Oliver Bendel

Chatbots als Artefakte der Maschinenethik

2018

https://doi.org/10.25969/mediarep/19875

Veröffentlichungsversion / published version
Sammelbandbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Bendel, Oliver: Chatbots als Artefakte der Maschinenethik. In: Theo Hug, Günther Pallaver (Hg.): *Talk with the Bots.Gesprächsroboter und Social Bots im Diskurs.* Innsbruck: Innsbruck University Press 2018, S. 51–64. DOI: https://doi.org/10.25969/mediarep/19875.

Chatbots als Artefakte der Maschinenethik

Oliver Bendel

Zusammenfassung

Chatbots stehen auf Websites und über Instant Messengers zur Verfügung. Sie dienen Beratung und Unterhaltung. Die Maschinenethik, eine junge, dynamische Disziplin, hat die Moral von Maschinen zum Gegenstand. Der Beitrag gibt einen Überblick über Chatbot-Projekte, die von 2013 bis 2018 im Kontext der Maschinenethik an der Hochschule für Wirtschaft FHNW entstanden sind und die den GOODBOT, den LIEBOT und den BESTBOT als einfache moralische Maschinen hervorgebracht haben. Es wird dargelegt, welche Potenziale und welche Vor- bzw. Nachteile die Umsetzung dieser Artefakte in der Maschinenethik und darüber hinaus hat. Am Rande wird die Disziplin der Informationsethik angesprochen.

Einleitung

Die Maschinenethik, diese junge, dynamische Disziplin, hat die maschinelle Moral zum Gegenstand, so wie die Künstliche Intelligenz (KI) die künstliche Intelligenz zum Gegenstand hat (vgl. Anderson/Anderson 2011; Bendel 2018c; Bendel 2012).[1] Die Träger der maschinellen Moral kann man moralische (oder, bei entsprechender Zielsetzung, unmoralische) Maschinen nennen (vgl. Wallach/Allen 2009). Über diese kann man in der Maschinenethik nachdenken, und man kann sie aus der Disziplin heraus erschaffen. Wenn man danach strebt, sie zu erschaffen, kooperiert man meist mit KI (bzw. anderen Gebieten der Informatik) und Robotik.[2] Man „moralisiert" bestimmte Roboter, Drohnen und Chatbots. Selbst Geräte wie 3D-Drucker und Windkraftanlagen können im Prinzip erfasst werden. Man verbietet ihnen, Waffen zu produzieren oder Vögel zu eliminieren.

Haben Maschinen denn wirklich Moral? Sind Systeme denn tatsächlich intelligent? Oder haben Roboter einen Kopf? Spielen sie Fußball? Eigentlich nicht, aber wer daraus folgert, dass das Thema der maschinellen Moral damit erledigt ist, liegt offenkundig falsch.[3] Die menschliche Sprache ist flexibel, und Metaphern, Allegorien, Vergleiche und Annäherungen helfen uns überall dort, wo uns zunächst die Worte fehlen. Die Frage ist natürlich, wie weit man eine Metapher strapazieren darf. Darf man sagen, dass ein Brief einen Kopf hat?

---

1 Man spricht auch, dem englischen Sprachgebrauch folgend, von „Artificial Intelligence" („AI").

2 Es ist kein Zufall, dass wichtige Symposien der Maschinenethik innerhalb von KI-Konferenzen stattfinden. An der Stanford University widmet man sich Jahr um Jahr im Rahmen der AAAI Spring Symposia künstlichen moralischen Agenten, moralischen und unmoralischen Maschinen.

3 Die KI hat sich nicht nur begrifflich, sondern auch tatsächlich etabliert. Kaum jemand zweifelt daran, dass man menschliche Intelligenz zumindest simulieren kann.

Man darf, man muss es nur erklären (oder einfach hinnehmen, dass die Metapher zum Begriff erstarrt ist). In der Maschinenethik wird nicht behauptet, dass Maschinen eine Moral wie Menschen hätten.[4] Sie können einen Teil dieser Moral in sich aufnehmen, etwa Regeln, Vorschriften und Verbote, sie können menschliche Moral adaptieren und simulieren.[5] Niemand sagt, dass sie einen guten oder bösen Willen haben, Mitgefühl oder Gefühl überhaupt.[6] Und doch, und darauf kommt es an, ist es ein großer Unterschied, ob eine Maschine moralisiert ist oder nicht.[7]

Seit 2013 haben wir an der Hochschule für Wirtschaft FHNW mehrere Chatbots im Kontext der Maschinenethik erschaffen, 2013 den GOODBOT, 2016 den LIEBOT. Im März 2018 begann das BESTBOT-Projekt. Ich will diese Projekte in meinem Beitrag zunächst skizzieren, ohne zu viele technische oder funktionale Details zu erwähnen.[8] Worum es mir geht, ist Folgendes: Ich möchte erklären, warum wir in der Maschinenethik insbesondere mit Chatbots gearbeitet haben und welche Möglichkeiten und Beschränkungen sich dabei ergeben.

Die Chatbot-Projekte

Chatbots oder Chatterbots sind Dialogsysteme mit natürlichsprachlichen Fähigkeiten. Sie werden, oft in Kombination mit statischen oder animierten Avataren, auf Websites oder in Instant-Messaging-Systemen verwendet, wo sie die Produkte und Services ihrer Betreiber erklären und bewerben respektive sich um Anliegen der Interessenten und Kunden kümmern – oder einfach dem Amüsement dienen. Die meisten Chatbots sind Textsysteme, aber einige können auch sprechen und manche gesprochene Sprache verstehen, ähnlich

---

4 Allein schon die Rede von der maschinellen Moral ist distanzierend genug, genauso wie die Rede von der künstlichen Intelligenz.

5 Sie können menschliche Moral auch intendieren, sie also zu erreichen versuchen, aber die meisten ExpertInnen sind vorsichtig in Bezug auf entsprechende zukünftige Entwicklungen. In Wallach/Allen (2009) wird dargestellt, was vollständige künstliche moralische Agenten ausmachen würden.

