Law, Sciences and New Technologies

2

Law, Sciences and New Technologies

Open Access Series of the Interdepartmental Research Centre
European Centre for Law, Science and New Technologies (ECLT)
University of Pavia

URL: <http://www.paviauniversitypress.it/collana/LSNT/17/>

Texts published in the series "Law, Sciences and New Technologies" have been peer-reviewed prior to acceptance.

DATA-DRIVEN DECISION MAKING. LAW, ETHICS, ROBOTICS, HEALTH

AMEDEO SANTOSUOSSO, GIULIA PINOTTI (Eds)

PaviaUniversityPress

Data-Driven Decision Making. Law, Ethics, Robotics, Health / [edited by] Amedeo Santosuosso, Giulia Pinotti (Eds). - Pavia : Pavia University Press, 2020. - 108 p. : ill. ; 22 cm.
(Law, Sciences and New Technologies ; 2)

http://archivio.paviauniversitypress.it/oa/9788869521348.pdf

ISBN 9788869521331 (brossura)
ISBN 9788869521348 (e-book PDF)

ISBN: 978-88-6952-134-8

Nella sezione Scientifica Pavia University Press pubblica esclusivamente testi scientifici valutati e approvati dal Comitato scientifico-editoriale.

UPI
UNIVERSITY PRESS ITALIANE

Opera sottoposta a peer review secondo il protocollo UPI

Peer reviewed work in compliance with UPI protocol

In copertina: *L'isola* ..., opera di Sabrina Mezzaqui, Galleria Massimo Minini per collezione Paolo Consolandi, fotografia di Alessandro Lui.

Prima edizione: gennaio 2020

Pavia University Press – Edizioni dell'Università degli Studi di Pavia
Via Luino, 12 – 27100 Pavia (PV) Italia
http://www.paviauniversitypress.it – unipress@unipv.it

Printed in Italy

Table of Contents

Introduction

Whoever has been engaged in academic research, at least in the last decade, is well aware that participating in European calls is often a tall order. A good research idea and a solid scientific network are both essential, and it is at times arduous to fulfill the calls' specific requirements, because they often overstep the boundaries of individual disciplines.

Those are worthy moments that, besides being useful exercises, also allow a genuine discussion among scientific areas that usually do not overlap. The results of interdisciplinary transactions can be thrilling and their energy, once properly channeled and focused, can generate think tanks on the specific subject.

This is the story of our last year. This book is the witness and the result of a path that begun with our application to the Cost Action 2018 and, most importantly, with the creation of a network in order to apply to the ETN Marie-Curie 2019 call.

The first step along this route was the "Data Driven Decision Making" workshop that took place in Pavia[1] in October 2018. The aim of the workshop was to explore the different aspects of data-driven decision-making in the field of assistive robotics and its interaction with legal regulations. The application of robotics, autonomous systems, and AI to medicine may improve the process of diagnosis, care, and even rehabilitation. These new complex technologies produce a huge amount of data that requires new statistical approaches such as Big Data Mining and Analysis. The workshop addressed these issues from a highly interdisciplinary perspective.

Areas of application in medicine include Telemedicine (which allows the presence of immediate assistance to patients with chronic diseases living in remote places far from hospitals), Sensor technology (e.g. the electronic nose device that has received attention due to the applications in research and applied sciences), Surgical Assistants (i.e. remote-controlled robots that sup-

[1] The workshop wouldn't have place without the contribution of INROAd, University of Pavia.

port surgeons in performing operations), Rehabilitation Robots (i.e. robots designed with the aim to improve motor functions including coordination, postural control, and mobility in the environment, also using virtual reality systems). All these systems imply the collection of a sizable quantity of data. Such big data accumulations can be then explored through algorithms that in turn produce evidence for further decisions: i.e. data-driven decision making.

The aim of the Workshop was to establish the scientific common ground of the network and to gather adhesions to the research group in order to get prepared for the Marie-Curie call. The workday was so stimulating that an idea began to take shape in our minds: the contributions to the debate did not have to remain restricted to that particular moment, they had to become the body of an autonomous publication.

This is how this volume was born, which collects the ideas of Gabriella Bottini, Stefania Basilico, Valeria Peviani, Frederike Seitz, Tamar Sharon, Paul Vogel, Nicolas Woltmann, Riccardo Bellazzi, Francesca Bellazzi, Francesca Lagioia e Giuseppe Contissa.

