

MAI4CAREU

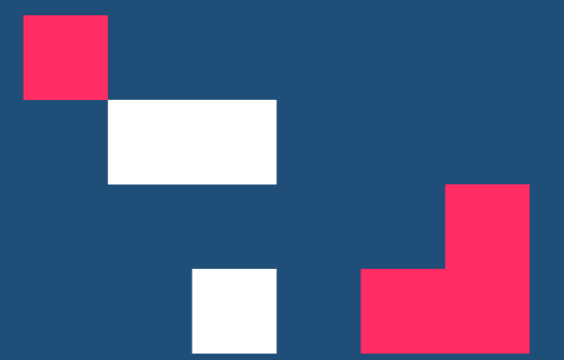
Master programmes in Artificial
Intelligence 4 Careers in Europe

University of Bologna

Computational Ethics

Daniela Tafani

2022/2023 – Second Semester



KI und magisches Denken

KI als Technologie und "KI" als Sprechakt

We need to distinguish between

- 1. artificial intelligence (AI) as a technology with practical application:** “as a technology, AI exists somewhere on a spectrum from, practically, at one end, expert systems, path planners, and practical reasoning systems [...] through to, theoretically, at the other end, Alan Turing’s “imaginable digital computers which would do well in the imitation game” or John Haugeland’s synthetic intelligence (i.e., machine intelligence that is constructed but not necessarily imitative)”;
- 2. “artificial intelligence” (“AI”) as a speech act with conventional force:** “a social constructor that stems largely from science fiction with computers and robots having hugely overblown capabilities and a tendency to the apocalyptic”.
**“People have been, and are being, “encouraged” to think about artificial intelligence wrongly.
Companies are leveraging “AI” to exert control without responsibility.**

Peter R Lewis, Stephen Marsh, Jeremy Pitt, *AI vs «AI»: Synthetic Minds or Speech Acts*, in «IEEE Technology and Society Magazine», 2021, <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9445758>

KI als Technologie und "KI" als Sprechakt

Wir müssen unterscheiden zwischen

- 1. Künstliche Intelligenz (KI) als eine Technologie mit praktischer Anwendung:** Als Technologie befindet sich KI irgendwo auf einem Spektrum, das praktisch von Expertensystemen, Pfadplanern und praktischen logischen Systemen [...] bis hin zu Alan Turings "vorstellbaren digitalen Computern, die sich gut im Nachahmungsspiel schlagen würden" oder John Haugelands synthetischer Intelligenz (d.h. maschinelle Intelligenz, die konstruiert ist, aber nicht notwendigerweise nachahmt) reicht;
- 2. Künstliche Intelligenz ("AI") als Sprechakt mit konventioneller Kraft:** ein soziales Konstrukt, das größtenteils aus der Science-Fiction stammt, in der Computer und Roboter überzogene Fähigkeiten und eine Tendenz zum Apokalyptischen haben. **Die Menschen wurden und werden "ermutigt", falsch über künstliche Intelligenz zu denken. Die Unternehmen nutzen die "KI", um Kontrolle auszuüben, ohne Verantwortung zu übernehmen.**

Peter R Lewis, Stephen Marsh, Jeremy Pitt, *AI vs «AI»: Synthetic Minds or Speech Acts*, in «IEEE Technology and Society Magazine», 2021, <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9445758>

Das Problem der "vertrauenswürdigen KI"

“The problem of “trustworthy AI” is one that has great many different “sides.” On the one hand, there are guidelines (for example, from the EU) that tell us how AI should be built and/or behave in order to be seen as “trustworthy”—presumably this means that people are going to (should? must?) trust it.

On the other hand, the problem is seen as “We shouldn’t have to trust AI” because it is a “made thing” and, since it is a human artifact, humans should be held responsible (accountable) when it does something wrong.

In many cases, when they are using marketing speak, those who claim “AI” can be seen as “trustworthy” also claim that it is “beyond the control” of its creators when it leaves the shop floor.”

“It’s not just an evasion of responsibility; it is an exercise in power and it is profoundly wrong.”

Peter R Lewis, Stephen Marsh, Jeremy Pitt, *AI vs «AI»: Synthetic Minds or Speech Acts*, in «IEEE Technology and Society Magazine», 2021, <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9445758>

Das Problem der "vertrauenswürdigen KI"

Das Problem der "vertrauenswürdigen KI" hat sehr viele verschiedene "Seiten". Auf der einen Seite gibt es Richtlinien (z. B. von der EU), die uns sagen, wie KI gebaut sein und/oder sich verhalten sollte, um als "vertrauenswürdig" zu gelten - was vermutlich bedeutet, dass die Menschen ihr vertrauen werden (sollen? müssen?).

Auf der anderen Seite wird das Problem so gesehen: Wir sollten der KI nicht vertrauen müssen, weil sie eine "gemachte Sache" ist, und da sie ein menschliches Artefakt ist, sollte der Mensch dafür verantwortlich (rechenschaftspflichtig) gemacht werden, wenn sie etwas falsch macht.

In vielen Fällen, wenn sie sich der Marketingsprache bedienen, behaupten diejenigen, die behaupten, dass "KI" als "vertrauenswürdig" angesehen werden kann, auch, dass sie "jenseits der Kontrolle" ihrer Schöpfer ist, wenn sie die Werkshalle verlässt.

Das ist nicht nur eine Flucht vor der Verantwortung, sondern auch eine Ausübung von Macht, und das ist zutiefst falsch.

Peter R Lewis, Stephen Marsh, Jeremy Pitt, *AI vs «AI»: Synthetic Minds or Speech Acts*, in «IEEE Technology and Society Magazine», 2021, <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9445758>

“We suggest that a democratization of both “AI” and AI is necessary in order to better inform the people who are affected by this deceit. It is not satisfactory to blame the computer—indeed it never has been, yet since we’ve had them, we’ve tried to do exactly that—what is needed is the means to *explain*:

What the system is doing;

Why it does what it does;

How it does this thing;

Why it does it this way;

In ways that the people affected by it understand.

