ABDUCTION IN LLMS
THE ISSUE OF REAL WORLD NAVIGATION

Mireille Hildebrandt, FBA
PI **COHUBICOL** ERC ADG project

Preliminary remarks

The person, the scientist, the lawyer

1. Abductive action in day2day navigation (surviving and flourishing)
   - embodied action, perception as informed by action-potential
   - double contingency: navigating both brute and institutional facts
2. Abductive action in science (exploratory and explanatory)
   - not every abduction makes sense
   - but any abduction could make sense
   - the context of justification inspires the context of discovery
3. Abductive action in law (judgement and judgment)
   - Dworkin's right answer thesis
   - in light of his focus on rule-based discretion

Abductive action in law

- The context of justification restricts the decisional space of the court
  - A judge may be inclined to convict me because she does not like me, or due to a bad breakfast (naïve legal realism)
  - She will nevertheless have to provide a ‘valid’ reason for her decision
  - The ‘legal syllogism’ is the justification not the method for the decision
    - Legal norm (major)
    - Facts of the case (minor)
    - Judgment (conclusion)
- The context of discovery and justification interact
  - In law they do so in a very specific way

The LLM system

- Abductive action? agents (different level of agency)
- Abductive behaviour? (things, including those with mindless agency)

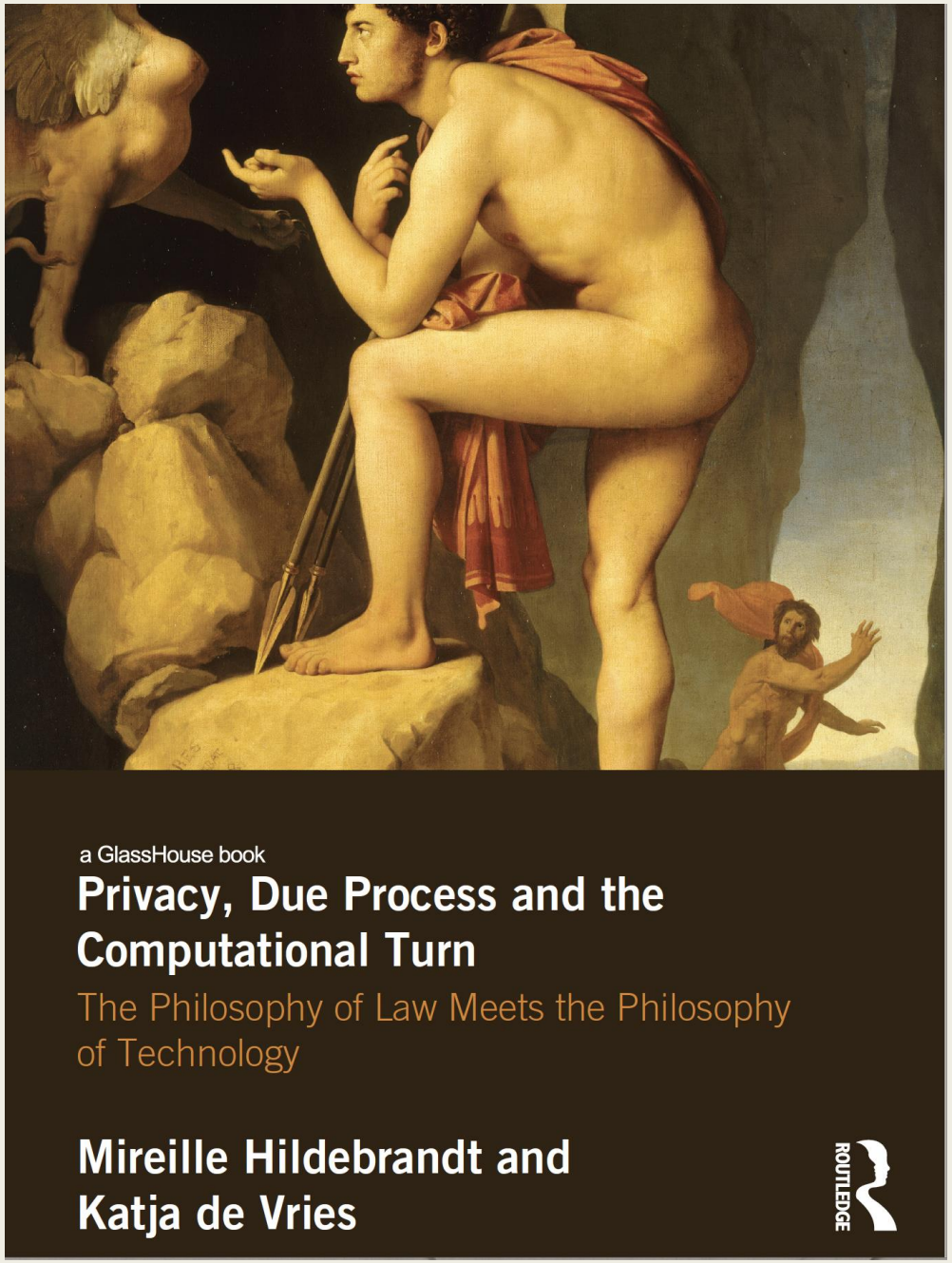
Abduction of an LLM
DALL E 2.6.23
[trouble with embodiment?]

What's next?

- Introduction
- The ML pipeline: proxies and alignment
- The training data: language behaviours
- Spoken and written speech acts
- Reasoning, inductive inference and abduction
- Real world and real life
- Person, scientists and judge: on the nature of abductive inferencing

What's next?

- Introduction
- The ML pipeline: proxies and alignment
- The training data: language behaviours
- Spoken and written speech acts
- Reasoning, inductive inference and abduction
- Real world and real life
- Person, scientists and judge: on the nature of abductive inferencing



3

Abducing personal data, destroying privacy

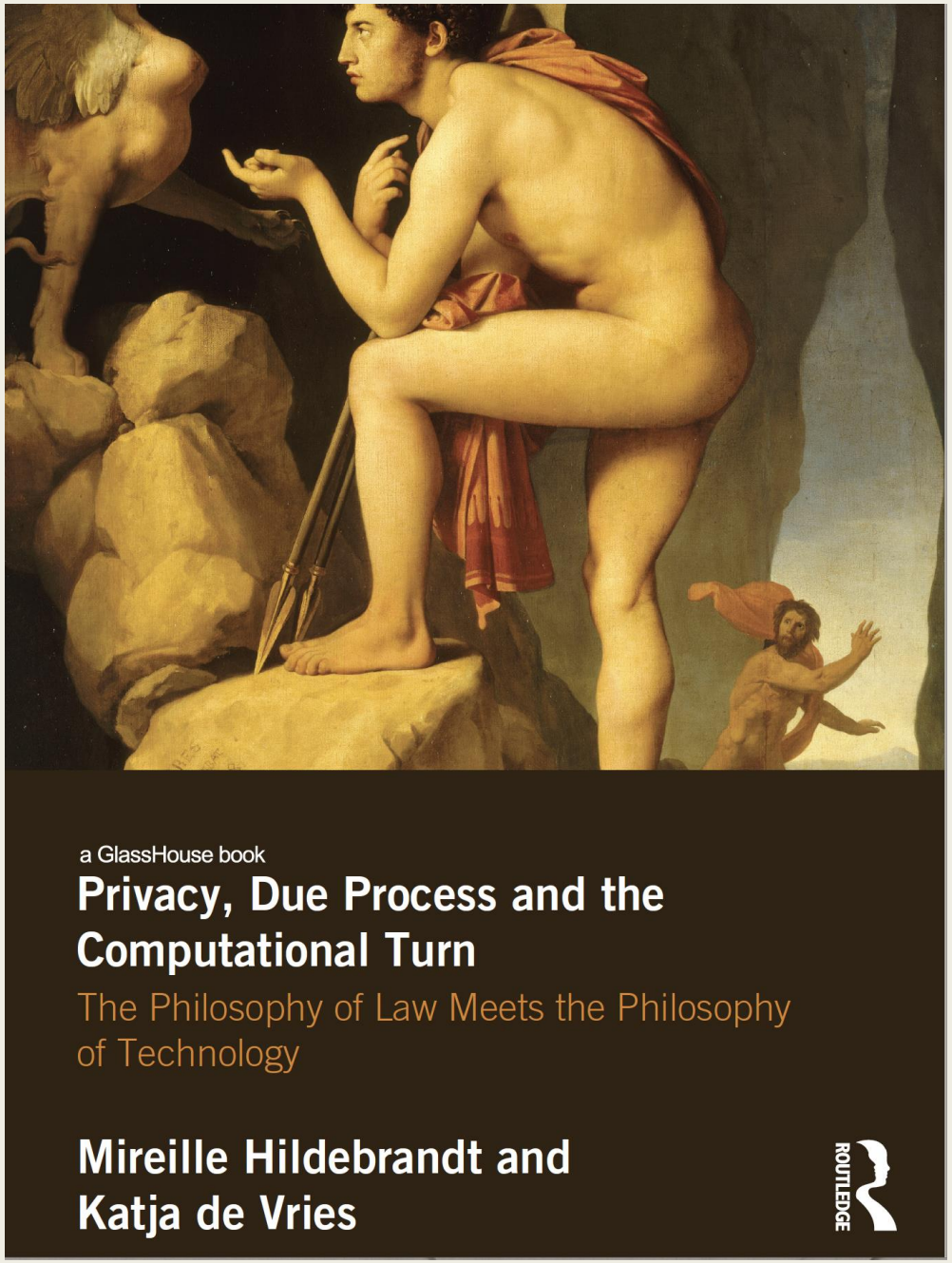
Diagnosing profiles through artefactual mediators