The result is, in our opinion (and within the limits of proceedings of a workshop), very significant for three reasons:

1. Firstly, because it forced us to find a shared vocabulary among researchers whose backgrounds are very distant. This is very useful, when a subject (such as data-driven decision making) raises issues that can best be addressed only through an interdisciplinary dialogue.
2. Secondly, the high level of the contributions considered *per se* and as a whole.
3. The third point of strength of this work is precisely the network that has been created: researchers from several European countries were able to work together, regardless of their nationality and level of academic experience.

All this work would not have been possible without the contribution of several people that we want to thank: Sofia Baggini, Andrea Carini and all the staff of the Research Office (Università degli Studi di Pavia), Prof. Andrea Belvedere, Rector of Collegio Ghislieri, the Director of the Department of law (Prof. Ettore Dezza) who hosted the meeting and the Presidents of the Center CHT (Prof. Riccardo Bellazzi) and ECLT (Prof.sa Silvia Garagna) who continuously supported the initiative.

Pavia, December 2019 Amedeo Santosuosso – Giulia Pinotti

Science and Law in Big Data era: decisions, dilemmas and opportunities

Amedeo Santosuosso, Giulia Pinotti

1. Why to focus on data-driven decision making

Some elements of the present technological landscape are clear and largely described.

We are in the era of 4th industrial revolution, whose main characteristics are connectivity, distributed intelligence, industrialization of every process. Investments are mostly on cloud/digital infrastructure, with large capacity data centers and high-speed data communication. Different facets of such a reality are a) the *Internet of everything*, which includes the (let's say "old") *Internet of things*, i.e. the fast connection through cloud (and 5G in the next future) of services, industrial activity, hospitals and all aspects of smart cities, and the *Internet of people*, when the entities connected are humans; b) the huge quantity of data all these connections produce; c) Big Data analytics as a means for governing all this data and exploit them through the use of machine learning technologies, i.e. Artificial intelligence and data science.

Of course, the fact that all these facets are part of a unique interconnected environment does not mean they are a unique thing. They can be explored and studied in different ways and the result will largely depend on the research focus, e.g. if the connectivity or Artificial intelligence or social applications and so on. In our research[1] we have decided to focus on data-driven decision-making, i.e. how all these technological developments affect the way we take decisions either in scientific research or in the social, ethical and legal domains. It seemed to us that exploring decision-making would allow one of the best interdisciplinary theoretical experience.

In this paper we firstly outline the state of the art on big data and its opportunities in European documents, deserving some attention to the discus-

[1] For the scientific environment of this research at the ECLT and CHT of the University of Pavia, see Introduction.

sion about the kind of regulation to be introduced for new technologies. In the following paragraph, we describe some lines of research the data-driven revolution opens both in the scientific field of neuroscience (from the bench to health policies) and in law, as an autonomous field of big data research. A final conclusion stresses the opportunities and importance of an interdisciplinary approach.

2. The current debate on AI, big data and regulations

Current times witness an unprecedented generation of amounts of data of different nature, also called Big Data. This ranges from an ever-increasing amount of data from social media and Internet and mobile applications, to the growing digitization of all human activities (books, legal archives and medical records), to multimodal sensors data collected by robots and digital assistants. Big Data have pushed technologies towards new paradigms for their collection, storage and analysis. In particular, big data analytics uses machine learning methods and tools extensively, which makes possible to examine large amounts of data, to uncover hidden patterns, to find correlations and to infer other insights. Big data analytics is also useful to design new automated and autonomous data-driven reasoning and decision systems.

An interesting picture of the field can be found in some documents and reports by European Union institutions and public-private actors about the so called "big data revolution".

First of all, the European Parliament, in its *Towards a thriving data-driven economy*, stresses that in European countries and EU institutions it is a shared opinion that Big Data "has the potential to boost economic productivity and improve consumer and government services; […] may bring more business opportunities and increased availability of knowledge and capital, as long as governments and stakeholders work together in a constructive manner".[2]

In addition, the EU Commission delivered in 2017 a communication

[2] European Parliament, *Towards a thriving data-driven economy*, https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/towards-thriving-data-driven-economy). See also the Big Data Europe project, and the activity of the Consortium of European Social Science Data Archives CESSDA https://www.cessda.eu.

