MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

University of Bologna

Computational Ethics

Daniela Tafani

2022/2023 – Second Semester

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

4 – Learning material

AI and magical thinking

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 as technology and “AI” as speech act

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

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

The problem of “trustworthy AI”

“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

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

“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

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

The animation of the inanimate

- 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 (forthcoming)

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

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.

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

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.

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

A “conversation” with 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.”

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

“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 accusa tions 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.

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

“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

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.

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

The fist-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

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

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.

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

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.

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 and the Everything in the Whole Wide World Benchmark

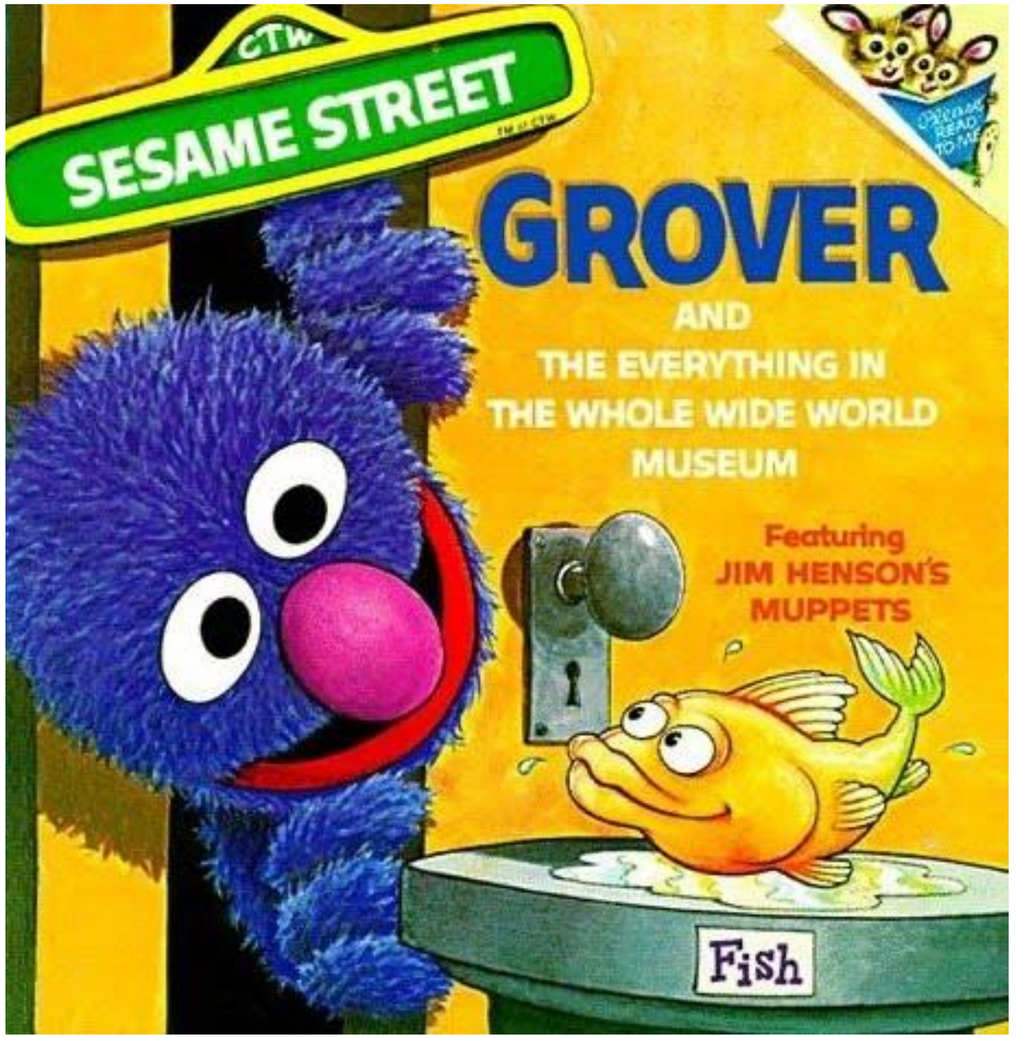
Co-financed by the European Union
Connecting Europe Facility

