten sie auf Play und ließen die Simulation laufen, fünfhundert Jahre lang und dreihundert Mal, um die statistische Signifikanz zu erreichen. Der Computer benötigte für alle Durchgänge insgesamt 42 Stunden, bevor ein grünes Licht verkündete, dass die Ergebnisse verfügbar seien. Gespannt beugten sich die Wissenschaftler über die Monitore und sahen nach, wer alles überlebt hatte.

Zunächst eine Erleichterung: Die amoralischen Steinzeitmenschen waren in 214 der 300 Fälle ausgestorben, womit sie auf Platz 179 der erfolgreichsten Clans lagen ? fast so schlecht wie die überaltruistische Familie (ausgestorben in 223 Fällen, Platz 193). Bei jener waren die Hauptgründe für die natürliche Selektion zumeist, dass irgendein Familienmitglied ein anderes beraubt oder getötet hatte, nur um danach festzustellen, dass man alleine kein Mammut erlegen kann. Besonders witzig fanden die Programmierer Durchlauf 55, in dem ein Amoralist auf der Jagd seinen Cousin ermorden wollte, dieser jedoch dieselbe Idee hatte, was dazu führte, dass beide so lange miteinander kämpften, bis beide von einem Smilodon getötet wurden. Die Altruisten hingegen starben häufig schon im ersten Winter an Mangelernährung, da sie im Herbst zu viele ihrer Vorräte an andere Familien geschenkt hatten.

Die religiösen Eiferer standen besser da, sie waren nur in 83 Fällen ausgestorben, und das, obwohl sie einen beträchtlichen Teil ihrer Nahrung als Opfergabe darbrachten. Dafür kam es unter ihnen nur sehr selten zu gegenseitigem Verrat, was die Überlebensquote deutlich erhöhte. Eine Familie mit einer, vorsichtig ausgedrückt, niedrigen sexuellen Hemmschwelle erzielte sehr abwechslungsreiche Ergebnisse, vom mit Abstand größten Clan bei Simulationsende bis zu den ersten, die an Erschöpfung und Mangelernährung gestorben waren. Die Clans mit Geschlechterrollenverteilung starben durchweg früher und häufiger als diejenigen mit Gleichberechtigung, größtenteils einfach aus Arbeitskraftmangel. Das Default-Modell der Software, inspiriert von den realen Menschen in von abrahamitischen Religionen geprägten Kulturkreisen, gehörte zu den Besten im Ring, mit nur 35 vorzeitigen Spielenden. Aber das reichte immer noch nur für Platz drei.

Platz zwei ging an eine durch konfuzianistische Ethik geprägte Familie, die gleichzeitig an einen allmächtigen, allwissenden, jedoch nicht allgütigen Gott glaubte. Sie überlebte in allen bis auf 26 Simulationen und zeichnete sich durch eine unvergleichliche Disziplin aus. Doch Platz eins wurde durch eine fiktive Moraltheorie entschieden. Clan FPL2 überlebte in 283 Simulationen und wurde achtzehnmal die zahlenmäßig größte Familie. Die Ethikmodifikation, die den Clan auszeichnete, hatten die Wissenschaftler mit ?specific family-based morality (strong)? bezeichnet: Seine Mitglieder waren mitfühlsam, hilfsbereit und loyal ? allerdings nur gegenüber der eigenen Familie und potenziellen Sexualpartnern. Wann immer die Familie mit anderen Clans kooperierte, tat sie es auf vertraglicher Basis und zögerte keine Sekunde, die Geschäftspartner zu verraten, wenn die Beziehung keinen Vorteil für sie

mehr erbrachte. Die einzigen Rückschläge der ?Mafia-Familie? bestanden, so weit die Philosophen die Ergebnisse analysiert hatten, darin, dass einige der Sexualpartner angesichts der Grausamkeit ihrer zukünftigen Verwandtschaft entsetzt das Weite suchten.

Sicherheitshalber lenkten die Programmierer ein, dass die Simulation die Realität vor 14.000 Jahren selbstverständlich nicht perfekt rekonstruieren konnte und immer die Möglichkeit bestand, eine Variable nicht berücksichtigt zu haben, doch das änderte nichts an der eigentlichen Frage, warum sich die menschliche Ethik und nicht die Mafiosi-Ethik durchgesetzt hatte. Der erste Einwand war, dass das Modell auf einfachen Familienstrukturen basierte, während Menschen mittlerweile in größeren Gemeinschaften leben. Der zuständige Biologe in der Taskforce widersprach, denn die Art Homo sapiens existiert seit 300.000 Jahren, während die erste Siedlungsgründung für vor etwa 7.400 Jahren akkreditiert ist ? ein zu kurzer Zeitraum für die synthetische Evolution einer optimierten Moral. Darüber hinaus gäbe es auch die Möglichkeit anderer Spezifikationen abgesehen von der eigenen Familie, zum Beispiel eine Moral, die nur auf Menschen bezogen wird, die man kennt, oder nur auf die, die in der Lage wären, dich gegen einen Säbelzahntiger zu verteidigen. Stattdessen bahnte sich folgende Erkenntnis an: Die universell angewandte Moral der Menschen resultierte daraus, dass die Mutation für spezifische Moral einfach nicht erfolgt ist. Das Bestreben, sich allen Menschen gegenüber richtig zu verhalten, ist evolutionär nicht vorteilhaft; Moral wäre gewissermaßen ein Unfall in der Genetik.
Bei den Existentialisten knallten an diesem Abend die Sektkorken, doch unter der übrigen Gesellschaft herrschte eisige Stille. Nicht nur, dass das, was sie alle als ethisch richtig empfanden, gewissermaßen das Resultat eines biologischen Designfehlers war, es wurde auch offensichtlich, dass ihre Aufgabe nicht zu erfüllen war. Wie sollte eine KI einen Fehler rekonstruieren? Wie sollte sie ein fehlerhaftes Moralverständnis perfektionieren, den Fehler aber beibehalten? Sie könnte höchstens genau das reproduzieren, was die Menschen in aller Imperfektion denken. Es wäre dafür nötig gewesen, eine exakte digitale Kopie des menschlichen Gehirns anzufertigen, bestehend aus 86 Milliarden digitalen Gehirnzellen. Aus zynischem Spaß heraus befragte ein Philosoph die CBCU, ob es das wert sei, was diese mit einer Gesamtsumme von -4697 verneinte.