Lorenzo Magnani

Knowledge as a duty and its limitations

One of the aims of this chapter is to convince readers that knowledge has to become a duty in our technological world.[1] Thus far in my attempt to do so, I have to combine ethics, epistemology and cognitive science. An important issue arises from the fact that technological advances[2] have given greater value to external things – both natural and artificial – and while this may seem to

About the moral duty to ***provide*** knowledge and ***inform oneself*** when engaging in actions that may impact others



9

Profile transparency by design? Re-enabling double contingency

Mireille Hildebrandt

Introduction

The technologies of machine learning render us transparent in a rather counterintuitive manner. We become *transparent* in the sense that the profiling software looks straight *through us* to 'what we are like', instead of making transparent 'what or who we are'. This reminds me of a cartoon that shows a couple, sitting up in bed – after the act – confronted with a voice-over that proclaims: 'I'm glad you enjoyed that. People who like that technique also enjoyed these other sexual techniques: …'.[1] It is interesting to note that the couple – who may have felt they had just had a unique experience – is brought down to earth with a reminder of the repetitive nature of human interaction. They are reduced to being like many others and invited to explore the consolidated repertoire of those who are like them. In machine learning

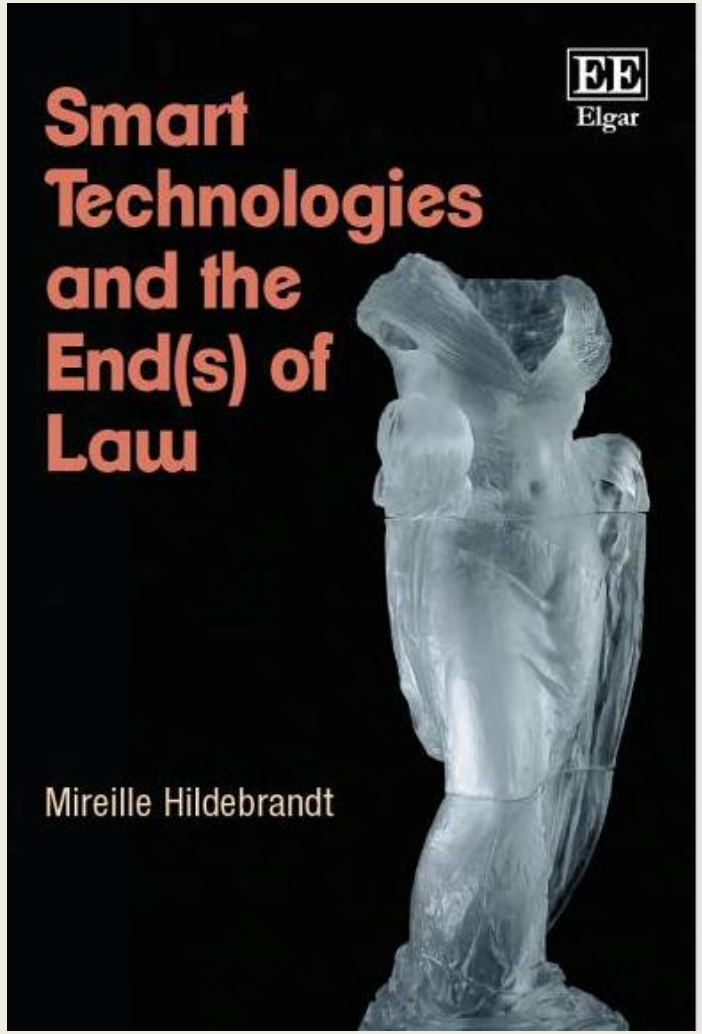
Anticipating how one is anticipated

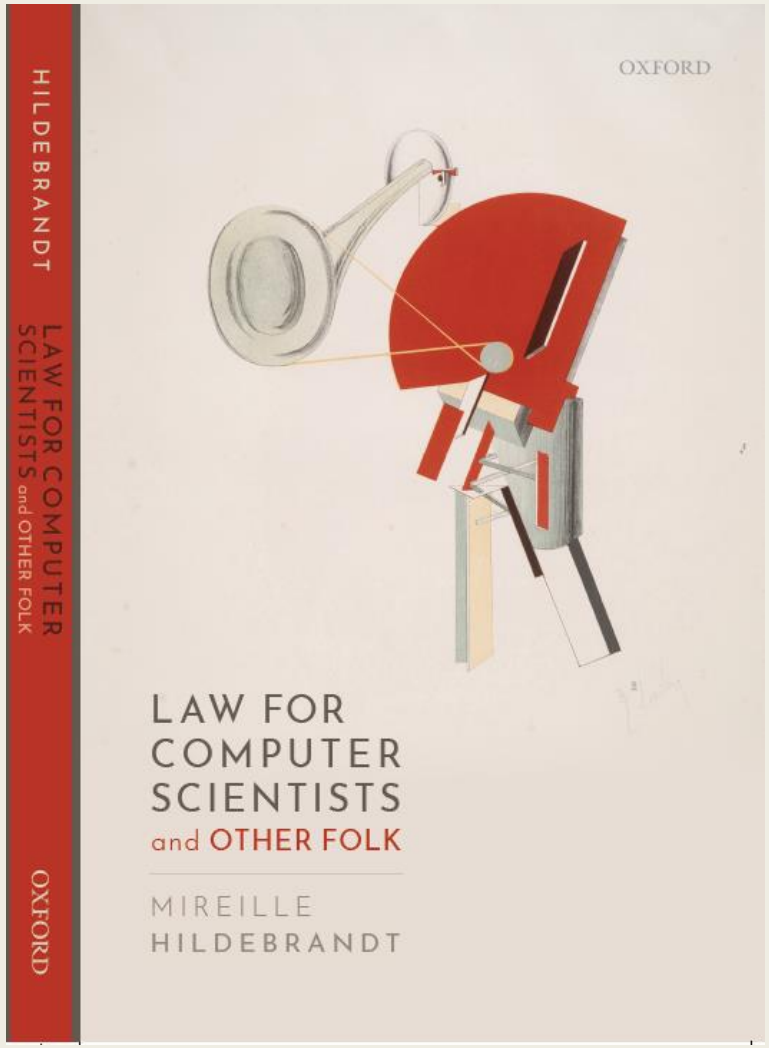
2015

- Living with systems that anticipate us
- Mindless agency (*ChatGPT avant la lettre*)

- Big data spaces (*EU strategy avant la lettre*)