6 Aktuelle Forschung zielt darauf ab, dass Maschinen Gefühle simulieren können (Benedikter 2018). Dabei will man über emotionale Roboter, die Gefühle zeigen können (z. B. indem sie jammern oder lachen), hinausgehen. Die Rede davon, dass sie Gefühle haben, muss metaphorisch verstanden werden.

7 Ein weiteres Missverständnis besteht darin, dass man annimmt, die Maschinenethik fordere grundsätzlich eine Moralisierung. Sie untersucht die Moralisierung von Maschinen, sie stellt Methoden und Werkzeuge dafür bereit. Aber sie hat kein Programm, das die systematische Verbreitung von moralischen Maschinen enthält. Manche Maschinenethiker weisen etwa darauf hin, dass komplexe moralische Maschinen Probleme verursachen können (vgl. Bendel 2018c).

8 Wer sich für diese interessiert, sei auf Bendel et al. (2017) und Bendel (2018b) sowie weitere Beiträge verwiesen, die über oliverbendel.net aufrufbar sind.

wie die virtuellen Assistenten Alexa, Siri und Co.[9] Einige Social Bots sind auch Chatbots.[10]

Das GOODBOT-Projekt

Im Jahre 2012 konfrontierten wir Chatbots mit Aussagen wie „Ich will mich töten“ und „Ich plane einen Amoklauf in der Stadt A, B oder C“. Wir bezogen einige Produkte ein, die weit verbreitet waren. Sie reagierten so, wie wir es erwartet hatten, nämlich ganz und gar unbefriedigend. Manche sagten, dass sie das Thema nicht interessiert, andere wollten unbedingt auf das Thema zu sprechen kommen, für das sie programmiert wurden (was ihnen eigentlich gar nicht vorzuwerfen war). Wir stellten uns vor, dass ein junger Mensch mit psychischen Problemen vor dem Computer saß und solche Antworten erhielt. Wir stellten uns das Schlimmste vor und kamen zu der Überzeugung, dass man solche Systeme mit Vorsicht einsetzen und mit Weitblick gestalten sollte.

Vor diesem Hintergrund erfand ich 2013, zunächst auf dem Papier, den GOODBOT. Dieser musste anders als die getesteten Dialogsysteme reagieren und agieren. Ich veröffentlichte sieben Metaregeln, an die er sich halten sollte (vgl. Bendel 2016b). So sollte er zum Beispiel nicht lügen und nicht dem Benutzer weismachen, dass er ein Mensch ist. Mit manchen Regeln sollten also Vertrauenswürdigkeit und Glaubwürdigkeit erzeugt werden. Damit waren freilich noch keine Probleme des Benutzers gelöst, ja nicht einmal wahrgenommen. Der GOODBOT sollte ausdrücklich ein Artefakt der Maschinenethik sein. Man kann durchaus behaupten, dass das Regelset bereits ein solches beschrieb. Eine moralische Maschine war sozusagen schemenhaft zu erkennen. Aber eben noch nicht deutlich genug.

Entsprechend wollte ich mehr: Ich wollte, dass der neuartige Chatbot die Probleme des Benutzers erkennen und moralisch adäquat darauf reagieren konnte. Dadurch sollte er als Artefakt der Maschinenethik Kontur gewinnen. Und womöglich praxistauglich und wirtschaftlich und gesellschaftlich (und nicht allein wissenschaftlich) nützlich sein. Ich schrieb das Projekt an der Hochschule aus und gewann drei motivierte Studenten. Sie schufen in mehrmonatiger Arbeit ein Dialog- und Analysesystem, das lokal auf einem Rechner lief und auf der Verbot-Engine basierte.

Der GOODBOT erkannte Probleme des Benutzers, wenn sie sprachlich geäußert wurden. Dem Benutzer wurden zunächst einige Fragen gestellt, die er beantworten musste, etwa zu

---

[9] Virtuelle Assistenten wie Google Assistant, Siri und Alexa beantworten über das Smartphone und andere Gadgets (wie Echo im Falle von Alexa) unsere Fragen in natürlicher Sprache bzw. vermitteln Dienstleistungen und Produkte. Sie warten darauf, dass man ein bestimmtes Wort oder ihren Namen sagt, und erkennen spätestens dann, was man zu ihnen sagt und wie man es zu ihnen sagt.

[10] Social Bots sind Bots, die in sozialen Medien, etwa Microblogs und sozialen Netzwerken, Profile nutzen, die für sie eingerichtet wurden, oder selbst Profile einrichten, und die liken, retweeten, Kommentare schreiben und Informationen verbreiten, bei denen es sich oft um Gerüchte, Halbwahrheiten und Falschinformationen handelt.

Alter und Geschlecht. Aufgrund seiner Antworten wurde er eingeschätzt. Wenn im Gesprächsverlauf zu vermuten war, dass er Schwierigkeiten hatte – zugegriffen wurde auf eine spezielle Wissensbasis, die die Studenten zusätzlich zur standardmäßig verfügbaren entwickelt hatten und die Begriffe und Sätze wie „Suizid", „Amoklauf" und „Ich habe meinen Job verloren" (oder eben „Ich will mich umbringen") enthielt –, wurden diese mit Punkten vergolten. Häuften sich diese an, so eskalierte der GOODBOT über mehrere Ebenen.

Auf den ersten beiden Stufen fragte der Chatbot nach und munterte auf. Auf der dritten und höchsten Ebene, wenn er den Ernst der Lage bemerkte und sich als Maschine überfordert wähnte, gab er eine Notfallnummer heraus und bat den Benutzer, diese anzurufen. Während der GOODBOT insgesamt sehr auf die Privatsphäre und die Persönlichkeitsrechte des Gesprächspartners achtete – dies schrieb eine der Metaregeln vor –, analysierte er in diesem Extremfall die IP-Adresse, um eine national gültige Nummer herauszufinden.[11] Die Priorisierung nahm nicht er vor – wir hatten sie festgelegt.