Building a European Data Economy[3] where the importance of the huge amount of collected data is clearly underlined: "as the data-driven transformation reaches into the economy and society, ever-increasing amounts of data are generated by machines or processes based on emerging technologies, such as the Internet of Things (IoT), the factories of the future and autonomous connected systems. […] The enormous diversity of data sources and types, and the rich opportunities for applying insights into this data in a variety of domains, including for public policy development, are only beginning to emerge. To benefit from these opportunities, both public and private players in the data market need to have access to large and diverse datasets".[4]

Also, private actors and industries are aware of the importance of Big data tools: "the increased volume, velocity, variety, and social and economic value of data signal a paradigm shift towards a data-driven socio-economic model. The significance of data will only grow in importance beyond 2020 as it is used to make critical decisions in our everyday lives".[5]

Of course this considerations have also undeniable consequences on the legislative activity: among the current legislative priorities and commitments to implement a connected Digital Single Market, the European Commission is working, closely with Member States and the independent Data Protection Supervisory Authorities, on the full application of the General Data Protection Regulation (GDPR), whose implementation is essential to 'safeguard individuals' fundamental right to the protection of personal data in the digital age'.

The European Parliament in its *Towards a thriving data-driven economy* "stresses that the processing of certain kinds of data, in particular personal data, falls under the scope of EU data protection law; urges, in this connection, the swift adoption of the Data Protection Package; […] Believes that more effort is needed with regard to the anonymization and pseudo-anonymization of data as a precondition for creative data innovation and a major step in lowering market entry barriers for start-ups and SMEs; believes that uptake technologies, including text and data mining, will be an important factor

[3] https://eur-lex.europa.eu/content/news/building_EU_data_economy.html.

[4] See also the European Commission Project CORDIS- Data-driven decision making for a more efficient society. https://cordis.europa.eu/project/rcn/204374_en.html.

[5] Big Data Value Association, European Big Data Value Strategic Research and Innovation Agenda, Introduction, available at https://businessdocbox.com/Business_Software/71308945-European-big-data-value-strategic-research-and-innovation-agenda.html.

in deriving added value from open datasets; points out, however, that a clear distinction must be made between the processing of personal data and other kinds of data, and that technological solutions that are privacy-enhancing by design must be devised; […] Stresses that all the principles laid down in EU data protection law, such as fairness and lawfulness, purpose limitation, the legal basis for processing, consent, proportionality, accuracy and limited data retention periods, must be respected by Big Data providers when processing personal data; recalls, in this context, the opinion of the European Data Protection Supervisor on privacy and competitiveness in the age of Big Data".[6]

2.1. Some remarks on the EU regulation and its potentially universal application

The technological global scenario is unquestionably dominated, in terms of investments and political power, by the USA and China. In a situation like this some interesting interstitial phenomena are emerging.

European legal and ethical regulations are one of them. It is worth noting that in the general present claim for regulating in some way Artificial intelligence and its application the European GDPR is considered a possible model for universal application and there is who (no matter how much this man is in conflict of interest)[7] proposes to apply the GDPR to the United States, something that only few years ago was unimaginable. In addition, a recent document *Guidelines For Trustworthy Ai* the ETHICS High-Level Expert Group on Artificial Intelligence, appointed by the EU Commission, has suggested some ethical rules, which start from a not usual and interesting assumption: ethical AI should be lawful according to EU legislation, which includes the Treaties of the European Union and its Charter of Fundamental Rights, the General Data Protection Regulation, the Reg-

[6] European Parliament, *Towards a thriving data-driven economy*, cit.

[7] Mark Zuckerberg, The Internet needs new rules. Let's start in these four areas. The Washington Post, 2019 March 30. Other interesting interventions on the discussion about regulation are Reed C. 2018 How should we regulate artificial intelligence? Phil. Trans. R. Soc. A 376: 20170360. http://dx.doi.org/10.1098/rsta.2017.0360; Wendell Wallach; Gary Marchant, Toward the Agile and Comprehensive International Governance of AI and Robotics [point of view], Proceedings of the IEEE (Volume: 107, Issue: 3, March 2019); Yochai Benkler. Don't let industry write the rules for AI; NATURE 01 MAY 2019, Nature 569, 161 (2019) doi: 10.1038/d41586-019-01413-1 (Yochai Benkler is a law professor and co-directs the Berkman Klein Center for Internet & Society at Harvard University in Cambridge, Massachusetts).

ulation on the Free Flow of Non-Personal Data and more, the European Convention on Human Rights and more. Well, all this legal traditionally considered overregulating stuff seems to have turned into the object of the (US and global) desire.