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

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

Co-financed by the European Union
Connecting Europe Facility

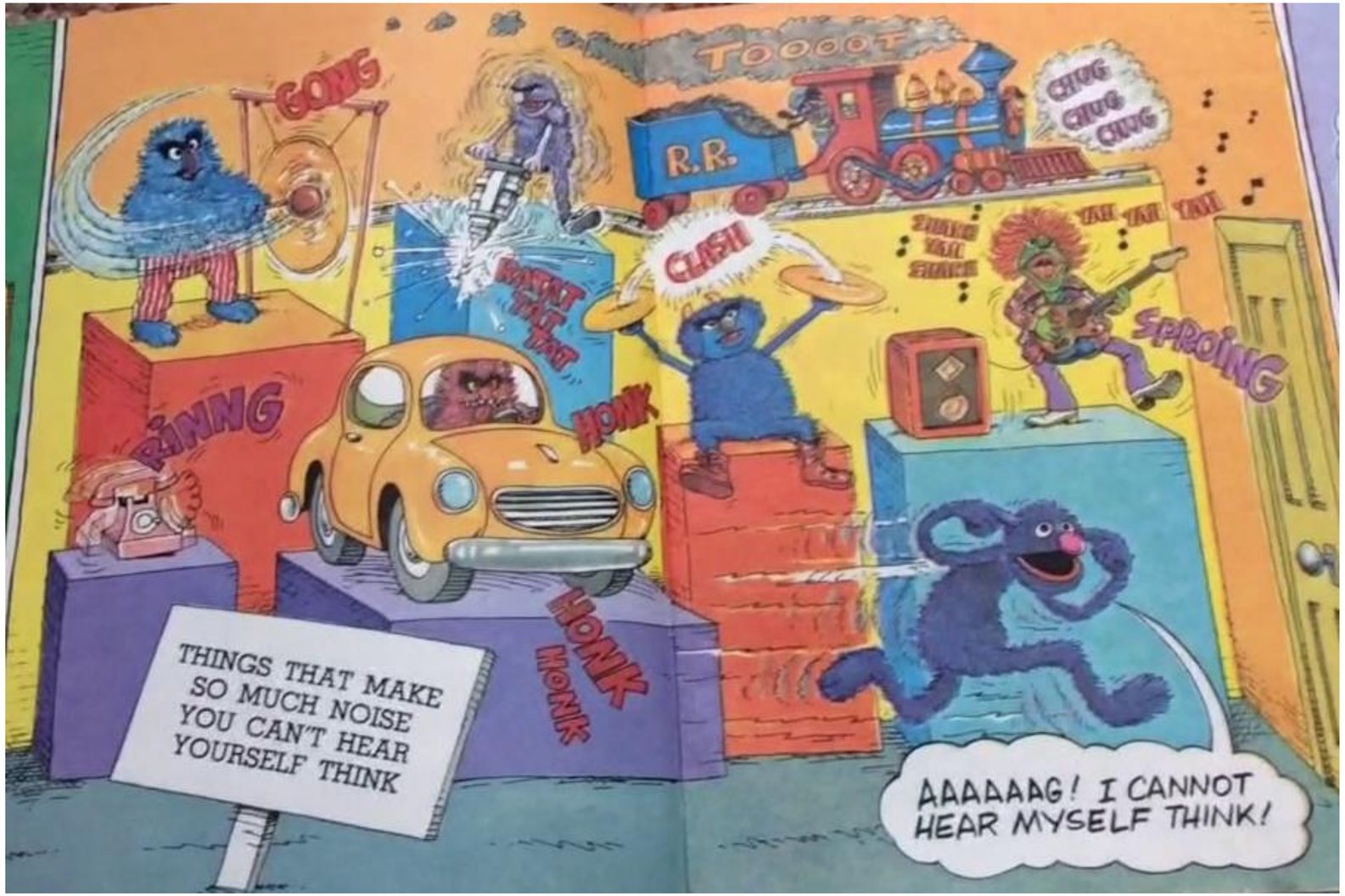
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

