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, tracing the building blocks of the architecture of such a space, including the required infrastructure and the relevant conditions for data quality. I conclude with a call to the health data science community to help the 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 in unreliable output and if used for decision making this can result in damage or even harm (especially in the case of medical artificial intelligence). To prevent harm, it is key to acknowledge the constructed nature of ground truth, foregrounding that it is the result of the selection of training data and the hard work of

[1] See: https://static-content.springer.com/openpeerreview/art%3A10.1186%2Fs12911-020-01224-9/12911_2020_1224_ReviewerReport_V0_R3.pdf

Hildebrandt, M.
Ground-Truthing in the European Health Data Space.
DOI: 10.5220/0011955900003414
In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - Volume 5: HEALTHINF*, pages 15-22
ISBN: 978-989-758-631-6; ISSN: 2184-4305

domain experts who label/annotate/rate the data in terms of preconceived labels/features/variables. This is of particular relevance for human decision-makers who consider using medical AI, notably because it enables them to be accountable to those subject to their decisions (patients). Considering the objectives of the General Data Protection Regulation and the upcoming legal framework of the EU for AI, keen attention to the upstream design decisions that define the ground truth will make the difference between lawful and unlawful design and deployment of medical AI. In this position paper, I focus on the proposed EU Regulation of the European Health Data Space, to mark out the discrepancy between an unsubstantiated faith in big health data and the need for high quality data in health and medical contexts.

Another way of framing the constructed nature of the ground truth is to acknowledge that the ground truth itself is not computable, though it can be approximated. In the case of supervised learning, this approximation depends on the annotations made by data scientists and domain experts that define the ground truth. Data scientists will be aware of the constructed and approximate nature of the ground truth and they will probably abstain from ontological truth claims when deciding on a specific ground truth. They are aware that they are 'merely' seeking a sufficiently corroborated point of departure to enable inferences and/or predictions.

However, once the outcome of ML research is implemented in decision (support) systems for medical diagnosis and treatment, it becomes pivotal that the trade-offs inherent in ground truthing are shared with those who consider deployment of the output.

3 PIERCING THE VEIL OF OBJECTIVIST ACCOUNTS OF GROUND TRUTHING

3.1 Supervised Machine Learning

Opening the black box of 'ground truthing' will also contribute to the domain of explainable AI (XAI), as it forces developers and providers of medical AI to account for the design decisions that are inherent in developing the ground truth, by showing the trade-offs inherent in such decisions. This should deflect from objectivist accounts of 'ground truthing' that suggest that the way it has been constructed equates it with 'the truth'. We should distinguish between 'objective' and 'objectivist' approaches to ground truthing. The first denotes a well argued, cross-disciplinary and contestable construction of a ground truth; the second denotes claims to truth that hide relevant assumptions and resist contestation.

One way to pierce the veil of objectivism that distracts attention from key design choices, would be to develop new types of metrics that highlight the choices made when labelling training data. Cabitza et al. (2020) suggest three such metrics, providing a more granular account of how true, how reliable and how informative the choice of a particular ground truth actually is, by estimating the 'trueness' in terms of the distance between the human annotation and the unknowable true annotation, which in turn depends on a new metric for the degree of concordance between those who did the labelling and another one for the degree of correspondence between sample and reference population.

3.2 Unsupervised Machine Learning

Some may believe that the constructive nature of the ground truth only concerns supervised learning and can be avoided by using unsupervised learning. They may point to the successes of deep learning (DL) in winning complex games such as chess and Go or to OpenAI, and its ability to generate seemingly well-designed sentences, arguments and narratives. As to success in games, this is connected with the fact that they have fixed rules and a final set of potential moves and outcomes. Even if that final set is intractable, it is incomparable with the complexities and uncertainties of real world, let alone real life, scenarios.