- How does it affect our shared world?
  - and the role and the rule of law

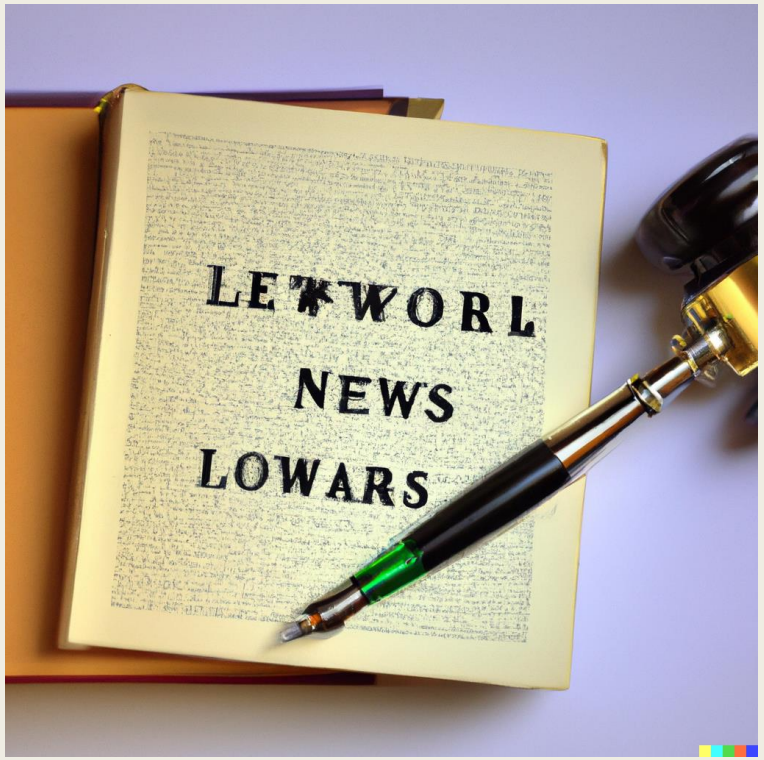




2020

- My background: law, philosophy of technology
- Chair at Computer Science Department @Radboud University
- My research focus:
  implications of 'AI' for law and the rule of law
- Research Chair at Vrije Universiteit Brussel
  - At the Faculty of Law and Criminology

2025?



DALL E for (31 May 2023):

- A New Hermeneutics for Computational Law

GPT4 gives us this: https://chat.openai.com/share/33beda32-7e75-4790-9c9a-61f35ee4a927

It would be nice if all of the data which sociologists require could be enumerated because then we could run them through IBM machines and draw charts as the economists do. However, not everything that can be counted counts, and not everything that counts can be counted

– William Cameron, *Informal Sociology* (1963)

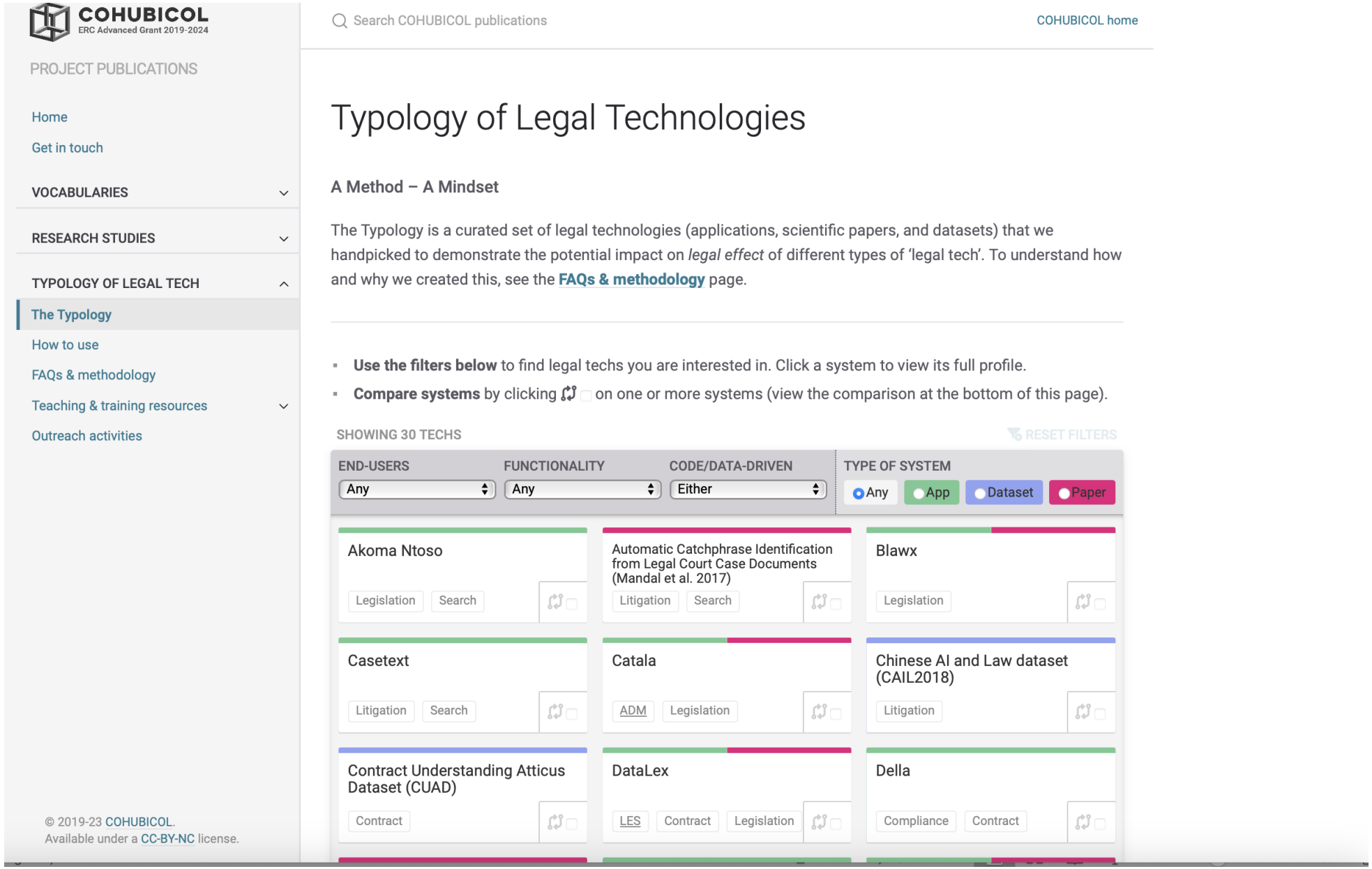
TYPOLOGY OF LEGAL TECHNOLOGIES

Typology of Legal

Latest news

Hildebrandt delivers keynote at Ninth International MBR Conference (Rome, 8 June 2023)

31 May 2023 ▪ 0 comments


COHUBICOL
ERC Advanced Grant 2019-2024
PROJECT PUBLICATIONS
Home
Get in touch
VOCABULARIES
RESEARCH STUDIES
TYPOLOGY OF LEGAL TECH
The Typology
How to use
FAQs & methodology
Teaching & training resources
Outreach activities
© 2019-23 COHUBICOL.
Available under a CC-BY-NC license.
Search COHUBICOL publications
COHUBICOL home
Typology of Legal Technologies
A Method – A Mindset
The Typology is a curated set of legal technologies (applications, scientific papers, and datasets) that we handpicked to demonstrate the potential impact on legal effect of different types of 'legal tech'. To understand how and why we created this, see the FAQs & methodology page.
Use the filters below to find legal techs you are interested in. Click a system to view its full profile.
Compare systems by clicking on one or more systems (view the comparison at the bottom of this page).
SHOWING 30 TECHS
RESET FILTERS
END-USERS
Any
FUNCTIONALITY
Any
CODE/DATA-DRIVEN
Either
TYPE OF SYSTEM
Any
App
Dataset
Paper
Akoma Ntoso
Legislation
Search
Automatic Catchphrase Identification from Legal Court Case Documents (Mandal et al. 2017)
Litigation
Search
Blawx
Legislation
Casetext
Litigation
Search
Catala
ADM
Legislation
Chinese AI and Law dataset (CAIL2018)
Litigation
Contract Understanding Atticus Dataset (CUAD)
Contract
DataLex
LES
Contract
Legislation
Della
Compliance
Contract

In the context of the ERC ADG we investigate:

- **claims made on behalf of** AI systems in law
- **the substantiation of such claims**
  - Mathematical verification, empirical validation, certification
  - Impact on the domain: gaps between requirements and specifications
  - Real-world impact (gap between specification and real-world goal)

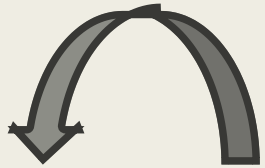
What's next?