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

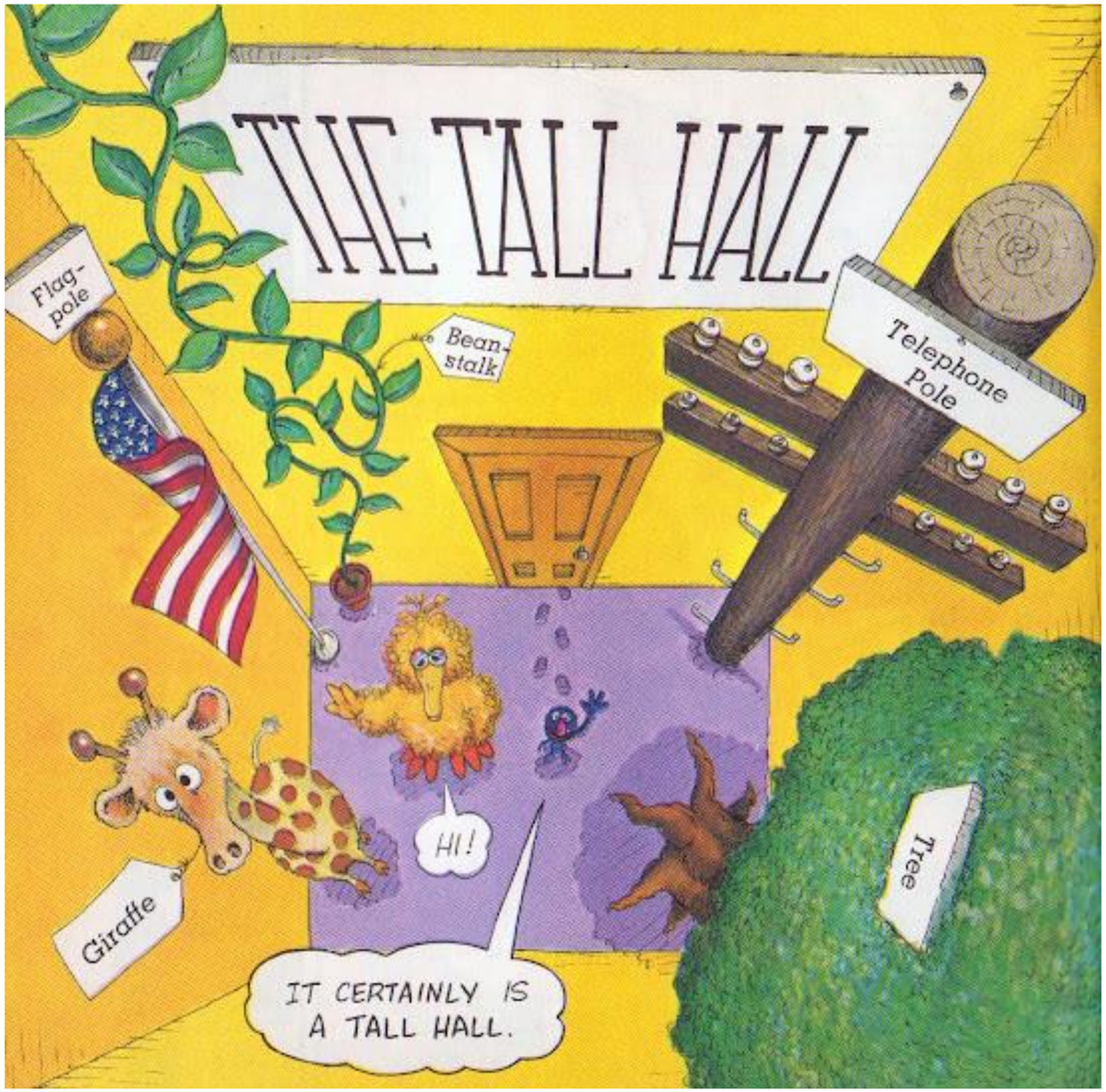
Co-financed by the European Union
Connecting Europe Facility