As to OpenAI's successes in pre-training large language models (LLMs) on 'the entire internet', we have seen how the absence of understanding and more precisely the absence of real world confrontation results in stochastic parrots (Bender et al., 2021) and in a challenging mix of pure nonsense and fascinating simulacra (Bogost, 2022). The constructive and approximate nature of the ground truth becomes even more obvious here, though perhaps hidden from popular imagination because it concerns the twofold challenge of deciding on (1) what data to use as a proxy for whatever it is one wants to achieve, on (2) how to curate the data (remove noise, structure, add, integrate and interoperationalise data) and on (3) what 'learner' to develop for training on the data. It seems that some people may seriously think that having 'all the data of the internet' implies that one now 'has' all there is to know about reality, mistaking the proxy for what it stands for. This kind of thinking is not merely naïve but dangerously so. To confuse knowing something

about the world with knowing how to survive and flourish in the world (Cantwell Smith, 2019) is a recipe for disaster, especially for medical research and medical treatment (see relevant caveats for the use of LLMs in clinical practice in Singhal et al., 2022).

More interesting, then, is DeepMind's AlphaFold (Jumper et al., 2021). This concerns the challenge of finding a 'method to reliably predict a protein's structure just from its sequence of amino acids' as the website tells us.[2] The claim is that 'the ability to predict a protein's shape computationally from its genetic code alone – as a complementary alternative to determining it through costly and time-consuming experimentation – could help dramatically accelerate research'.

Alpha Fold is collaborating with EMBL-EBI, that is "a not-for-profit international institute that helps scientists realise the potential of big data. The institute collaborates with scientists and engineers all over the world, and provides the infrastructure needed to share data openly and fairly in the life sciences. It also performs computational research and delivers bioinformatics training for the global scientific community. EMBL-EBI is part of the European Molecular Biology Laboratory (EMBL)." EMBL-EBI curates the AlphaFold dataset in a way that allows linking with other biological datasets, such as the Protein Data Bank Europe UniProt.[3] It is interesting to see how DeepMind frames the issues at stake, namely as a grand challenge to be solved. Considering the progress that has been enabled by AlphaFold one can understand the urge to think in terms of 'solutions', but clearly – as with all solutions to real life problems – these so-called solutions generate many new questions and may also create myriad new problems, such as the engineering of proteins that endanger entire ecosystems (with or without malicious intent).

It would help to not frame these tools as solutions but as tools, acknowledging that tools shape or reconfigure the goals they were aimed to achieve (Dewey, 1916). For instance, the goal may have been to help life sciences to speed up the experimental testing of protein architectures, whereas the success of the tool may instead achieve the replacement of experimental testing with computational predictions. The latter may not be helpful and bring along a plethora of risks to life on earth in ways that are difficult to foresee, even though it is not difficult to foresee that such risks may result in a catastrophe.

Back to unsupervised learning and groundtruthing. AlphaFold works with a transformer model and an attention architecture using multiple sequence alignment statistics, having moved from AlphaFold1 to AlphaFold2 with a leap in accuracy (Marcu et al., 2022). Some may believe that the problem of ground truthing can be solved by the use of transformer models, taking for granted that one could thereby drop the assumption that the distribution of future data is the same as that of the training data (Holzinger, 2016). Transformer models are based on calculating dependencies between distant sequential data, this way they can produce what is often qualified as context, thus developing sensitivity to the complex interactions between sequence and environments. By pre-training on very large data, such dependencies can be modelled and fine-tuned when further trained on more specific smaller data.

Not being a computer scientist my rendering here is probably somewhat retarded. However, even from the perspective of computer science, there is nothing final about the tool, which can interpolate and do some extrapolation but cannot accurately generate novel configurations. As Marcu et al. (2022) state: 'One limitation of approaches based on MSAs, such as AlphaFold2, is that they are constrained by our current knowledge and data sets'. This is a remarkable statement because any approach is necessarily based on current knowledge and data sets, the mere thought of escaping the laws of gravity that tie computational tools to available input data and known structures seems to confuse 'complex information processing' with human imagination or induction with abduction (Mooney, 2000).

Marcu et al. (2022) seem to believe that this problem can be solved by adding molecular dynamics, which would obviously create a more realistic picture, but would – also obviously – be constrained to known (or abducted) dynamics. Novel dynamics will depend on novel types of modelling *by researchers*, whose ability to imagine and abduct will make the difference.