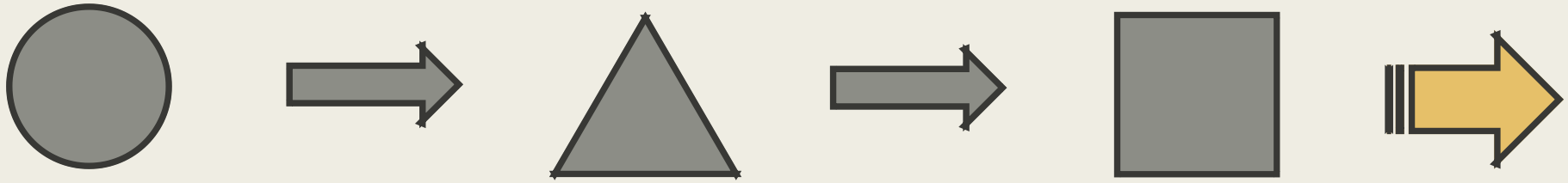
- Introduction
- The ML pipeline: proxies and alignment
- The training data: language behaviours
- Spoken and written speech acts
- Reasoning, inductive inference and abduction
- Abductive moments in LLM development and deployment
- Real world and real life
  - The plenitude of real world flux (B.C. Smith)
  - 'Intentionality' (Brentano) and 'Conatus' (Spinoza)
  - Two types of consequentialism

ML output-testing:
- **Accuracy**
- **Precision**
- **Recall**

validation

verification



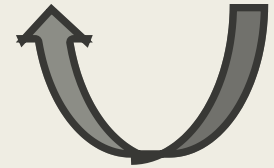


C O D E

Real World Goals
‘do justice’

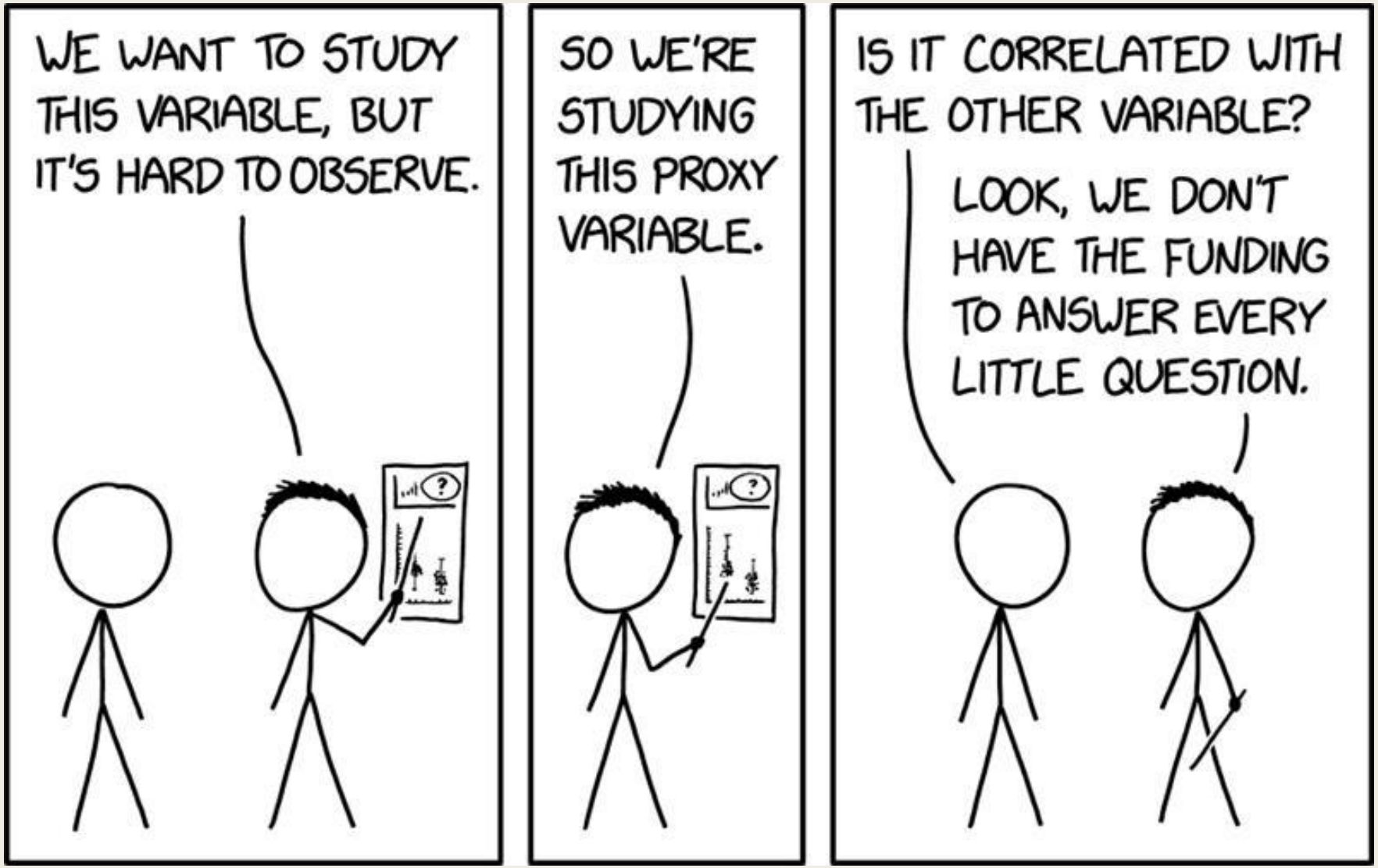
Requirements
‘like cases treated alike’

Specifications
‘select relevant features’
‘train an LLM on relevant case law’

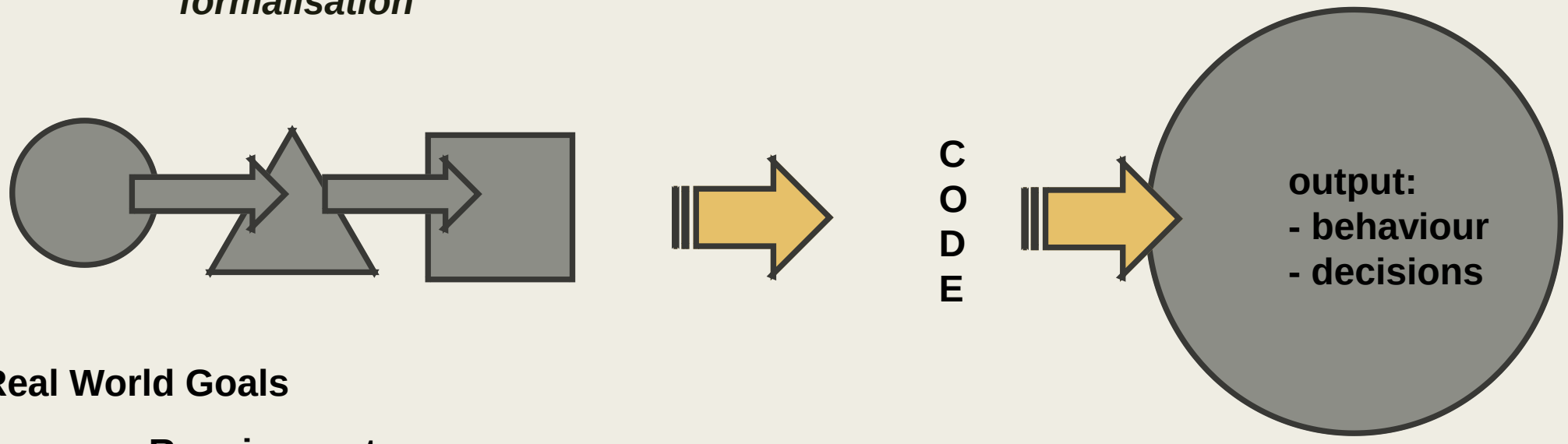


verification

validation


WE WANT TO STUDY THIS VARIABLE, BUT IT'S HARD TO OBSERVE.
SO WE'RE STUDYING THIS PROXY VARIABLE.
IS IT CORRELATED WITH THE OTHER VARIABLE?
LOOK, WE DON'T HAVE THE FUNDING TO ANSWER EVERY LITTLE QUESTION.