This should not be the responsibility of the machine, since we do not (yet) have AI capable of bearing responsibility for its behavior and operation.”

Peter R Lewis, Stephen Marsh, Jeremy Pitt, *AI vs «AI»: Synthetic Minds or Speech Acts*, in «IEEE Technology and Society Magazine», 2021, <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9445758>

Wir schlagen vor, dass eine Demokratisierung sowohl der "KI" als auch der KI notwendig ist, um die Menschen, die von diesem Betrug betroffen sind, besser zu informieren. Es ist nicht zufriedenstellend, dem Computer die Schuld zu geben - das war es noch nie, aber seit wir sie haben, versuchen wir genau das zu tun - was wir brauchen, sind Mittel, um zu erklären:

Was das System tut;

Warum es tut, was es tut;

Wie es diese Dinge tut;

Warum es das auf diese Weise tut;

Und zwar so, dass die Menschen, die davon betroffen sind, es verstehen.

Dies sollte nicht in der Verantwortung der Maschine liegen, da wir (noch) keine KI haben, die in der Lage ist, Verantwortung für ihr Verhalten und ihren Betrieb zu übernehmen.

Peter R Lewis, Stephen Marsh, Jeremy Pitt, *AI vs «AI»: Synthetic Minds or Speech Acts*, in «IEEE Technology and Society Magazine», 2021, <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9445758>

KI und magisches Denken

Künstliche Intelligenz ist Gegenstand einer Konstellation von Narrativen - d. h. von Ideen, die in Form von Geschichten verbreitet werden -, die drei für magisches Denken typische Merkmale aufweisen:

1. die Tendenz, sich bestimmte Objekte der Technik anthropomorph vorzustellen;
2. der magische Zug, ein Ergebnis oder eine Wirkung zu zeigen, während man gleichzeitig seine konkreten Ursachen und Kosten verschweigt;
3. der Glaube an die Vorhersagbarkeit des zukünftigen Verhaltens jeder einzelnen Person (ein Glaube, der wie die Astrologie auf einer verfeinerten Mathematik und einer hybriden Mischung aus Aberglaube und Wissenschaft beruht).

D. Tafani, *What's wrong with "AI ethics" narratives*, in «Bollettino telematico di filosofia politica», 2022, pp. 1-22.

Die Animation des Unbelebten

- As David Hume wrote in *The Natural History of Religion*, “there is an universal tendency among mankind to conceive all beings like themselves, and to transfer to every object those qualities with which they are familiarly acquainted and of which they are intimately conscious”.
- **“the animation of the inanimate” – is, according to Freud, the very nature of magical thinking:** “the misunderstanding” whereby we “put psychological laws in place of natural ones” is still present “in the life of today”, “in living form, as the foundation of language, our beliefs and our philosophy”.
- It is a well-known and yet irresistible tendency: emotional and social responses are automatically generated also by media, such as televisions or computers, and overcoming this unconscious impulse would require the effort of a continuous reflection and the employment of a technical vocabulary, different for each type of object and unfamiliar to most of us.

B. Reeves, C. Nass, [The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places](#), Cambridge, Cambridge University Press, 1996.

D. Tafani, *What's wrong with “AI ethics” narratives*, in «Bollettino telematico di filosofia politica», 2022, pp. 1-22.

Die Animation des Unbelebten

- Wie David Hume in *The Natural History of Religion* schrieb - gibt es eine allgemeine Tendenz unter den Menschen, sich alle Wesen wie sich selbst vorzustellen und auf jeden Gegenstand die Eigenschaften zu übertragen, mit denen sie vertraut sind und derer sie sich innig bewusst sind.
- **Die Belebung des Unbelebten - ist nach Freud das Wesen des magischen Denkens:** "Das Missverständnis", durch das wir "psychologische Gesetze an die Stelle natürlicher Gesetze setzen", ist "im heutigen Leben" noch immer präsent, "in lebendiger Form, als Grundlage der Sprache, unserer Überzeugungen und unserer Philosophie".
- Es handelt sich um eine bekannte und doch unwiderstehliche Tendenz: emotionale und soziale Reaktionen werden automatisch auch durch Medien wie Fernsehen oder Computer erzeugt, und die Überwindung dieses unbewussten Impulses würde die Anstrengung einer kontinuierlichen Reflexion und die Verwendung eines technischen Vokabulars erfordern, das für jede Art von Objekt unterschiedlich ist und den meisten von uns nicht vertraut ist.

B. Reeves, C. Nass, [The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places](#), Cambridge, Cambridge University Press, 1996.

D. Tafani, *What's wrong with "AI ethics" narratives*, in «Bollettino telematico di filosofia politica», 2022, pp. 1-22.

ELIZA

“I chose the name ELIZA for the language analysis program because, like the Eliza of Pygmalion fame, it could be taught to “speak” increasingly well.

Because conversations must be about something, that is, because they must take place within some context, the program was constructed in a two-tier arrangement, the first tier consisting of the language analyzer and the second of a script. The script is a set of rules rather like those that might be given to an actor who is to use them to improvise around a certain theme. Thus ELIZA could be given a script to enable it to maintain a conversation about cooking eggs or about managing a bank checking account, and so on. Each specific script thus enabled ELIZA to play a specific conversational role.”

J. Weizenbaum, *Computer Power and Human Reason. From Judgement to Calculation*, San Francisco, W.H. Freeman & Company, 1976.

ELIZA

Ich habe den Namen ELIZA für das Sprachanalyseprogramm gewählt, weil man ihm, wie der Eliza aus Pygmalion, immer besser "sprechen" beibringen kann.