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



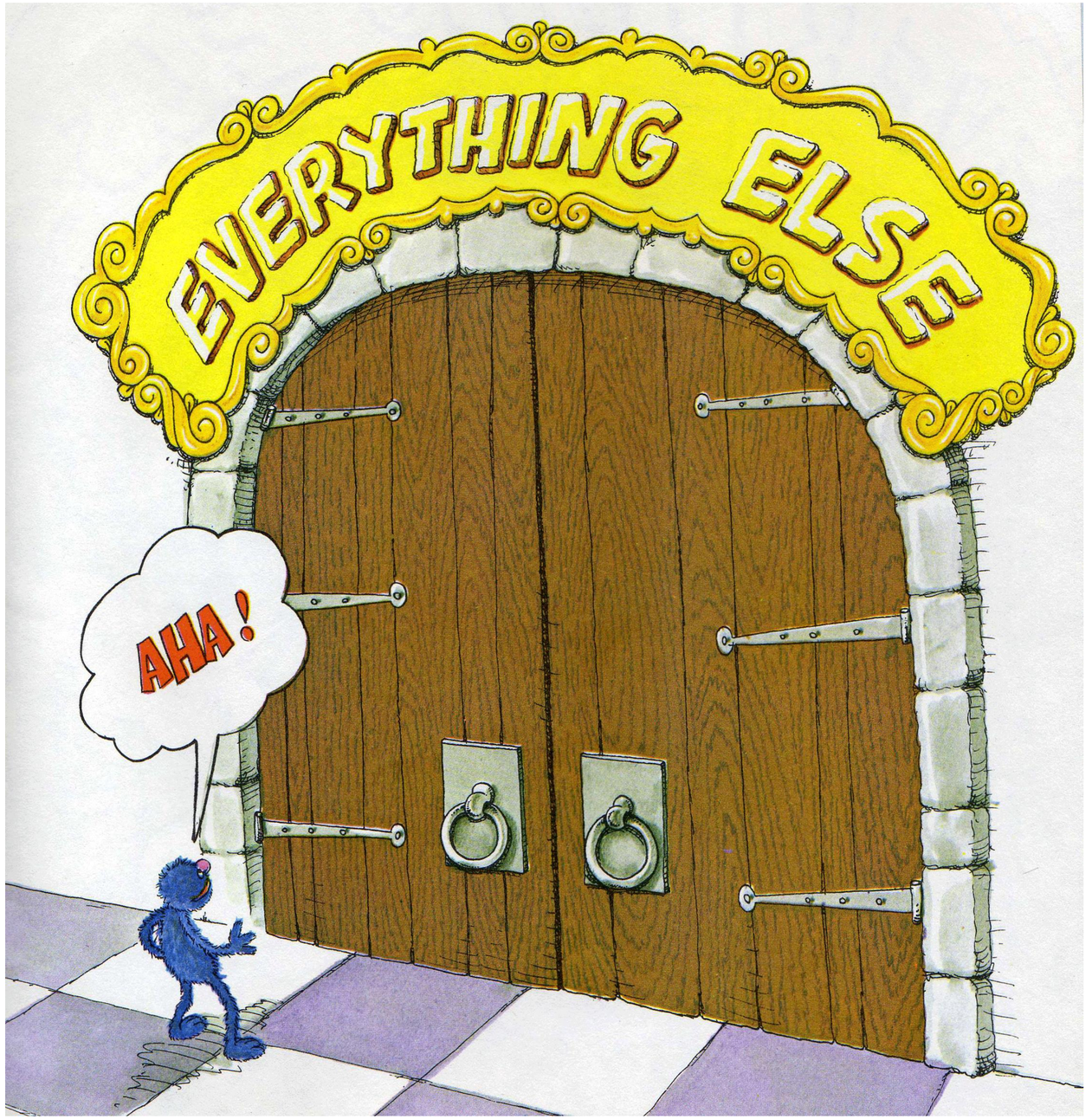
Co-financed by the European Union
Connecting Europe Facility

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



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



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 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

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 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

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

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Thank you. Any questions?

daniela.tafani@unibo.it

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