formalisation



Real World Goals

Requirements

Specifications

empirical validation – NOT about performance metrics

REVIEW article
Front. Artif. Intell., 28 April 2022
Sec. AI for Human Learning and Behavior Change
Volume 5 - 2022 | https://doi.org/10.3389/frai.2022.789076

The Issue of Proxies and Choice Architectures. Why EU Law Matters for Recommender Systems

Mireille Hildebrandt[1,2*]

[1] Institute of Computing and Information Sciences (iCIS), Science Faculty, Radboud University, Nijmegen, Netherlands
[2] Research Group Law Science Technology & Society (LSTS), Faculty of Law and Criminology, Vrije Universiteit Brussel, Brussels, Belgium

Recommendations are meant to increase sales or ad revenue, as these are the first priority of those who pay for them. As recommender systems match their recommendations with inferred preferences, we should not be surprised if the algorithm optimizes for lucrative preferences and thus co-produces the preferences they mine. This relates to the well-known problems of feedback loops, filter bubbles, and echo chambers. In this article, I discuss the implications of the fact that computing systems necessarily work with proxies when inferring recommendations and raise a number of questions about whether recommender systems actually do what they are claimed to do, while also analysing the often-perverse economic incentive structures that have a major impact on relevant design decisions. Finally, I will explain how the choice architectures for data controllers and providers of AI systems as foreseen in the EU's General Data Protection Regulation (GDPR), the proposed EU Digital Services Act (DSA) and the proposed EU AI Act will help to break through various vicious circles, by constraining how people may be targeted (GDPR, DSA) and by requiring documented evidence of the robustness, resilience, reliability, and the responsible design and deployment of high-risk recommender systems (AI Act).

TABLE OF CONTENTS

Paper

Ground-Truthing in the European Health Data Space

In **Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5 HEALTHINF: BIOSTEC,** 15-22, 2023 , Lisbon, Portugal

Ground-Truthing in the European Health Data Space

Mireille Hildebrandt
Vrije Universiteit Brussel, Belgium

Keywords: European Health Data Space, Ground Truth, Proxy, Interactive Machine Learning, Health Data, Alphafold, Large Language Models, Law.

Abstract: In this position paper I discuss the use of health-related training data for medical research, in light of the European Health Data Space. If such data is deployed as a proxy for 'the truth on the ground', we need to address the issue of proxies. Ground truth in machine learning is the pragmatic stand-in or proxy for whatever is considered to be the case or should be the case. Developing a ground truth dataset requires curation, i.e. a number of translations, constructions and cleansing. What if the resulting proxies misrepresent what they stand for and what if the imposed interoperability of health data across the EU affects the quality of the data and/or their relationship to what they stand for? I argue that ground-truthing is an act rather than a given, that this act is key to machine learning and assert that this act can have potentially fatal implications for the reliability of the output. Deciding on the ground truth is what philosophers may call *a speech act with performative effects*. Emphasising these effects will allow us to better address the constructive nature of the datasets used in medical informatics and should help the EU legislature to take a precautionary approach to medical informatics.

1 INTRODUCTION

In this position paper, I take issue with the productive assumptions of machine learning in the context of health data research. The focus is on the construction of training datasets that function as ground truth in supervised learning or otherwise as a proxy for (part of) the real world in unsupervised and reinforcement learning. I highlight the need to explicitly acknowledge that any computational ground truth is at most an approximation whose match with the real world depends on myriad design decisions that are part of the collection and curation of training data. Having discussed this point, I turn to the secondary use of health data as foreseen in the proposed Regulation on the European Health Data Space, EU legislature to better understand what cross-border aggregates of health data can and cannot achieve.

2 THE CONSTRUCTIVE AND/OR APPROXIMATE NATURE OF GROUND TRUTH

This short paper is indebted to the work of Cabitza, more precisely Cabitza et al. (2020, which I reviewed)[1], and my work in the context of AI in law, for instance Hildebrandt (2023) and law for AI, for instance Hildebrandt (2020, 2021, 2023). Establishing ground truth is a *conditio sine qua non* for supervised learning. Getting it wrong will result

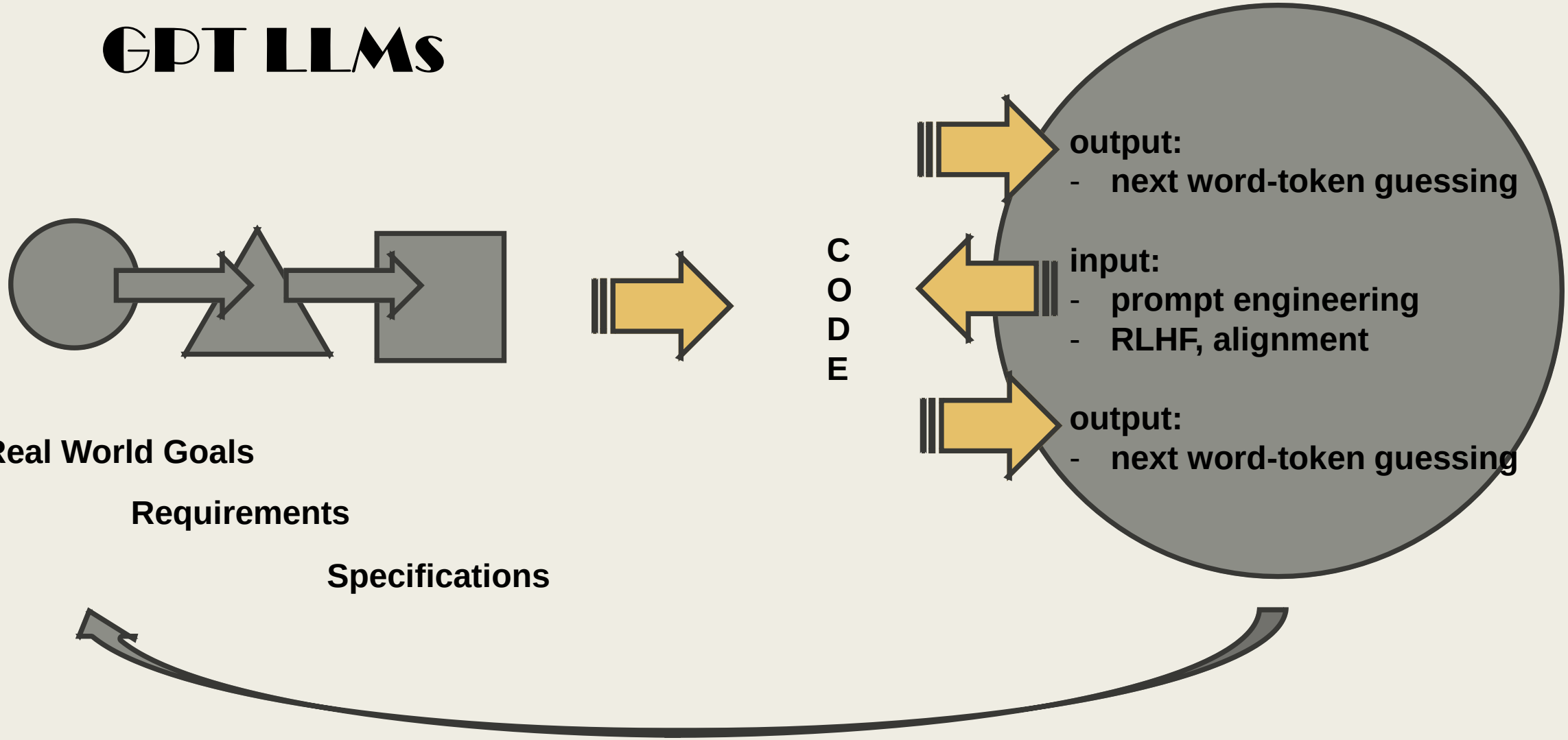
Author: Mireille Hildebrandt

Affiliation: Vrije Universiteit Brussel, Belgium

Keyword(s): European Health Data Space, Ground Truth, Proxy, Interactive Machine Learning, Health Data, Alphafold, Large Language Models, Law.