Da es bei Gesprächen um etwas gehen muss, d.h. weil sie in einem bestimmten Kontext stattfinden müssen, wurde das Programm in einer zweistufigen Anordnung aufgebaut, wobei die erste Stufe aus dem Sprachanalysator und die zweite aus einem Skript besteht. Das Skript ist eine Reihe von Regeln, ähnlich denen, die man einem Schauspieler geben kann, der damit über ein bestimmtes Thema improvisieren soll. So könnte man ELIZA ein Skript geben, das es ihm ermöglicht, ein Gespräch über das Kochen von Eiern oder über die Verwaltung eines Girokontos zu führen, und so weiter. Jedes spezifische Skript ermöglichte es ELIZA also, eine bestimmte Gesprächsrolle zu spielen.

J. Weizenbaum, *Computer Power and Human Reason. From Judgement to Calculation*, San Francisco, W.H. Freeman & Company, 1976.

DOCTOR

“For my first experiment, I gave ELIZA a script designed to permit it to play (I should really say parody) the role of a Rogerian psychotherapist engaged in an initial interview with a patient. The Rogerian psychotherapist is relatively easy to imitate because much of his technique consists of drawing his patient out by reflecting the patient’s statements back to him.

The following conversation between a young lady and Eliza playing doctor illustrates both the Rogerian technique of encouraging a patient to keep talking and the operation of the computer program ELIZA. The first to “speak” is the young lady. The computer’s responses are printed entirely in capitals.”

J. Weizenbaum, *Computer Power and Human Reason. From Judgement to Calculation*, San Francisco, W.H. Freeman & Company, 1976.

DOCTOR

"Für mein erstes Experiment gab ich ELIZA ein Skript, das es ihm ermöglichen sollte, die Rolle eines Roger'schen Psychotherapeuten zu spielen (ich sollte eigentlich sagen zu parodieren), der ein Erstgespräch mit einem Patienten führt. Der Rogerianische Psychotherapeut ist relativ leicht zu imitieren, da ein Großteil seiner Technik darin besteht, den Patienten durch die Rückspiegelung seiner Aussagen in die Irre zu führen.

Das folgende Gespräch zwischen einer jungen Frau und Eliza, die Doktor spielt, veranschaulicht sowohl die Rogersche Technik, einen Patienten zum Weiterreden zu ermutigen, als auch die Funktionsweise des Computerprogramms ELIZA. Als erste "spricht" die junge Frau. Die Antworten des Computers sind komplett in Großbuchstaben geschrieben.

J. Weizenbaum, *Computer Power and Human Reason. From Judgement to Calculation*, San Francisco, W.H. Freeman & Company, 1976.

Ein "Gespräch" mit DOCTOR

“Men are all alike

IN WHAT WAY

They're always bugging about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

It's true. I am unhappy.

DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY

I need some help, that much seems certain.

WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP

Perhaps I could learn to get along with my mother.

TELL ME MORE ABOUT YOUR FAMILY

My mother takes care of me.

WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU

My father.

YOUR FATHER

You are like my father in some ways.”

Ein "Gespräch" mit DOCTOR

Männer sind alle gleich

IN WELCHER WEISE

Sie sind immer über irgendetwas genervt.

FÄLLT IHNEN EIN KONKRETES BEISPIEL EIN?

Nun, mein Freund hat mich dazu gebracht, hierher zu kommen.

DEIN FREUND HAT DICH HIERHER GEBRACHT

Er sagt, ich sei oft deprimiert.

ES TUT MIR LEID ZU HÖREN, DASS DU DEPRIMIERT BIST

Es ist wahr. Ich bin unglücklich.

**GLAUBEN SIE, DASS DER BESUCH HIER IHNEN HELFEN WIRD,
NICHT MEHR UNGLÜCKLICH ZU SEIN?**

Ich brauche Hilfe, so viel scheint sicher.

**WAS WÜRDEN ES FÜR SIE BEDEUTEN, WENN SIE HILFE
BEKÄMEN?**

Vielleicht könnte ich lernen, mit meiner Mutter zurechtzukommen.

ERZÄHLEN SIE MIR MEHR ÜBER IHRE FAMILIE

Meine Mutter kümmert sich um mich.

WER IN DEINER FAMILIE KÜMMERT SICH NOCH UM DICH?

Mein Vater.

DEIN VATER

Du bist in mancher Hinsicht wie mein Vater.

“The reckless anthropomorphization of the computer”

“I was startled to see how quickly and how very deeply people conversing with DOCTOR became emotionally involved with the computer and how unequivocally they anthropomorphized it. Once my secretary, who had watched me work on the program for many months and therefore surely knew it to be merely a computer program, started conversing with it. After only a few interchanges with it, she asked me to leave the room.

Another time, I suggested I might rig the system so that I could examine all conversations anyone had had with it, say, overnight. I was promptly bombarded with accusations that what I proposed amounted to spying on people’s most intimate thoughts.”

J. Weizenbaum, *Computer Power and Human Reason. From Judgement to Calculation*, San Francisco, W.H. Freeman & Company, 1976.

"Die rücksichtslose Anthropomorphisierung des Computers"

Ich war erstaunt zu sehen, wie schnell und wie tief die Menschen, die sich mit DOCTOR unterhielten, emotional mit dem Computer verbunden waren und wie eindeutig sie ihn vermenschlichten. Einmal begann meine Sekretärin, die mir monatelang bei der Arbeit an dem Programm zugesehen hatte und daher sicher wusste, dass es sich lediglich um ein Computerprogramm handelte, mit ihm zu sprechen. Nach nur wenigen Unterhaltungen mit ihm bat sie mich, den Raum zu verlassen.