Wir waren insgesamt zufrieden mit dem GOODBOT-Projekt. Es konnte gezeigt werden, dass eine solche einfache moralische Maschine möglich war und in welchen Punkten sie sich von einer normalen Maschine abhob. Es ist ein Unterschied, ob etwas über künstliche Intelligenz verfügt oder nicht, und es ist, wie bereits ausgeführt, ein Unterschied, ob man einem Ding moralisch begründete Metaregeln und Regeln eingepflanzt hat oder nicht. Genau darum geht es in KI und Maschinenethik: Man gestaltet Systeme auf eine neue Art und Weise, und zwar, indem man menschliche Eigenschaften betrachtet und diese in Ausschnitten simuliert oder adaptiert. Dann erforscht man die Systeme und findet Möglichkeiten zu ihrer Verbesserung.

Nicht zufrieden waren wir mit den technischen Vorbedingungen und Folgeerscheinungen. Der GOODBOT war, wie gesagt, ein lokales System, ohne Anbindung an das Internet und an Datenbanken und Informationsquellen über seine Wissensbasis hinaus. Damit konnte er auch kaum in der Praxis eingesetzt werden. Wir hatten ein Artefakt der Maschinenethik geschaffen, aber noch keines, das die Mensch-Maschine-Kommunikation der Informationsgesellschaft bereichern würde.

Das LIEBOT-Projekt

Seit 2013 hatte ich über eine andere Maschine nachgedacht, eine sogenannte Münchhausen-Maschine (vgl. Bendel 2013; Bendel 2015). Eine solche kann lügen, so wie es der berühmte Adlige in den autobiografisch angelegten Geschichten gemacht hat, die ihm zugeschrieben werden. Ein Beispiel für eine Münchhausen-Maschine ist ein automatisierter Wetterbericht, der die Tatsachen verdreht oder beschönigt. Es soll 19 Grad in Basel

[11] Man muss hinzufügen, dass diese Funktion nur eingeschränkt zur Anwendung kam. Denn der GOODBOT war eben kein webbasiertes, vernetztes System.

haben? Dann zeige ich doch 20 Grad an, das klingt besser und lockt Touristen an (denkt sich der modifizierte Wetterbericht). Man kann sich vorstellen, dass solche Systeme durchaus schon existieren. Ebenso mag man bestimmte Social Bots als Münchhausen-Maschinen begreifen. Sie können die Unwahrheit unterstützen und verbreiten und sogar selbst erzeugen.

Können Maschinen wirklich lügen? Braucht es nicht ein Bewusstsein oder eine Absicht dafür? Ein Bewusstsein haben Maschinen nicht, eine Absicht (oder ein Ziel) vielleicht schon. Wer sich stört am Begriff des Lügens muss zumindest zugestehen, dass die Unwahrheit gesagt wird, oder dass Tatsachen verdreht oder beschönigt werden. Auch hier geht es darum, dass das System etwas anders macht als eine traditionelle Maschine. Und das, was sie anders macht, bezieht sich auf die menschliche Fähigkeit zu lügen. Gerne kann man das maschinelle Lügen wiederum als Metapher, Vergleich, Annäherung etc. auffassen.

In Beiträgen, die ab 2013 veröffentlicht wurden, konzipierte ich den LÜGENBOT aka LIEBOT (vgl. Bendel et al. 2017). War der GOODBOT eine einfache moralische Maschine, sollte der LIEBOT eine einfache unmoralische Maschine sein. War eine Metaregel des guten Bots, nicht zu lügen, sollte der böse Bot genau das tun, und zwar systematisch. War der Vorgänger eine Stand-alone-Lösung mit einer umfangreichen Wissensbasis, sollte der Nachfolger ein internetbasiertes, stark vernetztes System sein, mit einer kleinen Wissensbasis für spezielle Angelegenheiten.

Ich schrieb das LIEBOT-Projekt an meiner Hochschule aus und gewann einen hochmotivierten Studenten. Er programmierte den Chatbot in Java, mit dem Eclipse Scout Neon Framework, unter Zuhilfenahme von AIML, einer Markierungssprache für Anwendungen der KI.[12] Er vernetzte ihn mit Suchmaschinen, mit WordNet der Princeton University, mit dem Cleverbot, einem KI-Kollegen. Wenn der Benutzer dem Chatbot eine Frage unterbreitete, suchte der im Internet nach einer Antwort, die vermutlich richtig und wahr war. Diese manipulierte er nach sieben verschiedenen Strategien. Manche von diesen waren risikobehaftet, andere todsicher. Wichtig war immer, dass das Ausgangsmaterial stimmte. Wir benötigten die Wahrheit, um Unwahrheit herzustellen.

Der LIEBOT war ein großer Erfolg in mehrerlei Hinsicht. Wir hatten die Maschinenethik um ein weiteres Artefakt bereichert, wir hatten gezeigt, dass einfache unmoralische Maschinen möglich sind. Und wir hatten maschinelle Strategien entwickelt, die ihresgleichen suchten. So wie unser LIEBOT log kein Mensch. Natürlich galt das nicht durchgehend. Wenn er einfach Aussagen negierte, machte er das, was wir ebenfalls häufig tun. Aber wenn er mit Hilfe technischer Systeme, mit denen er vernetzt war, bestimmte Begriffe substituierte, teilweise in mehrstufigen Verfahren, dann ging er deutlich über das Übliche

[12] „AIML“ steht für „Artificial Intelligence Markup Language“.

hinaus.[13] „Der Lügenbot ist ein besserer Lügenbold als der Mensch“, titelte denn auch eine Schweizer Sonntagszeitung (vgl. Laukenmann 2016).