Marcu et al. (2022) then refer to yet another constraint, noting that Alphafold2 was trained on a specific database with specific drawbacks; this sounds like the constraint they already mentioned, namely limitation to current datasets.

In all cases, the datasets and the knowledge deployed to train the model are proxies for an assumed ground truth. The design, the engineering or

[2] https://alphafold.ebi.ac.uk/about

[3] https://www.uniprot.org/database/DB-0070

the construction of this proxy is not only hard work but makes all the difference – it enables the learning process due to the constraints it imposes, and whether those constraints are relevant and productive can only be decided by testing the output model in real world environments – which may be far removed from the virtual laboratories of protein fold mappings.

3.3 Reinforcement and Interactive Machine Learning

As Holzinger (2016) argues, reinforcement learning (RL) concerns systems that are built to interactively learn from the environment they navigate (my paraphrasing, see also Pfeifer and Bongard, 2007; Russell, 2019; Cantwell Smith, 2019). Holzinger's (2016) claim is that human beings can help to reduce the search space of unsupervised systems, thus greatly advancing reliable outcomes in domains where time constraints or limited availability of relevant data present NP-hard problems, taking medical treatment as a prime example. Holzinger (2016, at 124) highlights the salience of RL as

> the first field to seriously address the computational issues that arise when learning from interaction with an environment in order to achieve long-term goals, because it makes use of a formal framework defining the interaction between a learning agent and its environment in terms of states, actions, and rewards. This framework is intended to be a simple way of representing essential features of general AI problems and features including a sense of cause and effect, a sense of uncertainty and non-determinism, and the existence of explicit goals.

I could imagine that deployment of RL, and what Holzinger calls interactive machine learning (iML), has a better chance of 'getting things right' than OpenAI's stochastic parrots, especially when domain experts interact with these systems to reduce what Holzinger (2016, at 119) calls the otherwise 'exponential search space'.

He then defines iML as

> algorithms that can interact with both computational agents and human agents [mh: 'oracles'] and can optimize their learning behavior through these interactions.

Linking back to the previous section, this comes close to what is now being defined as 'prompt engineering' that highlights the crucial role of human interaction. Gilson et al (2022) have 'tested' ChatGPR performance on the Medical Licensing Exams, concluding that it could be an interesting educational and knowledge assessment tool. A similar test has been conducted for the Bar Exam by Bommarito II and Katz (2022), again with key attention to prompt engineering.

Clearly, the key role of human domain expertise in iML testifies to the need to mitigate risks inherent in ground truthing, especially (though not only) in the case of unsupervised and reinforcement learning. This confirms potential problems with the interoperability of medical data across different jurisdictions and healthcare systems, due to the fact that data have often been collected and stored based on different purposes which render aggregation in a shared data space hazardous at least.

As argued above, blind trust in the 'trueness', relevance and interoperability of health data is a very bad idea, even when the data is properly curated. From the perspective of medical science and its methodological integrity and from the perspective of individual patients and public healthcare, we need to develop methods and methodologies to better understand what 'properly curated' means in the context of ground truthing and how independent supervisors can test whether the interplay between data and human intervention results in reliable and contestable output. Referring back to Holzinger's definition of iML, we should acknowledge that computational agents depend on the data they train on, whereas human agents have access to real world and real life implications.

4 EUROPEAN HEALTH DATA SPACE

4.1 Secondary Use of Health Data

In the 90s of the last century, Van der Lei (1991) warned against using medical treatment data for other purposes than those for which they were collected. Not because he was concerned about violation of the fundamental right to data protection but because of his concern for the inherent unreliability of such data. For instance, data may have been configured in a way conducive to compensation by an insurance company or conducive to obtain permission for a specified test.

Also, the data is necessarily skewed by the fact that people with similar health problems do not necessarily all seek medical advice or treatment, due

to different access to healthcare, income or education. The latter means that specific types of data are absent from the training data and/or their distribution differs from real world distribution of the relevant health problems. This may differ between member states (MSs) of the European Union, causing incompleteness and bias in the data.