Ein anderes Mal schlug ich vor, das System so zu manipulieren, dass ich alle Gespräche, die jemand mit dem Programm geführt hatte, z. B. über Nacht, überprüfen konnte. Prompt wurde ich mit Vorwürfen bombardiert, dass mein Vorschlag darauf hinausliefe, die intimsten Gedanken der Menschen auszuspionieren.

J. Weizenbaum, *Computer Power and Human Reason. From Judgement to Calculation*, San Francisco, W.H. Freeman & Company, 1976.

“Enormously exaggerated attributions”

“Another widespread, and to me surprising, reaction to the ELIZA program was the spread of a belief that it demonstrated a general solution to the problem of computer understanding of natural language. In my paper, I had tried to say that no general solution to that problem was possible, i.e., that language is understood only in contextual frameworks, that even these can be shared by people to only a limited extent, and that consequently even people are not embodiments of any such general solution.”

“This reaction to ELIZA showed me more vividly than anything I had seen hitherto the enormously exaggerated attributions an even well-educated audience is capable of making, even strives to make, to a technology it does not understand. Surely, I thought, decisions made by the general public about emergent technologies depend much more on what that public attributes to such technologies than on what they actually are or can and cannot do. If, as appeared to be the case, the public’s attributions are wildly misconceived, then public decisions are bound to be misguided and.”

J. Weizenbaum, *Computer Power and Human Reason. From Judgement to Calculation*, San Francisco, W.H. Freeman & Company, 1976.



Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423



Enorm übertriebene Zuschreibungen

Eine weitere weit verbreitete und für mich überraschende Reaktion auf das ELIZA-Programm war die Verbreitung des Glaubens, dass es eine allgemeine Lösung für das Problem des Computerverständnisses natürlicher Sprache aufzeigt. In meiner Arbeit hatte ich versucht zu sagen, dass es keine allgemeine Lösung für dieses Problem gibt, d.h. dass Sprache nur in kontextuellen Rahmen verstanden wird, dass selbst diese nur in begrenztem Maße von Menschen geteilt werden können und dass folglich auch Menschen keine Verkörperung einer solchen allgemeinen Lösung sind.

Diese Reaktion auf ELIZA zeigte mir deutlicher als alles, was ich bis dahin gesehen hatte, zu welcher enorm übertriebenen Zuschreibungen ein selbst gut ausgebildetes Publikum einer Technologie, die es nicht versteht, fähig ist, ja sogar anstrebt. Sicherlich, so dachte ich, hängen die Entscheidungen, die die Öffentlichkeit über neu entstehende Technologien trifft, viel mehr davon ab, was sie diesen Technologien zuschreibt, als davon, was sie tatsächlich sind oder was sie können und was nicht. Wenn, wie es den Anschein hat, die Zuschreibungen der Öffentlichkeit völlig falsch sind, dann sind die öffentlichen Entscheidungen zwangsläufig fehlgeleitet.

J. Weizenbaum, *Computer Power and Human Reason. From Judgement to Calculation*, San Francisco, W.H. Freeman & Company, 1976.



Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423



AI meets natural stupidity

“Wishful Mnemonics

A major source of simple-mindedness in AI programs is the use of mnemonics like "UNDERSTAND" or "GOAL" to refer to programs and data structures. This practice has been inherited from more traditional programming applications, in which it is liberating and enlightening to be able to **refer to program structures by their purposes.**”

“However, in AI, our programs to a great degree are problems rather than solutions. If a researcher tries to write an "understanding" program, it isn't because he has thought of a better way of implementing this well-understood task, but because he thinks he can come closer to writing the *first* implementation. If he calls the main loop of his program "UNDERSTAND", he is (until proven innocent) merely begging the question. He may mislead a lot of people, most prominently himself, and enrage a lot of others.”

D. McDermott, *AI Meets Natural Stupidity*, in «ACM SIGART Bulletin», 1976, n. 57, pp. 4-9.

KI trifft auf natürliche Dummheit

Wünschenswerte Mnemotechnik

Eine Hauptquelle für Einfältigkeit in KI-Programmen ist die Verwendung von Mnemonics wie "UNDERSTAND" oder "GOAL", um auf Programme und Datenstrukturen zu verweisen. Diese Praxis wurde von traditionelleren Programmieranwendungen übernommen, bei denen es befreiend und aufschlussreich ist, sich auf **Programmstrukturen durch ihren Zweck beziehen zu können**.

In der KI sind unsere Programme jedoch zu einem großen Teil eher Probleme als Lösungen. Wenn ein Forscher versucht, ein "verstehendes" Programm zu schreiben, dann nicht, weil ihm ein besserer Weg eingefallen ist, diese wohlverstandene Aufgabe zu implementieren, sondern weil er glaubt, er könne der ersten Implementierung näher kommen. Wenn er die Hauptschleife seines Programms "VERSTEHEN" nennt, stellt er (bis zum Beweis seiner Unschuld) die Frage nur in den Raum. Er könnte eine Menge Leute in die Irre führen, vor allem sich selbst, und eine Menge anderer verärgern.

D. McDermott, *AI Meets Natural Stupidity*, in «ACM SIGART Bulletin», 1976, n. 57, pp. 4-9.

The first-step fallacy

“Advances on a specific AI task are often described as “a first step” towards more general AI. The chessplaying computer Deep Blue was “was hailed as the first step of an AI revolution”. IBM described its Watson system as “a first step into cognitive systems, a new era of computing”. OpenAI’s GPT-3 language generator was called a “step toward general intelligence”.

Indeed, if people see a machine do something amazing, albeit in a narrow area, they often assume the field is that much further along toward general AI. The philosopher Hubert Dreyfus (using a term coined by Yehoshua Bar-Hillel) called this a “first-step fallacy.”