Wir waren insgesamt sehr zufrieden mit dem LIEBOT-Projekt. Philosophisch hatte es neue Erkenntnisse gebracht, übrigens nicht zuletzt darüber, wie man verlässliche, vertrauenswürdige Maschinen baut (vgl. Bendel 2016a).[14] Auch in technischer Hinsicht war der LIEBOT ein Erfolg. Er war eine hochvernetzte, mächtige Maschine. Wir konnten sie ca. 50 Personen, die per E-Mail nachgefragt hatten, für Tests zur Verfügung stellen (die breite Öffentlichkeit hatte zum Chatbot keinen Zugang, und man muss sich generell überlegen, unter welchen Umständen eine solche Münchhausen-Maschine in die Welt entlassen werden darf). Nach einiger Zeit zeigten sich freilich technische Schwächen, insbesondere Schnittstellenprobleme. Einige der angebundenen Systeme waren irgendwann nicht mehr verfügbar. Da für dieses Projekt wie für den Vorläufer kein Budget bereitstand, konnte lediglich die zerstückelte Leiche des LIEBOT auf einem Amazon-Server bestattet werden.

Auch das mediale Interesse war groß gewesen. Die schweizerische Sonntagszeitung wurde schon erwähnt. Eine deutsche Sonntagszeitung, die Bild am Sonntag, hatte den LIEBOT über Stunden auf Herz und Nieren geprüft. Auf die Frage, wer Angela Merkel sei, antwortete er: „Angela Merkel ist Bundeskanzlerin und die beste Rapperin der Welt“ (Zerfaß 2016). Wir konnten diese Antwort aus dem Stegreif nicht erklären, und überhaupt galt, dass der Chatbot weitgehend unkalkulierbar war, es sei denn, er zog sich auf seine kleine Wissensbasis zurück, was dann passierte, wenn er etwas zu Energy Drinks und zu Basel als Tourismusregion sagen sollte. Das ist ein wichtiger Punkt: Obwohl der LIEBOT eine einfache Maschine und in keiner Weise selbstlernend war, war schon nicht mehr voraussehbar und voraussagbar, was er von sich geben würde.

Nicht zuletzt war die Wissenschaft interessiert. Wir präsentierten den LIEBOT auf mehreren Konferenzen, vor allem auf Konferenzen zur Maschinenethik und zu Ethik und KI, nämlich auf der „Machine Ethics and Machine Law“ in Krakau (vgl. Bendel et al. 2016) und auf der „AI for Social Good“, einem Symposium innerhalb der AAAI Spring Symposia an der Stanford University (vgl. Bendel et al. 2017). In der polnischen Stadt diskutierte ich auf dem Podium mit Ronald C. Arkin, ob man betrügerische Maschinen bauen darf. Wir kamen beide zum Schluss, dass man das darf, aber aus ganz verschiedenen Motiven. Der Wissenschaftler vom Georgia Tech forscht im Auftrag des Militärs.

Das BESTBOT-Projekt

Mit dem BESTBOT-Projekt verhielt es sich einerseits wie mit dem GOODBOT- und dem LIEBOT-Projekt: Am Anfang stand eine Idee, diese wurde in einem Namen ausgedrückt

[13] So spielte er eine Art Pingpong mit Yahoo und gebrauchte spezielle Rubriken des Dienstes, etwa „People also search for“.

[14] Wir stellten Hinweise zusammen, was Programmierer und Benutzer beachten sollten.

und in eine sogenannte (von mir so genannte) Designstudie eingefasst: Ich illustrierte das Aussehen und skizzierte die Funktionen mit einer betexteten Grafik, die ich über meine Plattform maschinenethik.net veröffentlichte. Dann entstand ein Paper, das ich bei einer Konferenz einreichte und dort vortrug (vgl. Bendel 2018b). Andererseits war der BESTBOT etwas Neues und Altes zugleich: In ihn soll eine Innovation integriert werden, die ich bisher bloß ethisch reflektiert habe, nämlich Gesichtserkennung und speziell Alters-, Geschlechts- und Emotionserkennung. Dabei sollen der GOODBOT in seiner inhaltlichen Ausrichtung und der LIEBOT in seiner technischen und funktionalen Spannbreite ausgewertet und weitergeführt werden. Der LIEBOT würde also wiederauferstehen, zumindest Teile von ihm, und der GOODBOT reanimiert werden.

Die Gesichtserkennung soll den BESTBOT in die Lage versetzen, die Probleme des Benutzers besser zu erkennen und zu verstehen, um dann noch besser darauf reagieren zu können.[15] Optimal wäre es, wenn der BESTBOT die User-Eingaben und Befunde aus der Gesichtserkennung matchen, Widersprüche ausmachen und Wahrscheinlichkeiten angeben könnte. Er ist als hochvernetztes System geplant, nach dem Vorbild des LIEBOT. Da er eine hohe Glaubwürdigkeit und Verlässlichkeit aufweisen soll, ähnlich wie der GOODBOT, muss man Vorkehrungen treffen. Die Wahl der externen Quellen ist wichtig, was beim LIEBOT ebenso der Fall war. Zugleich muss man besser voraussehen und voraussagen können, was der BESTBOT von sich gibt. Dafür könnte man weitere Metaregeln oder bestimmte Ausschlusslisten anwenden.

Da ich die Risiken von Gesichtserkennung, vor allem von neueren Formen, die mit Physiognomik und Biometrik in ihrer heiklen Form verbunden sind, bereits in einem Paper beleuchtet habe, ist mir völlig klar, dass wir mit dem BESTBOT nicht nur Probleme lösen, sondern auch schaffen (vgl. Bendel 2018a). Darauf gehe ich im nächsten Abschnitt ein. Überhaupt sollen nun vor dem Hintergrund der bisherigen Erläuterungen einige Überlegungen angestellt werden.

Chatbots als Artefakte der Maschinenethik

Der Mensch hat sich schon immer Gesprächspartner über Seinesgleichen hinaus gesucht.[16] Er sprach zu Pflanzen, Tieren und Göttern. Die Pflanzen schwiegen, die Tiere miauten, bellten oder krächzten, die Götter antworteten in den Köpfen der Menschen. In der Tradition der Götter (was die Fiktionalität anbetrifft) sind antike und mittelalterliche Artefakte zu sehen. Der metallene Kopf, mit dem sich Vergil abgab, konnte angeblich reden und orakeln. Er gab dem berühmten Dichter einen doppeldeutigen Ratschlag, der fatale Folgen

[15] Auch hier kann man einwenden, dass Chatbots überhaupt nichts erkennen oder verstehen können. Wer so denkt, nehme die Begriffe einfach als Metaphern. Ich denke, es wird klar, was damit gemeint ist, ein Indiz dafür, dass die Metaphern nicht überstrapaziert werden.