On top of that, the incentives to configure treatment data in one way or another depend on the way a national healthcare system has been organised, and overlooking how this relates to their accuracy and relevance will result in massive misinterpretation an 'mismodelling'. Cross-border aggregation will exacerbate these problems.

4.2 The Proposed Regulation on the EDHS

In 2022 the European Commission has launched the proposal for a Regulation on the European Health Data Space,[4] aiming to establish 'rules, common standards and practices, infrastructures and a governance framework for the primary and secondary use of electronic health data' (art. 1). This includes establishment of 'a mandatory cross-border infrastructure for the secondary use of electronic health data' (art. 2(e)).

The proposal defines 'data quality' as 'the degree to which characteristics of electronic health data are suitable for secondary use' (art. 2(ad)), and 'data quality and utility label' as 'a graphic diagram, including a scale, describing the data quality and conditions of use of a dataset' (art. 2(ae)).

Art. 33.1 reads: 'Data holders shall make the following categories of electronic data available for secondary use in accordance with the provisions of this Chapter:', listing a broad set of categories of health related data, such as '(a) EHRs [electronic health record systems]; (b) data impacting on health, including social, environmental behavioural determinants of health; (c) relevant pathogen genomic data, impacting on human health; (d) health-related administrative data, including claims and reimbursement data; (e) human genetic, genomic and proteomic data; (f) person generated electronic health data, including medical devices, wellness applications or other digital health applications;' and many more.

Art. 33 continues in paragraph 3 by stating that 'The electronic health data referred to in paragraph 1 shall cover data processed for the provision of health or care or for public health, research, innovation, policy making, official statistics, patient safety or regulatory purposes, collected by entities and bodies in the health or care sectors, including public and private providers of health or care, entities or bodies performing research in relation to these sectors, and Union institutions, bodies, offices and agencies.'

Though the purposes for which secondary use is permitted are limited, their articulation is very broad (art. 34), e.g. including scientific research related to health or care sectors; development and innovation activities for products or services contributing to public health or social security, training, testing and evaluating of algorithms and for providing personalised healthcare.

Though some purposes are explicitly prohibited (art. 35, for instance taking decisions that are detrimental for a natural person, decisions that exclude certain groups from insurance or marketing to health professionals) it is unclear how this could be monitored and enforced, knowing that the enforcement of purpose limitation in the context of the GDPR has been notoriously difficult.

The governance of the EHDS is attributed to Health Data Access Bodies (art. 36-43) that can issue data permits to access data to potential data users, provided a number of procedural and material conditions are fulfilled (including purpose limitation).

The proposed Regulation requires that a 'cross-border infrastructure for secondary use of electronic health data' is set up by designated contact points in the MSs (art. 52). Datasets available for cross-border access must contain a metadata catalogue that describes e.g. 'the source, the scope, the main characteristics, nature of electronic health data and conditions for making electronic health data available'. The European Commission will set up 'an EU Datasets Catalogue connecting the national catalogues of datasets established by the health data access bodies and other authorised participants' (art. 57.1).

Data made available through the health data access bodies may have a 'data quality and utility label', which is compulsory when processed 'with the support of Union or national public funding' (art. 56).

[4] Proposal for a Regulation of the European Parliament and of the Council for the European Health Data Space, 3.5.2022 COM(2022) 197 final, see https://health.ec.europa.eu/publications/proposal-regulation-european-health-data-space_en#details

The label must comply with the following elements (art. 56.3):

> (a) for data documentation: meta-data, support documentation, data model, data dictionary, standards used, provenance;
>
> (b) technical quality, showing the completeness, uniqueness, accuracy, validity, timeliness and consistency of the data;
>
> (c) for data quality management processes: level of maturity of the data quality management processes, including review and audit processes, biases examination;
>
> (d) coverage: representation of multi-disciplinary electronic health data, representativity of population sampled, average timeframe in which a natural person appears in a dataset;
>
> (e) information on access and provision: time between the collection of the electronic health data and their addition to the dataset, time to provide electronic health data following electronic health data access application approval;
>
> (f) information on data enrichments: merging and adding data to an existing dataset, including links with other datasets;