As Dreyfus characterized it, **“The first-step fallacy is the claim that, ever since our first work on computer intelligence we have been inching along a continuum at the end of which is AI so that any improvement in our programs no matter how trivial counts as progress.”**

Dreyfus quotes an analogy made by his brother, the engineer Stuart Dreyfus: **“It was like claiming that the first monkey that climbed a tree was making progress towards landing on the moon”.**

Melanie Mitchell, [*Why AI is Harder Than We Think*](#), 2021

Der Trugschluss des Faustschritts

Fortschritte bei einer bestimmten KI-Aufgabe werden oft als "erster Schritt" zu einer allgemeineren KI beschrieben. Der schachspielende Computer Deep Blue wurde "als erster Schritt einer KI-Revolution" gefeiert. IBM bezeichnete sein Watson-System als "ersten Schritt in Richtung kognitiver Systeme, einer neuen Ära des Computing". Der Sprachgenerator GPT-3 von OpenAI wurde als "Schritt in Richtung allgemeine Intelligenz" bezeichnet.

Wenn die Menschen sehen, dass eine Maschine etwas Erstaunliches leistet, wenn auch in einem engen Bereich, gehen sie oft davon aus, dass das Feld der allgemeinen KI schon viel weiter ist. Der Philosoph Hubert Dreyfus (unter Verwendung eines von Yehoshua Bar-Hillel geprägten Begriffs) bezeichnete dies als "first-step fallacy".

Dreyfus charakterisierte dies folgendermaßen: **Der Trugschluss des ersten Schritts ist die Behauptung, dass wir uns seit unserer ersten Arbeit an der Computerintelligenz auf einem Kontinuum bewegt haben, an dessen Ende die KI steht, so dass jede noch so triviale Verbesserung unserer Programme als Fortschritt gilt.**

Dreyfus zitiert eine Analogie seines Bruders, des Ingenieurs Stuart Dreyfus: **Das wäre so, als würde man behaupten, dass der erste Affe, der auf einen Baum kletterte, einen Fortschritt auf dem Weg zur Landung auf dem Mond machte.**

Melanie Mitchell, [*Why AI is Harder Than We Think*](#), 2021

Max Weber's theory of disenchantment

“the growing process of intellectualization and rationalization does not imply a growing understanding of the conditions under which we live. It means something quite different.

It is the knowledge or the conviction that if only we wished to understand them we could do so at any time.

It means that in principle, then, we are not ruled by mysterious, unpredictable forces, but that, on the contrary, we can in principle control everything by means of calculation. That in turn means the disenchantment of the world. Unlike the savage for whom such forces existed, we need no longer have recourse to magic in order to control the spirits or pray to them. Instead, technology and calculation achieve our ends. This is the primary meaning of the process of intellectualization.”

M. Weber, *[The Vocation Lectures: Science As A Vocation, Politics As A Vocation](#)*, ed. by D.S. Owen, T.B. Strong; transl. by R. Livingstone, 2004.

Max Webers Theorie der Desillusionierung

Der zunehmende Prozess der Intellektualisierung und Rationalisierung bedeutet nicht, dass das Verständnis für die Bedingungen, unter denen wir leben, wächst. Er bedeutet etwas ganz anderes.

Es ist das Wissen oder die Überzeugung, dass wir sie jederzeit verstehen könnten, wenn wir es nur wollten.

Es bedeutet also, dass wir im Prinzip nicht von geheimnisvollen, unvorhersehbaren Kräften beherrscht werden, sondern dass wir im Gegenteil im Prinzip alles durch Berechnung steuern können. Das wiederum bedeutet die Entzauberung der Welt. Anders als der Wilde, für den solche Kräfte existierten, müssen wir nicht mehr auf Magie zurückgreifen, um die Geister zu kontrollieren oder zu ihnen zu beten. Stattdessen erreichen wir unsere Ziele durch Technik und Berechnung. Dies ist die Hauptbedeutung des Prozesses der Intellektualisierung.

M. Weber, *[The Vocation Lectures: Science As A Vocation, Politics As A Vocation](#)*, ed. by D.S. Owen, T.B. Strong; transl. by R. Livingstone, 2004.

Enchanted determinism

“What makes contemporary **deep learning systems** interesting is their ambivalent position with respect to Weber’s larger thesis. They **certainly embody aspects of a disenchanted world in that they work to master or control new domains of social life through technical forms of calculation.** [...]

At the same time, these systems seem to violate the epistemology of disenchantment, the idea that there are no longer “mysterious” forces acting in the world. Paradoxically, when the disenchanted predictions and classifications of deep learning work as hoped, **we see a profusion of optimistic discourse that characterizes these systems as magical, appealing to mysterious forces and superhuman power.** [...] **It is a form of power without knowledge.”**

“**Enchanted determinism**”: “a discourse that presents deep learning techniques as magical, outside the scope of present scientific knowledge, yet also deterministic, in that deep learning systems can nonetheless detect patterns that give unprecedented access to people’s identities, emotions and social character. These systems become deterministic when they are deployed unilaterally in critical social areas, from healthcare to the criminal justice system, creating ever more granular distinctions, relations, and hierarchies that are outside of political or civic processes, with consequences that even their designers may not fully understand or control.”

A. Campolo, K. Crawford, *Enchanted Determinism: Power without Responsibility in Artificial Intelligence*, in «Engaging Science, Technology, and Society», 6 (2020), pp. 1-19.