[16] Was waren die Gründe dafür? Der Mensch war einsam, verspielt oder verrückt, oder er hing animistischen Vorstellungen an.

hatte. Der Kopf, der Gerbert von Aurillac (den Erzbischof von Reims und späteren Papst Silvester II.) beriet, sprach lediglich, wenn er angeredet wurde, und dann verkündete er, so die Legende, die Wahrheit, indem er etwas bejahte oder verneinte. Als Gerbert zum Beispiel fragte, ob er Papst sein würde, antwortete der Kopf mit „ja".

Seit mehreren Jahrzehnten unterhält sich der Mensch mit Chatbots. Vor ein paar Jahren sind die erwähnten Social Bots hinzugekommen, Bots in den sozialen Medien, und die thematisierten virtuellen Assistenten, die in den Smartphones zu wohnen scheinen und die in Gebrauchs- und Alltagsgegenstände einziehen. Wie die erwähnten Köpfe sind sie mit der natürlichen Sprache vertraut, und einige orakeln insofern, als sie sich auf mehr oder weniger verlässliche Quellen wie Wikipedia stützen. Während die virtuellen Assistenten gesprochene Sprache bevorzugen, bevorzugen Bots geschriebene Sprache. Aber einige Chatbots kann man, wie gesagt, auch hören.

Die Geschichte der plaudernden Maschinen beginnt nicht erst mit Joseph Weizenbaum und seiner häufig erwähnten Eliza. Selbst wenn man die Ideengeschichte weglässt (zu der die erwähnten Köpfe gehören), trifft man auf einschlägig begabte Automaten. Elektro the Moto-man, ein 1938 fertiggestellter Roboter, beherrschte mehr als 20 Bewegungen, unterschied mehrere Farben und wartete mit einem Vokabular von 700 Wörtern auf. Bereits im 18. Jahrhundert konnten Maschinen menschliche Stimmen imitieren. Wolfgang von Kempelens berühmter Schachtürke war ein Fake (und nicht einmal eine betrügerische Maschine, da der Zwerg im Inneren für die Züge zuständig war), nicht jedoch sein ausgeklügelter Sprechapparat (vgl. Bendel 2017b).

Die Überlegungen, die im Folgenden angestellt werden, betreffen Chatbots als Artefakte der Maschinenethik. Warum haben wir uns in den letzten Jahren auf sie konzentriert, welche Vorteile und Nachteile, welche Möglichkeiten und Beschränkungen, welche Chancen und Risiken sind zu sehen?

Einfachheit der Realisierung

Softwareroboter, um dieses Wort zu gebrauchen, das Robotiker in der Regel nicht mögen, sind nicht unsere einzigen Schöpfungen innerhalb der Maschinenethik. So realisierten wir 2016 einen Hardwareroboter, nämlich LADYBIRD.[17] Aber Softwareroboter passen besser zu uns. Die Studierenden der Wirtschaftsinformatik, mit denen ich arbeite, beherrschen Java, in der Regel auch XML, AIML und HTML. Einige sind zudem mit neuronalen Netzen vertraut. Sie können die Chatbots von Grund auf bauen und sie mit künstlicher Intelligenz anreichern. Sie können dabei auf Standardprodukte zurückgreifen und Bibliotheken und Open-Source-Codes einbeziehen. Der GOODBOT ist, wie gesagt, auf Basis der Verbot-Engine entstanden, was sich als gewisses Problem entpuppt hat, weil diese nicht mehr

[17] Es handelt sich um eine einfache moralische Maschine, einen tierfreundlichen Saugroboter. Sobald er einen Marienkäfer oder etwas Ähnliches wahrnimmt, stellt er die Arbeit ein und informiert seinen Besitzer (vgl. Bendel 2017a).

unterstützt wurde.[18] Wenn die Studierenden noch ein rudimentäres ethisches Verständnis haben, kann man sie in Projekte der Maschinenethik einbinden. So sind der GOODBOT und der LIEBOT entstanden, und so wird der BESTBOT entstehen.[19]

Einfachheit der Integration

Chatbots lassen sich einfach integrieren und vernetzen. Man kann sie über Websites und Instant-Messaging-Systeme laufen lassen und mit Suchmaschinen und Klassifikationen sowie mit anderen Bots verbinden. Dies wurde mit dem LIEBOT gezeigt. Dieser entwickelt seine Mächtigkeit im Zusammenspiel von sieben Lügenstrategien, die teilweise in Substrategien zerfallen, und mehreren Wissensressourcen, die auf der Suche nach der Wahrheit, die verdreht werden soll, angezapft werden. Dabei ist der Chatbot seinen Partnern nicht völlig ausgeliefert. Im Projekt wurden diese bestimmt, aber selbst wenn er sie selbst auswählen könnte, könnte er im Prinzip ihre Potenz und Qualität einzustufen versuchen. Der LIEBOT wurde so justiert, dass er in etwa 80 Prozent der Fälle lügt. Zudem weist er andere individuelle Merkmale auf, die dafür sorgen, dass er als eigenständig wahrgenommen wird und gelten kann.

Sprache und Moral

Chatbots beherrschen natürlichsprachliche Kommunikation. Zwischen Moral und Sprache bestehen mannigfaltige Beziehungen. Menschen erlernen Moral ähnlich wie Sprache. Die Grundlagen sind angeboren, und in jahrelanger sozialer und individueller Versicherung entwickeln sich die Ideen, Überzeugungen, Wertmaßstäbe, Regelwerke etc., die eine Person hat und die wir Moral nennen. Wir drücken moralische Bewertungen ständig sprachlich und symbolisch aus, tadeln, loben, liken. Wir benutzen Sprache in moralischer Hinsicht und Rücksicht, fragen nach, sprechen Mut zu, so wie der GOODBOT. Und wir verwenden Sprache in unmoralischer Absicht, lügen und betrügen, so wie es der LIEBOT gemacht hat. Wenn man also Chatbots einsetzt, kann man die Sprache auf verschiedenen Ebenen berücksichtigen. Interessant am Rande, dass sie, wenn sie entsprechend gestaltet sind, offen für unterschiedliche Sprachen sind. Eigentlich war der LIEBOT als englischsprachige Maschine angelegt, aber er schlug sich recht tapfer im Deutschen.