I challenge health data scientists to figure out whether these requirements can be met, and what it means that commercial entities need not comply with them. I also challenge them to explain what it would mean if compliance with these requirements is not feasible: does this imply that the requirements make no sense or that the attempt to develop medical training data at scale across MS borders is doomed to result in misinterpretation, mismodelling and damage to individual and public health. Maybe the delays that are foreseen for the establishment of the health data infrastructure connecting the national catalogues of datasets established by the health data access bodies and other authorised participants (Pištorová and Plevák, 2022), indicate the need to reconsider what is wisdom in the context of medical research and big data.

5 A PRECAUTIONARY APPROACH TO THE EDHS

The EU's quest to find ever more data to train ever larger training datasets, thus hoping to compete with other geopolitical regions, should not result in massively noisy datasets that cannot even trace – let alone resolve - the data drift and concept drift that are implied in this kind of research (Rahmani et al., 2022; Toor et al., 2020). The European legislature should take a precautionary approach to such aggregation, instead of assuming that more aggregated data provides for better science. A precautionary approach should take into account the caveats that e.g. Peek and Pereira Rodrigues (2018) develop with regard to the use of medical treatment data for health data science.

Starting with Van der Lei's (1991) warning against repurposing of training data in the context of health, they continue to discuss to what extent and on what conditions randomised clinical trials could be replaced by Big Data and finally they highlight the need for patients' informed consent for secondary use of their treatment data. In all three cases, they show the complexities and the drawbacks of secondary use. More precisely they demonstrate how such data should and should not be used for medical research. The proposed Regulation seems to combine challenging quality requirements with stringent obligations to share health data across MS borders.

Together with Peek and Pereira Rodrigues (2018), I urge the community of health data scientists to develop a research agenda that addresses these concerns, acknowledging that 'ground truthing' is hard work and involves decisions that are non-obvious and may have major impact on individual and public health.

I also urge the community to explain to the EU legislature what can and cannot be expected from the use of cross-border aggregates of health data, highlighting the gap between claims made on behalf of data-driven medical technologies and the substantiation of such claims, taking the example of our own Typology of legal technologies (Diver et al., 2023).

REFERENCES

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), Association for Computing

Machinery, New York, NY, USA, 610–623. DOI:https://doi.org/10.1145/3442188.3445922

Ian Bogost. 2022. ChatGPT Is Dumber Than You Think. *The Atlantic*. Retrieved January 9, 2023 from https://www.theatlantic.com/technology/archive/2022/12/chatgpt-openai-artificial-intelligence-writing-ethics/672386/

Michael Bommarito II and Daniel Martin Katz. 2022. GPT Takes the Bar Exam. DOI:https://doi.org/10.48550/arXiv.2212.14402

Federico Cabitza, Andrea Campagner, and Luca Maria Sconfienza. 2020. As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI. *BMC Medical Informatics and Decision Making* 20, 1 (September 2020), 219. DOI:https://doi.org/10.1186/s12911-020-01224-9

John Dewey. 1916. The Logic of Judgments of Practice Chapter 14. In *Essays in Experimental Logic*, John Dewey (ed.). University of Chicago, Chicago, 335–442.