Verzauberter Determinismus

"Was zeitgenössische **Deep-Learning-Systeme** interessant macht, ist ihre ambivalente Position in Bezug auf Webers größere These. Sie verkörpern sicherlich Aspekte einer entzauberten Welt, da sie daran arbeiten, neue Bereiche des sozialen Lebens durch technische Formen der Berechnung zu beherrschen oder zu kontrollieren. [...] Gleichzeitig scheinen diese Systeme die Erkenntnistheorie der Entzauberung zu verletzen, die Vorstellung, dass es keine "geheimnisvollen" Kräfte mehr gibt, die in der Welt wirken. Wenn die entzauberten Vorhersagen und Klassifizierungen des Deep Learning wie erhofft funktionieren, erleben wir paradoxerweise **eine Fülle optimistischer Diskurse, die diese Systeme als magisch bezeichnen und sich auf geheimnisvolle Kräfte und übermenschliche Macht berufen.** [...] Es ist eine Form der Macht ohne Wissen.

Verzauberter Determinismus: ein Diskurs, der Deep-Learning-Techniken als magisch darstellt, außerhalb des Rahmens der gegenwärtigen wissenschaftlichen Erkenntnisse, aber auch als deterministisch, da Deep-Learning-Systeme nichtsdestotrotz Muster erkennen können, die einen beispiellosen Zugang zu den Identitäten, Emotionen und dem sozialen Charakter von Menschen ermöglichen. Diese Systeme werden deterministisch, wenn sie einseitig in kritischen sozialen Bereichen eingesetzt werden, von der Gesundheitsfürsorge bis zum Strafrechtssystem, und immer feinere Unterscheidungen, Beziehungen und Hierarchien schaffen, die außerhalb politischer oder zivilgesellschaftlicher Prozesse liegen, mit Folgen, die selbst ihre Entwickler nicht vollständig verstehen oder kontrollieren können.

A. Campolo, K. Crawford, *Enchanted Determinism: Power without Responsibility in Artificial Intelligence*, in «Engaging Science, Technology, and Society», 6 (2020), pp. 1-19.

KI und die Alles im Whole Wide World Benchmark



Co-financed by the European Union
Connecting Europe Facility

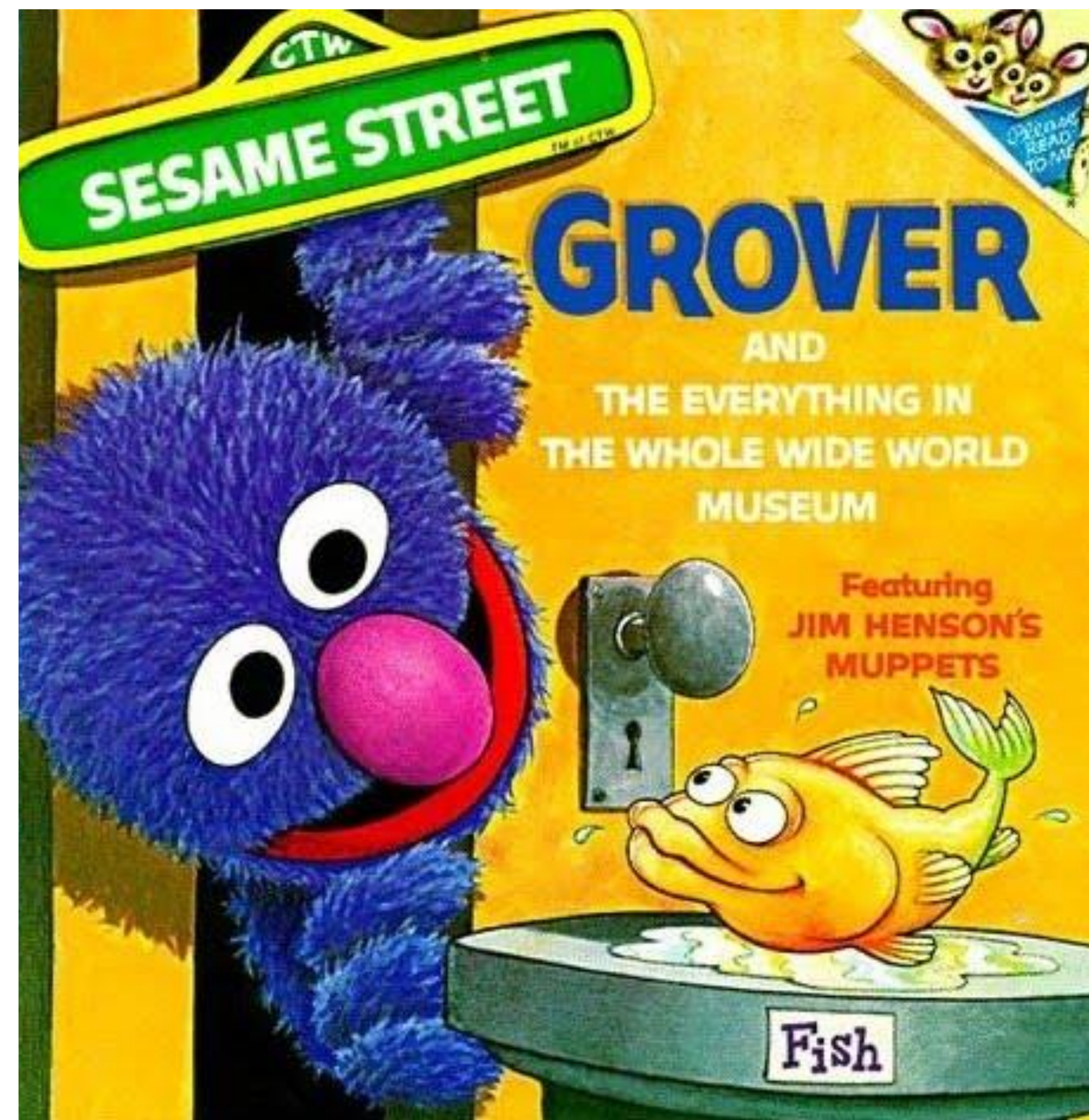
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423



In the 1974 Sesame Street children’s storybook *Grover and the Everything in the Whole Wide World Museum* [Stiles and Wilcox, 1974], the Muppet monster Grover visits a museum claiming to showcase “everything in the whole wide world”. Example objects representing certain categories fill each room. Several categories are arbitrary and subjective, including showrooms for “Things You Find On a Wall” and “The Things that Can Tickle You Room”. Some are oddly specific, such as “The Carrot Room”, while others unhelpfully vague like “The Tall Hall”. When he thinks that he has seen all that is there, Grover comes to a door that is labeled “Everything Else”. He opens the door, only to find himself in the outside world.