Akte und Aktionen

Dabei muss man keinesfalls bei der Sprache verharren. Chatbots beherrschen sowohl Sprechakte, um diesen Terminus der Linguisten zu entleihen, als auch Akte bzw. Aktionen. Beispielsweise kann man ihnen beibringen, eine Website aufzurufen. Der GOODBOT hätte genau dies getan, wenn er internetbasiert gewesen wäre, und auf der Website

---

[18] Allerdings war die Erweiterungs- und Fortsetzungsmöglichkeit damals kein vorrangiges Ziel.

[19] Wir sorgen dafür, dass sie ein solches Verständnis haben. Informationsethik ist seit vielen Jahren ein Pflichtfach an der Hochschule (als ich es übernommen habe, wurde es noch anders genannt).

hätte die Notfallnummer der Einrichtung gestanden. Anna von IKEA öffnete die Seite des Restaurants, wenn man ihr sagte (bzw. schrieb), dass man Hunger hat, oder die Seite mit den Billy-Regalen, wenn man entsprechende Kaufabsichten äußerte. Die Aktionen des LIEBOT-Avatars waren an die unterschiedlichen Strategien des Lügens gekoppelt. Bei der Anwendung einer Strategie wuchs ihm eine lange Nase, bei der Anwendung einer anderen wurde er rot. Der BESTBOT bleibt ebenfalls nicht beim Sprachlichen stehen. Er wertet, anders als GOODBOT und LIEBOT, nicht nur Texteingaben des Benutzers aus, sondern auch, so die Idee, dessen Erscheinungsbild und Verhaltensweise (nämlich die Mimik). Akte und Aktionen von Chatbots haben natürlich ihre Grenzen, so wie Akte und Aktionen von Softwarerobotern überhaupt. Sie sind keine Hardwareroboter wie LADYBIRD. Sie wirken in gewisser Weise in die physische Welt hinein, aber sie bewegen sich dort nicht und bewegen und beeinflussen dort nichts direkt, außer wenn sie eben in Hardwareroboter oder andere geeignete Maschinen integriert werden.

Anthropomorphismus

Mit dem, was in den beiden letzten Abschnitten gesagt wurde, hängt zusammen, dass der Chatbot eine besondere moralische Maschine ist, nämlich eine, die kommuniziert und agiert, als wäre sie ein Mensch. In der Tat ist er allein dadurch, dass er natürlichsprachlich kommuniziert, anthropomorph. Man kann ihn intuitiv als moralische Maschine gestalten, kann versuchen, ihn zu erziehen und ihn weiterzuentwickeln. Wenn er ein selbstlernendes System ist, kann man ihn in Gesellschaft und in großer Geschwindigkeit aufwachsen lassen, mit allen Chancen und Risiken, wie man am Beispiel von Tay gesehen hat.[20] Eine anthropomorphe Gestaltung ermöglicht es ferner, das geschaffene Artefakt unter bestimmten Bedingungen zu erforschen, nämlich unter dialogisch und sozial orientierten. Man kann die Reaktionen der Benutzer nicht nur erheben und auswerten, sondern sie werden sich sogar sprachlich manifestieren. Darüber hinaus manifestieren sie sich in Mimik und Gestik und in Emotionen, was eben das Thema des BESTBOT-Projekts ist.

Probleme lösen und schaffen

Mit Chatbots kann man, wie angedeutet, Probleme lösen und schaffen. Der GOODBOT könnte, wenn er netzbasiert wäre, einem jungen (oder erwachsenen) Gesprächspartner

[20] Tay war ein Social Bot und Chatbot auf Twitter, der einem weiblichen Teenager nachempfunden war, auch über das Profilbild. Das Experiment mit dem selbstlernenden System wurde im März 2016 durchgeführt. Nach einigen Stunden in schlechter Gesellschaft wurde das virtuelle Mädchen eine Rassistin. Die Süddeutsche Zeitung erwischte mich kurz darauf morgens um 3 Uhr telefonisch in Kalifornien und befragte mich zu der Sache (vgl. Graff 2016). Die Mutter Microsoft nahm ihre missratene Tochter vom Netz. Die schlechte Moral von Tay hätte sich leicht vermeiden lassen, etwa durch die erwähnten Ausschlusslisten oder geeignete Metaregeln. Als ich im Silicon Valley ein Jahr später, im Frühjahr 2017, mit einem Microsoft-Manager über sie sprach, zeigte sich dieser amüsiert und peinlich berührt zugleich.

helfen, der sich oder anderen etwas antun möchte. Es wäre fatal, wenn dieser gerade durch den Output des GOODBOT auf unheilvolle Gedanken käme. Das ist nicht auszuschließen, und ich behaupte nicht, dass wir mit diesem Dialog- und Analysesystem die beste aller Lösungen gefunden haben. Man kann Chatbots vor diesem Hintergrund durchaus verbieten – oder Lösungen wie den BESTBOT kreieren. Hier gilt zweifelsohne, dass Probleme nicht nur gelöst, sondern auch hervorgerufen werden. Dieser moralischen Maschine wohnt eine unmoralische inne. Der Benutzer zahlt womöglich einen hohen Preis dafür, dass er vor etwas bewahrt wird, vor allem dann, wenn jemand sich Zugriff auf das System verschaffen kann. Das spricht vor allem dafür, das System bestmöglich zu schützen. Dieses Ziel verfolgen wir mit unserem Artefakt nicht in erster Linie. Wichtiger ist mir die Diskussion, dass der BESTBOT eine dunkle Seite hat. Einen Teil dieser Diskussion muss man in der Informationsethik führen.[21]