Laurence Diver, Pauline McBride, Masha Medvedeva, Arjun Banerjee, Eva D'hondt, Tatiana Duarte, Desara Dushi, Emilie van den Hoven, Paulus Meessen, and Mireille Hildebrandt. 2022. The Typology of Legal Technologies. *COHUBICOL publications*. Retrieved January 9, 2023 from https://publications.cohubicol.com/typology/

Aidan Gilson, Conrad Safranek, Thomas Huang, Vimig Socrates, Ling Chi, R. Andrew Taylor, and David Chartash. 2022. How Does ChatGPT Perform on the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment. 2022.12.23.22283901. DOI:https://doi.org/10.1101/2022.12.23.22283901

Mireille Hildebrandt. 2020. *Law for Computer Scientists and Other Folk*. Oxford University Press, Oxford. Retrieved August 10, 2019 from https://global.oup.com/academic/product/law-for-computer-scientists-and-other-folk-9780198860884?cc=be&lang=en&

Mireille Hildebrandt. 2021. The issue of bias. The framing powers of machine learning. In *Machines We Trust: Perspectives on Dependable AI*, Marcello Pelillo and Teresa Scantamburlo (eds.). The MIT Press, Cambridge, Massachusetts.

Mireille Hildebrandt. 2022. The Issue of Proxies and Choice Architectures. Why EU Law Matters for Recommender Systems. *Frontiers in Artificial Intelligence* 5, (2022). Retrieved June 18, 2022 from https://www.frontiersin.org/article/10.3389/frai.2022.789076

Mireille Hildebrandt. 2023. Boundary Work between Computational 'Law' and 'Law-as-We-Know-it.' In *Data at the Boundaries of European Law*, Deirdre Curtin and Mariavittoria Catanzariti (eds.). Oxford University Press, Oxford.

Andreas Holzinger. 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Inf.* 3, 2 (June 2016), 119–131. DOI:https://doi.org/10.1007/s40708-016-0042-6

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Applying and improving AlphaFold at CASP14. *Proteins: Structure, Function, and Bioinformatics* 89, 12 (2021), 1711–1721. DOI:https://doi.org/10.1002/prot.26257

J. van der Lei. 1991. Use and abuse of computer-stored medical records. *Methods Inf Med* 30, 2 (April 1991), 79–80.

Ştefan-Bogdan Marcu, Sabin Tăbîrcă, and Mark Tangney. 2022. An Overview of Alphafold's Breakthrough. *Frontiers in Artificial Intelligence* 5, (2022). Retrieved December 30, 2022 from https://www.frontiersin.org/articles/10.3389/frai.2022.875587

Raymond J. Mooney. 2000. Integrating Abduction and Induction in Machine Learning. In *Abduction and Induction: Essays on their Relation and Integration*, Peter A. Flach and Antonis C. Kakas (eds.). Springer Netherlands, Dordrecht, 181–191. DOI:https://doi.org/10.1007/978-94-017-0606-3_12

Niels Peek and Pedro Pereira Rodrigues. 2018. Three controversies in health data science. *Int J Data Sci Anal* 6, 3 (November 2018), 261–269. DOI:https://doi.org/10.1007/s41060-018-0109-y

Rolf Pfeifer and Josh Bongard. 2007. *How the Body Shapes the Way We Think. A New View of Intelligence*. MIT Press, Cambridge, MA - London, England.

Barbora Pištorová and Ondřej Plevák. 2022. Stakeholders doubtful EU health data space will launch on schedule. *EURACTIV*. Retrieved January 9, 2023 from https://www.euractiv.com/section/health-consumers/news/stakeholders-doubtful-eu-health-data-space-will-launch-on-schedule/

Keyvan Rahmani, Rahul Thapa, Peiling Tsou, Satish Casie Chetty, Gina Barnes, Carson Lam, and Chak Foon Tso. 2022. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. *medRxiv* (June 2022), 2022.06.06.22276062. DOI:https://doi.org/10.1101/2022.06.06.22276062

Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Books.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj

Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large Language Models Encode Clinical Knowledge. DOI:https://doi.org/10.48550/arXiv.2212.13138

Brian Cantwell Smith. 2019. *The promise of artificial intelligence: reckoning and judgment*. The MIT Press, Cambridge, MA.

Affan Ahmed Toor, Muhammad Usman, Farah Younas, Alvis Cheuk M. Fong, Sajid Ali Khan, and Simon Fong. 2020. Mining Massive E-Health Data Streams for IoMT Enabled Healthcare Systems. *Sensors* 20, 7 (January 2020), 2131. DOI:https://doi.org/10.3390/s20072131