Abstract: In this position paper I discuss the use of health-related training data for medical research, in light of the European Health Data Space. If such data is deployed as a proxy for 'the truth on the ground', we need to address the issue of proxies. Ground truth in machine learning is the pragmatic stand-in or proxy for whatever is considered to be the case or should be the case. Developing a ground truth dataset requires curation, i.e. a number of translations, constructions and cleansing. What if the resulting proxies misrepresent what they stand for and what if the imposed interoperability of health data across the EU affects the quality of the data and/or their relationship to what they stand for? I argue that ground-truthing is an act rather than a given, that this act is key to machine learning and assert that this act can have potentially fatal implications for the reliability of the output. Deciding on the ground truth is what philosophers may call a speech act with performative (More)

GPT LLMs


output:
- next word-token guessing
input:
- prompt engineering
- RLHF, alignment
output:
- next word-token guessing
CODE
Real World Goals
Requirements
Specifications
validation

- What matters is not computable
- It can, however, be made computable
- This can always be done in different ways
- And those differences matter

- Key issues here are the selection and construction of the proxies
  - Training data, feature selection, hypothesis space, goals
- And, in the case of RLHF, the prompts provided to achieve alignment

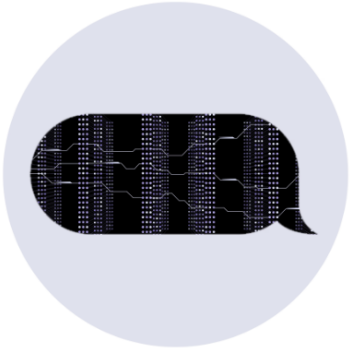
What's next?

- Introduction
- The ML pipeline: proxies and alignment
- The training data: language behaviours
- Spoken and written speech acts
- Reasoning, inductive inference and abduction
- Real world and real life
- Person, scientists and judge: on the nature of abductive inferencing

GPT training data

- ML, including LLMs, means:
  - detecting ‘relevant’ patterns in the distribution of ‘relevant’ data
- The assumption is:
  - past/current distributions are equivalent to future distributions
- In the real world that is not the case:
  - data is a proxy, the translation is both productive and defective, *and:*
    - if you use a measure as a target, it ceases to be a good measure (Goodhart effect)
    - the best way to predict the future is to create it (Gabor)
    - if machines define a situation as real, it may be real in its consequences (Thomas Merton Hildebrandt)

The Washington Post
Democracy Dies in Darkness



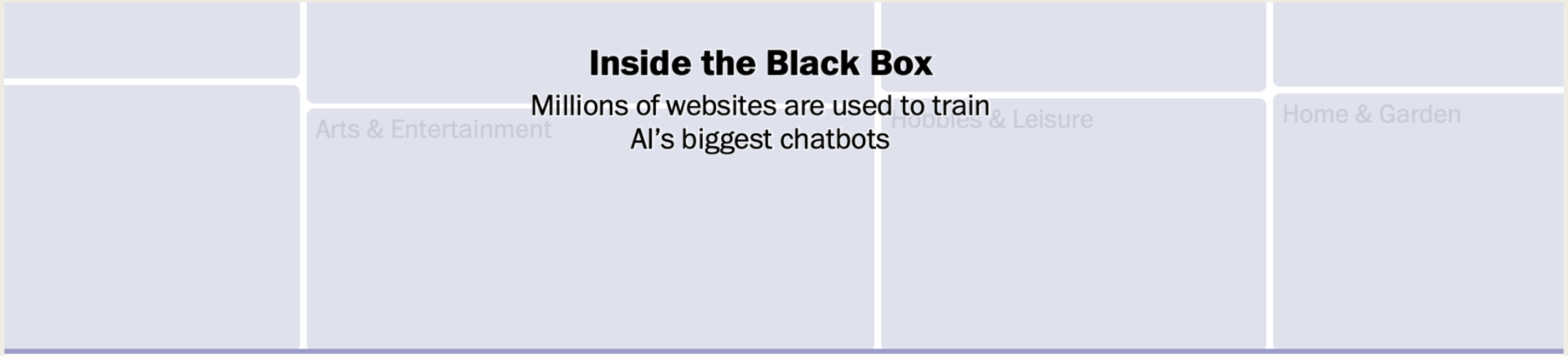
EXCLUSIVE

Inside the secret list of websites that make AI like ChatGPT sound smart

By Kevin Schaul, Szu Yu Chen and Nitasha Tiku

April 19 at 6:00 a.m.

AI chatbots have exploded in popularity over the past four months,



To look inside this black box, we analyzed Google's C4 data set, a massive snapshot of the contents of 15 million websites that have been used to instruct some high-profile English-language AIs, called large language models, including Google's T5 and Facebook's LLaMA. (OpenAI does not disclose what datasets it uses to train the models backing its popular chatbot, ChatGPT)

The Post worked with researchers at the Allen Institute for AI on this investigation and categorized the websites using data from Similarweb, a web analytics company. About a third of the websites could not be

LLM training data

- Ricoeur:
  - Langue (a language system) and Parole (language use)
- Wittgenstein:
  - Language usage is rule-based, but rules don't interpret themselves
- Kant:
  - Between a rule and its applicability there is a chasm
- Ricoeur:
  - Combining semiology with speech act theory

LLM training data

- Semiology ▻ language usage connects:
  - Intralinguistic references (signs) with
  - Extralinguistic references (real world phenomena)

- Speech act theory ▻ language use is a form of acting:
  - Locutionary speech acts: meant to describe
  - Illocutionary speech acts: performatives that constitute what they refer to
  - Perlocutionary speech acts: meant to influence, convince, steer

LLM training data

- Language use assumes (extra-linguistic reference):
  - Deictic or ostensive reference (Umwelt) pointing *at*
  - Non-deictic reference (Welt) pointing *out*

- Language use refers to (combining extra- and intra-linguistic reference):
  - Brute facts: a tree, the sun, a door
  - Institutional facts: doors, daylight, schools, marriage
  - Note the difference is relative

LLM training data

- LLMs are trained on behavioural data
  - Non-deictic intralinguistic reference
  - No intentionality (Brentano)
  - No access to a world outside the data (BC Smith)

- LLM's induction/abduction parasitises on our language behaviours
  - It thus lags behind (cannot train on future data)
  - As it is not coping with the real world and the real life consequences of its behaviours, it does not learn in the sense that living things learn

What's next?

- Introduction
- The ML pipeline: proxies and alignment
- The training data: language behaviours
- Spoken and written speech acts
- Reasoning, inductive inference and abduction
- Abductive moments in LLM development and deployment
- Real world and real life
  - The plenitude of real world flux (B.C. Smith)
  - 'Intentionality' (Brentano) and 'Conatus' (Spinoza)
  - Two types of consequentialism

Spoken and written speech acts

- Speech acts (Anscombe rather than Austin, rather than Searle):
  - Locutionary: descriptive (you are now husband and wife)
  - Illocutionary: performative (you are now husband and wife)
  - Perlocutionary: influencing (I urge you to get married)

- Performative speech acts:
  - Do what they say
  - Are constitutive of institutional facts

- The civil servant who declares two people husband and wife
  - Does not cause them to marry (I hope)
  - Is not describing their marriage (not a matter of propositional logic)
  - Is constituting the marriage (= an institutional fact)

Spoken and written speech acts

- Speech acts (Anscombe rather than Austin, rather than Searle):
  - Are about *language use:* 'to do things with words'
  - Are about *'making' our shared institutional world*

- Spoken speech acts
  - May refer to brute facts (locutionary)
  - May be intended to influence (perlocutionary)
  - May create institutional facts (illocutionary)
  - The difference between brute and institutional facts is relative

- Written speech acts
  - Have a different performative effect compared to spoken speech acts
  - Distantiation between author/text, text/reader, author/reader, author/meaning
  - The scope and reach are extended and an interpretation problem is generated

Spoken and written speech acts

- Spoken speech acts
  - Can entail ostensive (deictic) reference, pointing at (things in a) shared *Umwelt*
  - Can entail non-ostensive reference, pointing out what is absent
    - Thus creating a shared *Welt* between interlocutors