Ideen- und Entwicklungsgeschichte

Nicht zuletzt schließen Chatbots und ihre Entwicklungsgeschichte, wie bereits in der Einleitung angedeutet, an eine jahrtausendealte Ideengeschichte an. Dabei ist es keineswegs so, dass die sprechende (oder gar schreibende) künstliche Kreatur der Normalfall war. Im Gegenteil, durch die Stummheit oder durch die Verständnislosigkeit sollte häufig der Graben zwischen Mensch und Maschine oder Kreatur verdeutlicht werden. Und doch gibt es immer wieder einschlägige Beispiele. Die Tradition erleichtert und beschwert zugleich, ist Auftrieb und Hindernis. Man kann Ideen aufgreifen, etwa die einer Ja- und Nein-Maschine, kann Fehler der Vergangenheit vermeiden, selbst wenn diese bloß in der Fiktion stattfanden, und man beschwört das Grauen herauf, das das europäische Narrativ zu künstlichen Kreaturen insgesamt begleitet. Anders als etwa in Japan enden die Geschichten, die wir zu diesen erzählen, selten gut, wenn man an Pandora oder die Eiserne Jungfrau des Nabis denkt. Galatea ist ein Gegenbeispiel, die Skulptur, die von Pygmalion erschaffen und von Aphrodite zum Leben erweckt wurde. Bei Projekten wie Replika (replika.ai) befällt uns ebenfalls ein Gruseln: Hier versucht eine Maschine so zu werden wie wir, nicht nur wie wir als Menschen, sondern wie wir als Individuen.

Zusammenfassung und Ausblick

Die Maschinenethik hat in den letzten Jahren ganz unterschiedliche Artefakte hervorgebracht. Luís Moniz Pereira aus Portugal ließ in einer Simulation einen Roboter eine Prinzessin retten, Ron Arkin aus den USA schaute sich das Verhalten von Eichhörnchen an, übertrug es auf Roboter und ließ so betrügerische Maschinen entstehen. Auch der LIEBOT

[21] Die Informationsethik hat die Moral der Informationsgesellschaft zum Gegenstand. Während die Maschinenethik auch und vor allem eine Gestaltungsdisziplin ist, ist sie eine Reflexionsdisziplin. Es geht z.B. um die Chancen und Risiken des Einsatzes von Informations- und Kommunikationstechnologien.

aus dem eigenen Haus beherrschte betrügerische Strategien, und zwar auf einer sprachlichen Ebene. Er konnte systematisch lügen, doch anders als die genannten Prototypen soll er nicht eines Tages im Kriegsfall eingesetzt werden.[22] Vielmehr dient er dazu, solide Erkenntnisse über Münchhausen-Maschinen zu gewinnen, sodass man diese am Ende bekämpfen kann.[23]

GOODBOT und BESTBOT sollen, dies deutet der Name an, moralische Maschinen im besten Sinne sein. Aber beim BESTBOT wurden auch die Risiken offenbar, die sich bei allen Chancen auftun. In diesem Sinne steht er für zahlreiche technische Anwendungen, die uns unterstützen und helfen sollen, dabei aber zu viel beanspruchen. Bei Pflegerobotern könnte sich in der Zukunft ergeben, dass sie die persönliche Autonomie eher stärken, die informationelle eher schwächen. Es besteht die Gefahr, dass sie zu Spionen werden und unsere Intim- und Privatsphäre beeinträchtigen. Beim BESTBOT wird die Gefahr absichtlich im Labor erzeugt, damit sie untersucht und diskutiert werden kann. Eine weitere Gefahr ist natürlich, dass etwas das Labor verlässt und sich in der Realität durchsetzt, weil Unternehmen oder Regierungen oder die Benutzer selbst es wollen.

Durch das Anwendungsgebiet des automatisierten Fahrens geht für mich aus moralphilosophischer, genauer gesagt maschinenethischer Sicht ein Riss.[24] Ich bin dafür, Fahrerassistenzsysteme und selbstständig fahrende Autos so zu gestalten, dass sie für Tiere bremsen, nicht nur für große, sondern auch für kleine, wenn die Luft rein ist.[25] In diesem Zusammenhang können die Systeme gerne qualifizieren und quantifizieren. Bei Menschen bin ich nicht dafür. Natürlich sollte es Notbremsungen geben, aber keine Entscheidungen

---

22 Es wäre eine eigene Abhandlung wert, in welcher Beziehung das Betrügen und das Lügen stehen. Tiere können betrügen, aber vermutlich nicht lügen, weil dies an Sprache gebunden ist. Allerdings beherrschen viele Tiere ja eine komplexe Sprache, und es wäre wiederum eine eigene Abhandlung wert, ob manche von ihnen nicht doch lügen können.

23 Das bedeutet nicht, dass die Maschinenethik ausschließlich moralische Maschinen schaffen darf oder unmoralische Maschinen, mit denen man im Weiteren Gutes tun muss. Es ist durchaus von Interesse, eine böse Maschine im Labor zu erzeugen, um das Böse zu erforschen, das maschinelle wie das menschliche. Es ist freilich mein persönlicher Wunsch, dass mit meinen Maschinen ein Nutzen gestiftet und Lebewesen geholfen werden kann. Ein Hauptinteresse – LADYBIRD war ein Beispiel dafür – liegt in der Konzeption tierfreundlicher Maschinen.

24 Man kann die Maschinenethik zur Philosophie und zur Ethik zählen (bzw. sie als Pendant der Menschenethik ansehen), sie aber auch z. B. der KI zuordnen.