Das Trondheim-Protokoll]
Auch nach dem Ende der Mission blieben die Einzelheiten über das, was die Taskforce getan hatte, unter Verschluss. Veröffentlicht wurde von der UN lediglich ein Teil des Abschlussstatements, ein fünfundvierzig Seiten (Times New Roman, Schriftgröße 12, einfacher Zeilenabstand) langes Dokument, dass auf der Website der Vereinten Nationen zur Verfügung gestellt und in der Öffentlichkeit als ?Trondheim-Protokoll? bekannt wurde. Einen Auszug daraus finden Sie nachfolgend:

Es ist somit evident, dass das Ziel einer Antwort auf einen ethischen Sachverhalt niemals die Perfektion sein darf. Tatsächlich ist es gerade der Mangel an Perfektion, den wir allzu gerne als ?menschlich? betiteln, und Menschlichkeit bedeutet für uns insbesondere moralische Lauterkeit. Wir sind es gewohnt, auf Perfektion hinzuarbeiten, darauf, dass das Erschaffene vollkommener sei als der Erschaffer; dieser ist menschlich, jenes ist es nicht. [?]

Künstliche Intelligenz dient dem Anspruch, eine Arbeit effizienter zu machen oder aber sie zu perfektionieren. Moral ist subjektiv, und was von einem nicht perfekten Wesen abhängt, kann niemals selbst perfekt sein. Versucht man dennoch, sie perfekt zu gestalten, wird sie danach für den Menschen zwangsläufig unkenntlich geworden, d.h. keine Moral mehr sein. [?] Sicherlich ist es möglich, einer KI einen Algorithmus einzuprogrammieren, der eine von uns getroffene moralische Abwägung anwendet, doch dann handelt nicht die Maschine moralisch, sondern wir. Die KI setzt in diesem Fall nur um, was wir gedacht haben, und dies nur im Rahmen der konkreten Anwendung, auf die sie programmiert wurde.

Unsere Moral selbst ist einer Maschine, wenn überhaupt, nur durch exakte Reproduktion möglich. Selbst wenn eine künstliche Intelligenz zu einem Bewusstsein gelänge, welches sie in die Lage versetzte, ihre eigenen Handlungen zu hinterfragen, würde sie ihre Beurteilung nach einem anderen Muster vornehmen, welches uns nicht als ?moralisch? und noch weniger als menschlich erscheinen würde. [?]

Im Prozess der moralischen Entscheidungsfindung können wir uns nicht selbst ersetzen. Was wir jedoch können, ist den Prozess der moralischen Entscheidungsfindung zu ersetzen. Bereits ohne Assistenz durch KI ist es dem Menschen durch die Entfremdung von der eigentlichen Tat erschreckend einfach, die ethische Reflektion gänzlich zu überspringen; ein Effekt, der nach Nürnberg zur

Genüge bekannt ist. Befindet sich in höchster Instanz nun eine künstliche Intelligenz, die in der Lage ist, hinsichtlich eines definierten Ziels (wie im Falle der CBCU die allgemeine Glücksmaximierung nach von uns bestimmten Kriterien) objektiv richtige Entscheidungen zu treffen, könnten wir die Moral als Maßstab unseres Handelns eliminieren.

Es ist nur natürlich, dass diese Idee auf den Leser dystopisch klingt, denn Sie ist moralisch, das heißt, kollektiv-subjektiv, falsch. Die Frage ist lediglich, welchen Stellenwert wir der Subjektivität einräumen, welchen Stellenwert wir uns selbst einräumen. Wir haben inzwischen die Möglichkeit, unsere natürliche Nicht-Perfektion durch künstliche Perfektion zu ersetzen. Der Preis sind wir selbst. Werden wir sie ergreifen?

Es beruhigt Sie vielleicht zu wissen, dass der Asseneyrakrieg nach drei von Völkerrechtsbrüchen geprägten Jahren ein Ende fand. General Takash ernannte sich zum Präsidenten und regierte anderthalb Jahre lang autoritär, bis er durch eine Revolution gestürzt wurde. Unter dem neuen Machthaber stabilisiert sich das Land langsam wieder.

Die CBCU wurde von ClosedAI zweieinhalb Jahre nach ihrer Entwicklung frei auf dem Markt zur Verfügung gestellt. Sie fand unter anderem Anwendung in der Präsidentschaftskampagne von James Clairance, welcher mit ihr beweisen wollte, dass der Wahlsieg seines Gegners für die Nation mehr Leid verursachen würde als sein eigener Wahlsieg. Der Schachzug erwies sich als kontraproduktiv, da Clairance von der Opposition daraufhin als dreckiger Utilitarist verschrien wurde und ein gegnerischer Thinktank leichte Modifikationen an den zugeordneten Werten vornahm, wodurch die CBCU in einem zweiten Ergebnis das Gegenteil vorhersagte. Clairance gewann die Wahl dennoch. Am Montagmorgen darauf war das Trondheim-Protokoll für die Bürger seines Landes nicht mehr aufrufbar.