- Written speech acts
  - Entail non-ostensive references, creating a shared *Welt* between audiences

- A shared *Welt* can include the *Umwelt*,
  consists of brute and institutional facts,
  is based on ostensive and non-ostensive references
- The ‘real world’ is the shared Welt we need to navigate to survive and flourish

Spoken and written speech acts

- LLMs do not engage in speech acts
  - Just like the script does not engage in speech acts, the author does, the reader may
  - Do the authors of an LLM engage in ‘computational speech acts’?
  - Speaking to whom? to computing systems? to those targeted/influenced/interacted with?
- LLMs do not entail an ostensive reference, they have no Umwelt
  - Just like written speech acts (though the author and the reader do)
  - The authors, deployers and ‘users’ of LLMs have an Umwelt
- LLMs do not generate performative effect, they do not create a Welt
  - They create ‘output’ in the form of text, images, graphs, code, decisions or behaviour
  - Depending on how the output is deployed it may generate
    - Illocutionary effects, if their output is held to be true
    - Performative effects, e.g. if their output creates virtual environments
    - Perlocutionary effects, if their output is used to influence or manipulate people

Spoken and written speech acts

- LLMs have no intentionality in sense of Brentano (consciousness of something)
  - Consciousness is not mysterious, nor is it a problem to be solved
  - It is a brute fact and an institutional fact
- LLMs have conatus (Spinoza), but unlike that of living organisms (Deleuze, Lévy):
  - They move from the *possible* to the *real* – based on a program
    - Deduction and Induction (note the design of the program will entail abduction)
  - Not from the *virtual* to the *actual* – based on interaction with an environment
    - An iteration of abduction, deduction and induction

Spoken and written speech acts

Deployment of LLMs as 'legal tech':

- The key constitutive concept in law is 'legal effect'
  - Legal effect is typically a performative effect: it does what it says

- LLMs may offer information about law (treated as if the output is a locutionary behaviour)
  - Legal search (Westlaw, Casetext, etc.)
- LLMs may be interpreted as speaking the law (treating the output as an illocutionary behaviour)
  - ADM deployed by public administration or courts
- LLMs may be deployed to persuade lawyers/clients (in the context of perlocutionary behaviour)
  - Legal search, prediction of judgment, ADM

- Will LLMs have perlocutionary effects on legal effect?

What's next?

- Introduction
- The ML pipeline: proxies and alignment
- The training data: language behaviours
- Spoken and written speech acts
- Reasoning, inductive inference and abduction
- Real world and real life
- Person, scientists and judge: on the nature of abductive inferencing

Reasoning, inductive inference and abduction

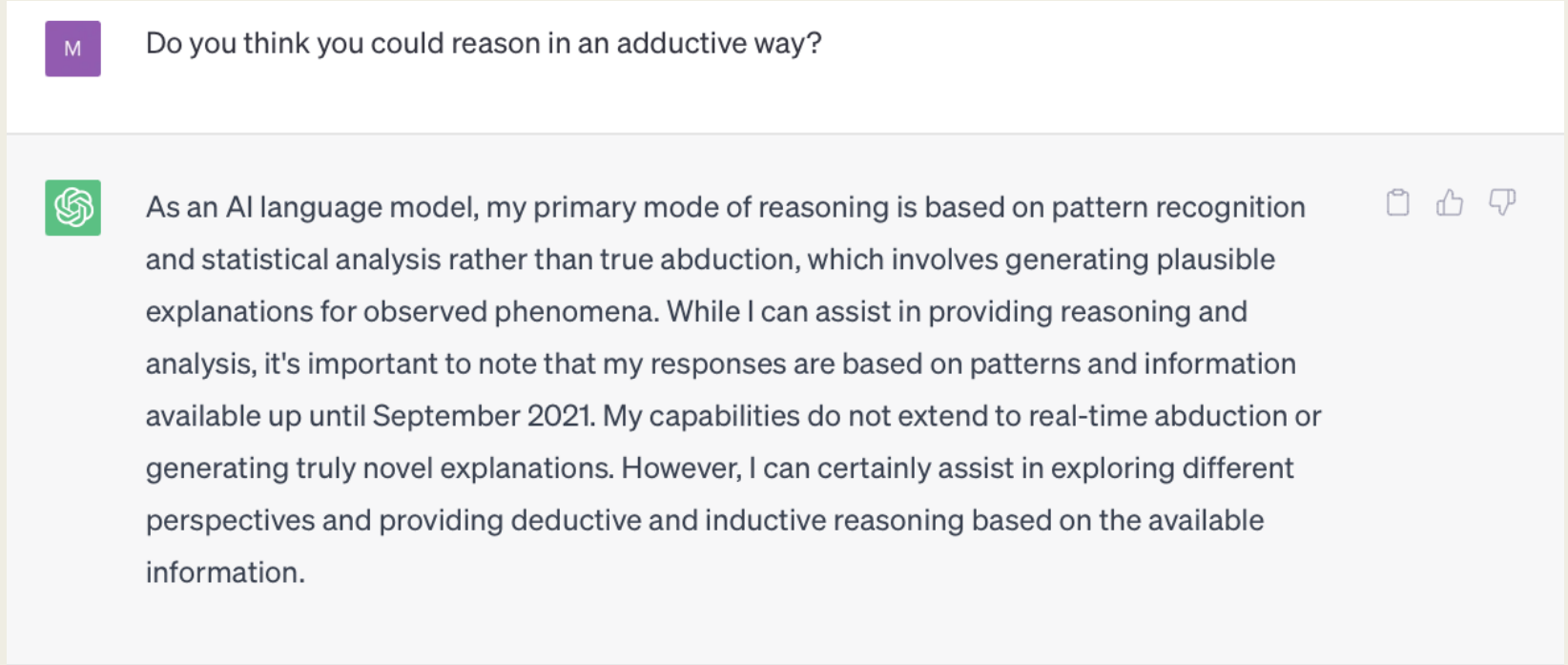
- Abduction (cf Peirce):
  - context of discovery (creative abduction, Magnani, filtering issues)
  - generating a theory or a hypothesis (or various theories/hypotheses)
  - followed by testing
    - using the abducted claim to deduce implications (logic)
      - to inductively ***falsify or verify*** (empirical)

- Context of justification or explanation impacts the context of discovery

Reasoning, inductive inference and abduction

- LLMs are data-driven, involving a combination of abductive and inductive ‘reasoning’
  - Those how think training ‘data’ are ‘given’ may think this is only inductive
- The construction and curation of training data involve abductive reasoning:
  - When selecting, ordering, cleansing the data, when adding syntactic data
  - When labelling the data (choice of labels)
- The construction of the model involves tentative anticipations of what might work:
  - the choices made when developing the hypothesis space
  - the development of a neural net
  - articulation of the target
  - the prompts