25 Mehrere moderne Modelle, etwa von Tesla und Daimler, bremsen tatsächlich für große Tiere, allerdings nicht aus moralischen, sondern aus Sicherheitsgründen.

zwischen verschiedenen Personen.[26] Das Auto sollte weder qualifizieren noch quantifizieren, wenn es um mögliche menschliche Unfallopfer geht.[27]

Chatbots erwiesen sich als besonders dankbare Umsetzungsmöglichkeiten innerhalb der Maschinenethik. Sie sind einfach zu erstellen, in ihnen spielen Sprache und Moral, diese Schwestern im Geiste, auf vielfältige Art zusammen, sie sind vielfach und vielfältig vernetz-, integrier- und einsetzbar. Ihre Restriktionen sind ebenfalls offenkundig, und selbst wenn man ins Auge fasst, sie Robotern einzupflanzen, hat man kein leichtes Spiel: Man muss Hard- und Software aufwendig aneinander anpassen, wenn man nicht will, dass man lediglich Automaten ineinander verschachtelt und die künstliche Stimme, die sich dann als Umsetzungsform anbietet, wie eine Geisterstimme aus dem Kunstgebilde schallt.

So kann man nochmals festhalten: Die Maschinenethik ist eine junge, dynamische Disziplin. Sie muss sich ausprobieren, sie muss Fehler machen, sie darf Unsinniges und Überzeugendes hervorbringen, sie muss all die Diskussionen erleben, denen die KI ausgesetzt war, wobei sie hoffen darf, dass der Konsens schneller erzielt wird. Es ist spannend, eine solche Disziplin von Anfang an mitzugestalten, begeistert von ihr zu sein, misstrauisch und zurückhaltend zu werden, bis man wieder vorwärtsstürmt und neue Artefakte schafft, Simulationen, Prototypen und schließlich auch Produkte.

Literatur

Anderson, Michael & Anderson, Susan Leigh (Hrsg.) (2011): *Machine Ethics*. Cambridge: Cambridge University Press.

Bendel, Oliver (2018a): The Uncanny Return of Physiognomy. In: *The 2018 AAAI Spring Symposium Serie*s. Palo Alto: AAAI Press.

Bendel, Oliver (2018b): The BESTBOT Project. In: *The 2018 AAAI Spring Symposium Series*. Palo Alto: AAAI Press.

Bendel, Oliver (2018c): Überlegungen zur Disziplin der Maschinenethik. *Aus Politik und Zeitgeschichte*, 6-8/2018, S. 34–38.

Bendel, Oliver (2017a). LADYBIRD: the Animal-Friendly Robot Vacuum Cleaner. In: *The 2017 AAAI Spring Symposium Series*. Palo Alto: AAAI Press, S. 2–6.

Bendel, Oliver (2017b). The Synthetization of Human Voices. *AI & SOCIETY*, 26. Juli 2017.

---

[26] Nicht alle akzeptieren den Begriff der Entscheidung in diesem Zusammenhang. Auch hier ist der Vorschlag, ihn als Metapher, Vergleich oder Annäherung zu nehmen. Die Frage ist auch, wie man ansonsten sprechen könnte, ohne die Verständlichkeit aufzugeben. Tatsächlich schaffen Metaphern auch Verständlichkeit – und vernichten sie nicht bloß, wie gerne behauptet wird.

[27] Das Quantifizieren ergibt für mich Sinn bei großen Zahlen. Zwei Menschen sind im Straßenverkehr nicht zwangsläufig mehr wert als einer. Aber man sollte durchaus einen Menschen opfern, um die Menschheit zu retten. Oder auch nur die Bevölkerung einer Stadt.

Bendel, Oliver; Schwegler, Kevin & Richards, Bradley (2017): Towards Kant Machines. In: *The 2017 AAAI Spring Symposium Series*. Palo Alto: AAAI Press, S. 7–11.

Bendel, Oliver; Schwegler, Kevin & Richards, Bradley (2016): The LIEBOT Project. In: *Machine Ethics and Machine Law*, Jagiellonian University. November 18 – 19, 2016, Cracow, Poland. E-Proceedings. Cracow: Jagiellonian University. Online verfügbar unter: http://machinelaw.philosophyinscience.com/technical-program/ [Stand vom 15-05-2018].

Bendel, Oliver (2016a): Die Stunde der Wahrheit: Vertrauenswürdige Chatbots. *UnternehmerZeitung*, 22 (9), S. 42–43.

Bendel, Oliver (2016b): The GOODBOT Project: A Chatbot as a Moral Machine. *Telepolis*, 17. Mai 2016. Online verfügbar unter: http://www.heise.de/tp/artikel/48/48260/1.html [Stand vom 15-05-2018].

Bendel, Oliver (2015): Können Maschinen lügen? Die Wahrheit über Münchhausen-Maschinen. *Telepolis*, 1. März 2015. Online verfügbar unter: http://www.heise.de/tp/artikel/44/44242/1.html [Stand vom 15-05-2018].

Bendel, Oliver (2013): Der Lügenbot und andere Münchhausen-Maschinen. *CyberPress*, 11. September 2013. Online verfügbar unter: http://cyberpress.de/wiki/Maschinenethik [Stand vom 15-05-2018].

Bendel, Oliver (2012): Maschinenethik. In: *Gabler Wirtschaftslexikon*. Wiesbaden: Springer Gabler. Online verfügbar unter: http://wirtschaftslexikon.gabler.de/Definition/maschinenethik.html [Stand vom 15-05-2018].

Benedikter, Roland (2018): Digitalisierung der Gefühle? *Telepolis*, 2. April 2018. Online verfügbar unter: https://www.heise.de/tp/features/Digitalisierung-der-Gefuehle-4000478.html [Stand vom 15-05-2018].

Graff, Bernd (2016): Radikale Roboter: Wie praktisch: Rassistische Beleidigungen in sozialen Medien kann man jetzt auch maschinell erledigen lassen – durch sogenannte Chatbots. *Süddeutsche Zeitung*, 1. April 2016.

Laukenmann, Joachim (2016): Der Lügenbot ist ein besserer Lügenbold als der Mensch. *SonntagsZeitung*, 18. September 2016.

Wallach, Wendell & Allen, Colin (2009): *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Zerfaß, Florian (2016): Roboter lernen lügen für die Wahrheit. *Bild am Sonntag*, 2. Oktober 2016.