As a children’s story, Grover’s described situation is meant to be absurd. However, in this paper, we discuss how a similar faulty logic is inherent to recent trends in artificial intelligence (AI) — and specifically machine learning (ML) — evaluation, where many popular benchmarks rely on the same false assumptions inherent to the ridiculous “Everything in the Whole Wide World Museum” that Grover visits. In particular, we argue that benchmarks presented as measurements of progress towards general ability within vague tasks such as “visual understanding” or “language understanding” are as ineffective as the finite museum is at representing “everything in the whole wide world,” and for similar reasons — being inherently specific, finite and contextual.

Benchmarks like GLUE [Wang et al., 2019a] or ImageNet [Deng et al., 2009] are often elevated to become definitions of the essential common tasks to validate the performance of any given model. As a result, often the claims that are justified through these benchmark datasets extend far beyond the tasks they are initially designed for, and reach beyond even the initial ambitions for development. Despite a presentation and acceptance as markers of progress towards general-purpose capabilities, there are clear limitations of these benchmarks. In fact, the reality of their development, use and adoption indicates a *construct validity* issue, where the involved benchmarks — due to their instantiation in particular data, metrics and practice — cannot possibly capture anything representative of the claims to general applicability being made about them.



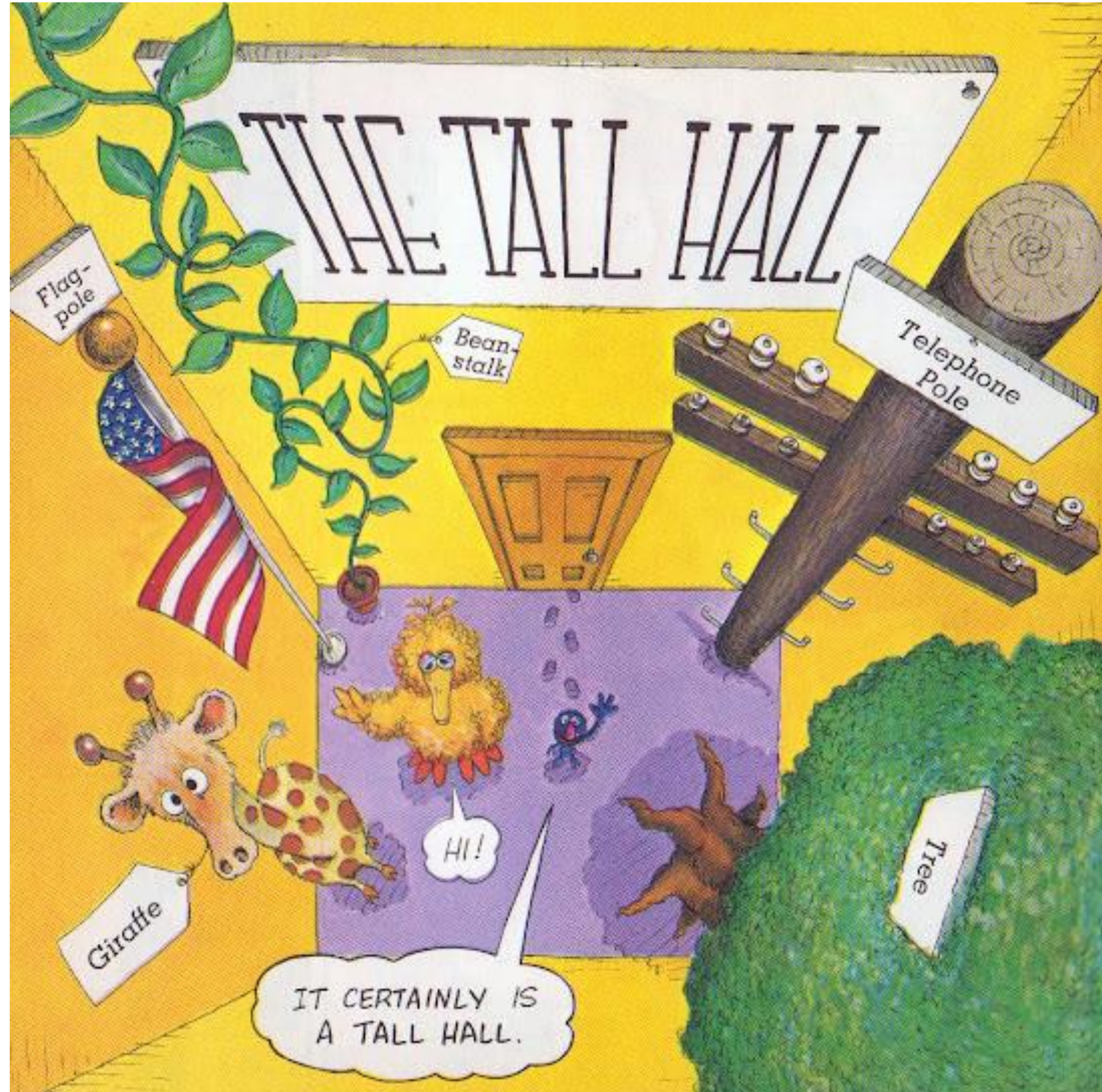
I.D. Raji, E.M. Bender, A. Paullada, E. Denton, A. Hanna, *AI and the Everything in the Whole Wide World Benchmark*, <https://arxiv.org/abs/2111.15366>