Reasoning, inductive inference and abduction

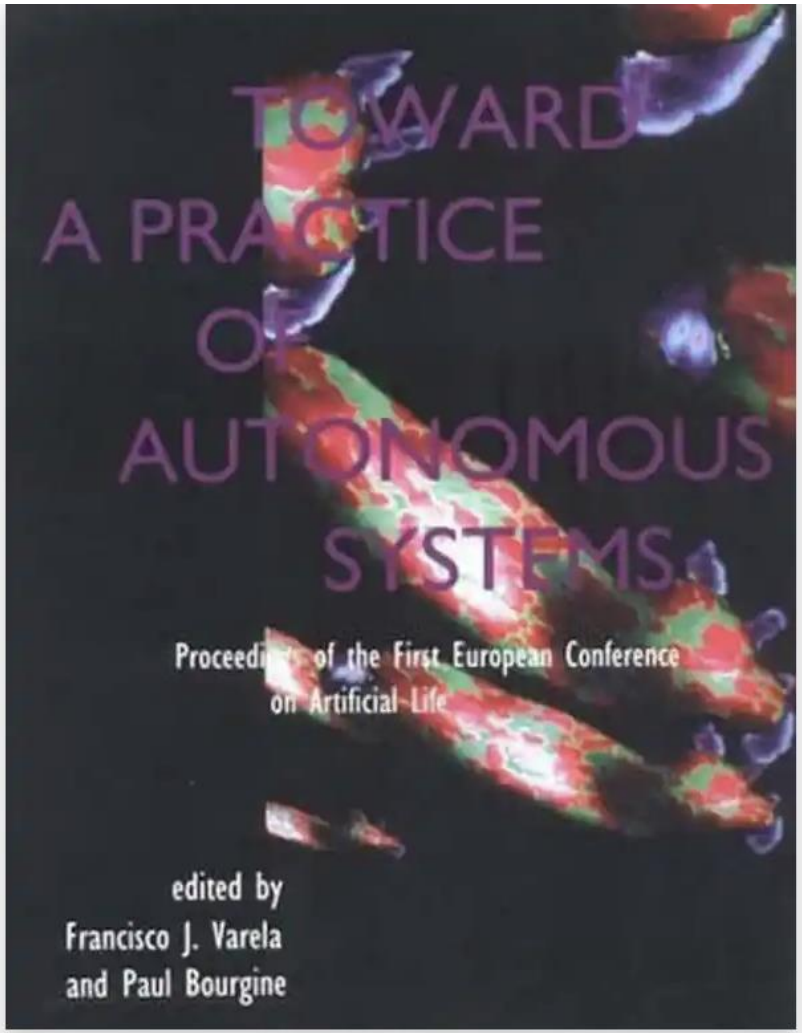


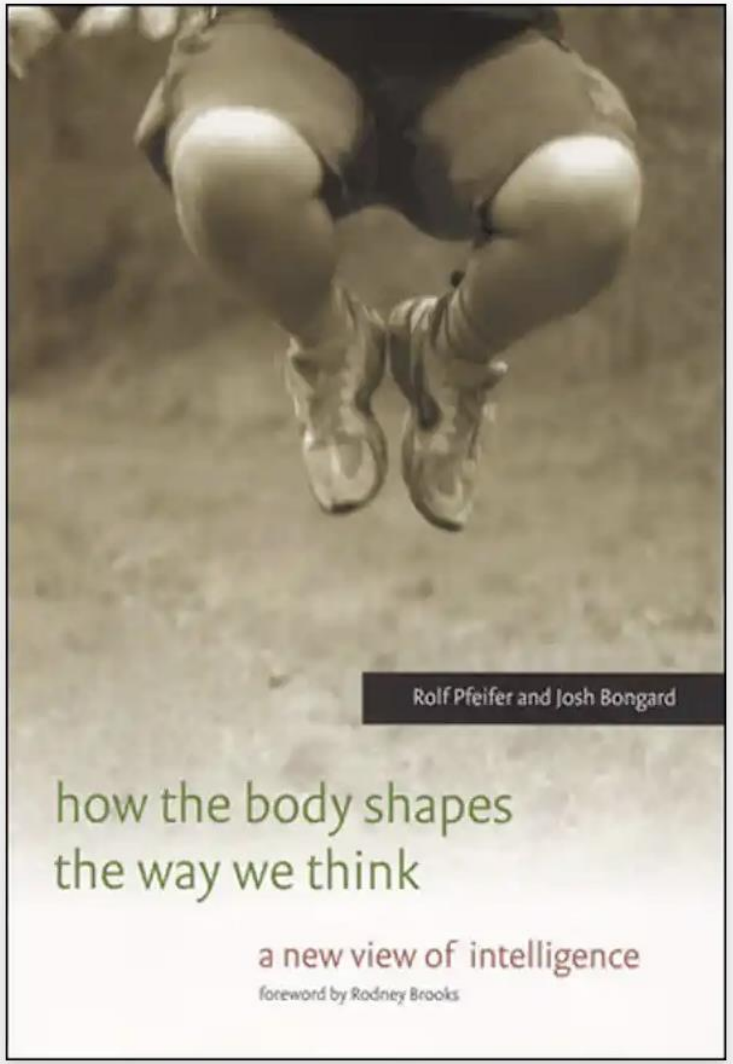
- https://chat.openai.com/share/23cde457-a2ac-4707-967c-460099b51b62

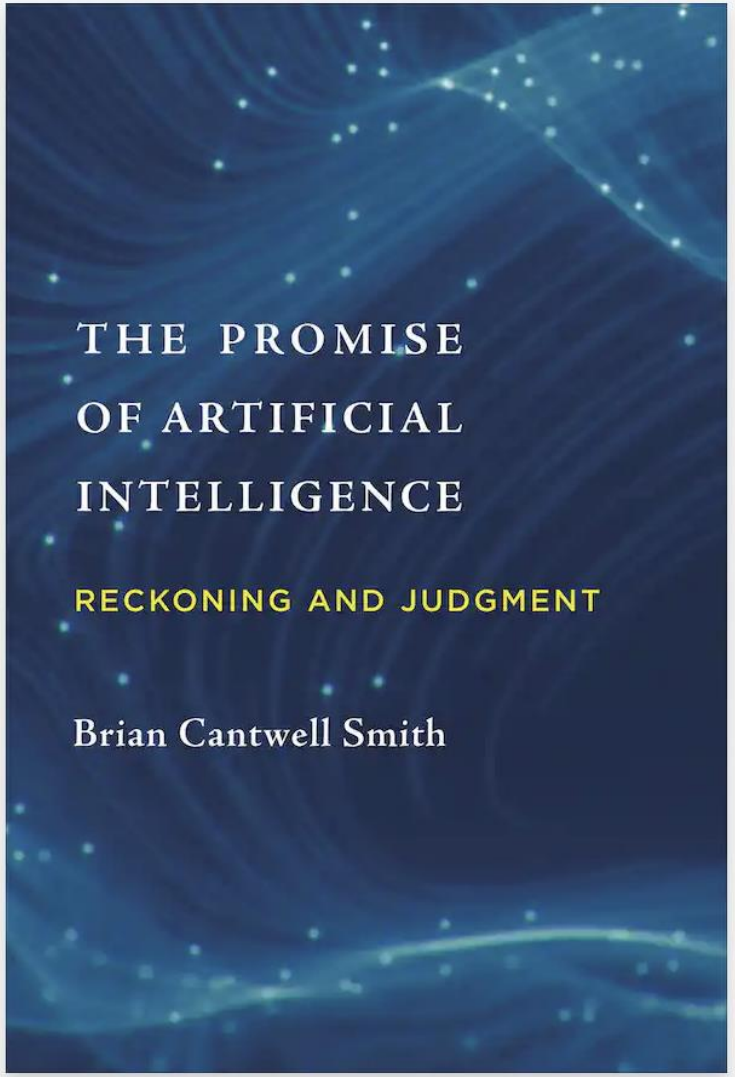
  (trained for modesty, and not distracted by the typo in the question)

What's next?

- Introduction
- The ML pipeline: proxies and alignment
- The training data: language behaviours
- Spoken and written speech acts
- Reasoning, inductive inference and abduction
- Real world and real life
- Person, scientists and judge: on the nature of abductive inferencing


TOWARD
A PRACTICE
OF
AUTONOMOUS
SYSTEMS
Proceedings of the First European Conference
on Artificial Life
edited by
Francisco J. Varela
and Paul Bourgine


Rolf Pfeifer and Josh Bongard
how the body shapes
the way we think
a new view of intelligence
foreword by Rodney Brooks


THE PROMISE
OF ARTIFICIAL
INTELLIGENCE
RECKONING AND JUDGMENT
Brian Cantwell Smith

What's next?

- Introduction
- The ML pipeline: proxies and alignment
- The training data: language behaviours
- Spoken and written speech acts
- Reasoning, inductive inference and abduction
- Real world and real life
- Person, scientists and judge: on the nature of abductive inferencing

The nature of abductive inferencing

- Living organisms: abductive movement to survive/flourish in an environment
  - Interacting to discover, tweak, ‘make the best’ of the affordances (Gibson)
- Scientists: abductive intuitions, made explicit in terms of reasoning
  - Thus enabling deduction of falsifiable hypotheses to inductive test (Peirce)
- Lawyers: abductive intuitions, tested against potential justification
  - Achieving a reflective equilibrium as to the best fitting applicable norm and interpretation
- LLMs: stochastic inferences, based on developers’ abductive intuitions
  - To be aligned by way of prompt engineering and tested in terms of real world validation …


THE END IS NEAR
ACTUALLY, THIS IS JUST THE BEGINNING
m.e.mcnair