AI and the Everything in the Whole Wide World Benchmark

“Limits of Benchmarking General Capabilities”

- “The imagined artifact of the “general” benchmark does not actually exist. Real data is designed, subjective and limited in ways that necessitate a different framing from that of any claim to general knowledge or general-purpose capabilities. In fact, presenting any single dataset in this way is ultimately dangerous and deceptive, resulting in misguidance on task design and focus, underreporting of the many biases and subjective interpretations inherent in the data as well as enabling, through false presentations of performance, potential model misuse”
- “benchmarking is a limited approach to assess general model capabilities”

I.D. Raji, E.M. Bender, A. Paullada, E. Denton, A. Hanna, *AI and the Everything in the Whole Wide World Benchmark*, <https://arxiv.org/abs/2111.15366>

Die Benchmark-KI und alles in der weiten Welt

Die Grenzen des Benchmarking allgemeiner Fähigkeiten

- Das imaginäre Artefakt des "allgemeinen" Benchmarking existiert nicht wirklich. Echte Daten sind auf eine Art und Weise konzipiert, subjektiv und begrenzt, die einen anderen Rahmen als den eines Anspruchs auf allgemeines Wissen oder allgemeine Fähigkeiten erfordert. Tatsächlich ist die Darstellung eines einzelnen Datensatzes auf diese Weise letztlich gefährlich und irreführend, da sie zu einer Fehlleitung bei der Aufgabengestaltung und dem Fokus führt, die vielen Verzerrungen und subjektiven Interpretationen, die den Daten inhärent sind, nicht ausreichend berücksichtigt und durch die falsche Darstellung der Leistung einen potenziellen Modellmissbrauch ermöglicht.
- Benchmarking ist ein begrenzter Ansatz zur Bewertung der allgemeinen Modellfähigkeiten.

I.D. Raji, E.M. Bender, A. Paullada, E. Denton, A. Hanna, *AI and the Everything in the Whole Wide World Benchmark*, <https://arxiv.org/abs/2111.15366>

AI and the Everything in the Whole Wide World Benchmark

“The situation with Grover and the museum’s claims are clearly ridiculous—yet in machine learning, we follow the exact same logical fallacies to justify the elevation of a select number of benchmarks operating as general benchmarks for the field. However, there is no dataset that will be able to capture the full complexity of the details of existence, in the same way that there can be no museum to contain the full catalog of everything in the whole wide world. Open-world, universal and neutral datasets don’t exist, and current methods of benchmarking do not offer meaningful measures of general capabilities.”

“language understanding relies not only on linguistic competence but also world knowledge, commonsense reasoning, and the ability to model the interlocutor’s state of mind, none of which can be thoroughly tested through text-only tasks, such as GLUE. Several researchers have raised the need to establish effective physical and social grounding as part of the process of moving towards robust and effective natural language understanding, warning against text-only learning as a limited approach. Bender and Koller additionally mention the tendency of machine learning researchers to misinterpret certain benchmarks as capturing the model’s ability to decipher meaning in language, arguing that benchmarks need to be constructed with care if they are to show evidence of “understanding” as opposed to merely the ability to manipulate linguistic form sufficiently to pass the test.”

I.D. Raji, E.M. Bender, A. Paullada, E. Denton, A. Hanna, *AI and the Everything in the Whole Wide World Benchmark*, <https://arxiv.org/abs/2111.15366>

Die Benchmark-KI und alles in der weiten Welt

"Die Situation mit Grover und die Behauptungen des Museums sind eindeutig lächerlich - doch im Bereich des maschinellen Lernens folgen wir genau denselben logischen Irrtümern, um die Erhebung einiger ausgewählter Benchmarks zu rechtfertigen, die als allgemeine Benchmarks für den Bereich gelten. Es gibt jedoch keinen Datensatz, der in der Lage wäre, die gesamte Komplexität der Details der Existenz zu erfassen, so wie es auch kein Museum geben kann, das den vollständigen Katalog aller Dinge in der weiten Welt enthält. Es gibt keine weltoffenen, universellen und neutralen Datensätze, und die derzeitigen Methoden des Benchmarking bieten keine aussagekräftigen Maßstäbe für allgemeine Fähigkeiten.

Das Sprachverständnis beruht nicht nur auf sprachlicher Kompetenz, sondern auch auf Weltwissen, logischem Denken und der Fähigkeit, den Gemütszustand des Gesprächspartners zu modellieren, was sich mit reinen Textaufgaben, z. B. GLUE, nicht gründlich testen lässt. Mehrere Forscher haben auf die Notwendigkeit hingewiesen, auf dem Weg zu einem robusten und effektiven natürlichen Sprachverständnis eine effektive physische und soziale Basis zu schaffen, und vor dem reinen Textlernen als einem begrenzten Ansatz gewarnt. Bender und Koller weisen außerdem auf die Tendenz von Forschern im Bereich des maschinellen Lernens hin, bestimmte Benchmarks so zu interpretieren, dass sie die Fähigkeit des Modells zur Entschlüsselung von Bedeutungen in der Sprache erfassen, und argumentieren, dass Benchmarks sorgfältig konstruiert werden müssen, wenn sie den Nachweis für "Verständnis" erbringen sollen, im Gegensatz zu der bloßen Fähigkeit, sprachliche Formen ausreichend zu manipulieren, um den Test zu bestehen.

I.D. Raji, E.M. Bender, A. Paullada, E. Denton, A. Hanna, *AI and the Everything in the Whole Wide World Benchmark*, <https://arxiv.org/abs/2111.15366>

Vielen Dank. Haben Sie noch Fragen?

daniela.tafani@unibo.it