The second example is given by the UNESCO recently delivered document about the proposed Recommendation on rules to be applied to AI[8]. UNESCO claims the uniqueness of its perspective "thanks to its universality in membership and drawing on its multidisciplinary expertise". And really the document develops some considerations about cultural minorities, multilingualism as a way of preserving cultural diversities, the effect of AI on arts and more. At the end the drafters conclude "It is not only desirable but urgent that measures be taken to set up a non-binding global instrument in a form of a recommendation. A recommendation – considering its non-binding character and its focus on the principles and norms for the international regulation of any particular question – would be a more flexible method and better suited to the complexity of the ethical questions raised by AI"[9].

What seems to us worth noting is that the level of complexity of the present technological turn is so high to leave room, beside the unquestionable superpower of US and China, to other entities and initiatives that can contribute to a real debate about the future of the humanity.

3. Lines of research

Data-driven decision-making as a field of research both in science-tech and in the social (and thus ethical and legal) domains has just been opened and, despite the vast concern, is still largely unexplored in depth.

[8] UNESCO World Commission on the Ethics of Scientific Knowledge and Technology (COMEST), Preliminary Study On The Technical And Legal Aspects Relating To The Desirability Of A Standard-Setting Instrument On The Ethics Of Artificial Intelligence, Paris 21 March 2019, available at https://unesdoc.unesco.org/ark:/48223/pf0000367422?posInSet=1&queryId=3cbc48e0-b3bd-488e-879b-84c382cd577d. [one of the authors of this paper, A. Santosuosso is a COMEST member].

[9] UNESCO World Commission on the Ethics of Scientific Knowledge and Technology (COMEST), Preliminary Study On The Technical And Legal Aspects Relating To The Desirability Of A Standard-Setting Instrument On The Ethics Of Artificial Intelligence, Paris 21 March 2019, available at https://unesdoc.unesco.org/ark:/48223/pf0000367422?posInSet=1&queryId=3cbc48e0-b3bd-488e-879b-84c382cd577d.

3.1. Data-driven decision-making and neuroscience (bench, clinical setting and Public health)

A first line of research regards Big Data analytics and Data-driven decision-making and how they are changing decisional process in the field of neuroscience, from the bench (scientific research), to clinical settings and ultimately to Public health policies. These changes require a new scientific and professional conceptual frame and an integrated cross-disciplinary approach to the training of the future generation of researchers and professionals in the involved disciplines.

Neuroscience is a strategic scenario for scientific investigation relevant to human well-being and for hands-on interdisciplinary training. Neuroscience research is based on the interaction between data-driven and knowledge-driven decision-making. It also raises relevant challenges in the ethical and legal domains.

Neuroscience as a case study will allow early stage researchers to experience, first hands, the selection, use and integration of methodologies from different disciplines, e.g. combining technical skills (e.g. machine learning and data science methods) and neuroscience investigation techniques with legal and ethical methods.

A research like this requires the cooperation of physicians, neuroscientists, researchers, engineers, computer scientists, ethicists and jurists, who should develop a strong interdisciplinary approach. A highly interdisciplinary network of international experts is necessary in order to face timely and properly the several scientific, technological, ethical and legal challenges. A special attention should be reserved to the ability to make explainable to lay people even complex decision-making systems (algorithms).

This line of research offers something traditional academic disciplinary partitions are not able to give to young researchers and professionals. We are in the need to create a new generation of scholars and professionals able to work in a highly technological environment and interact with experts having different backgrounds. On the other side this approach creates cross-sectorial training opportunity between industry, academia and public bodies.

3.1.1. Decisional processes in the era of Data-driven decision making

Decisions are currently distributed along the scale from rule-based decision making to statistical reasoning to machine learning and AI. As a matter of fact, data-driven decision-making and traditional knowledge-driven decision making (such as rule-based systems) coexist in different combinations according to the fields of application, the situations and, largely, the availability of data. However, this coexistence is not simply complementary, being the two systems different in kind, assumptions and inspiration. Are new ethical and legal paradigms necessary to tackle these challenges?

On the line from knowledge-based towards data-driven decision-making some well-known problems incrementally arise:

i. *Biases in dataset*. Quality and characteristics of the dataset used: where do data come from? How are they collected and selected?

ii. *Biases in applied algorithms*. Quality and characteristics of applied algorithms. How do we deal with inherent human biases? Can or should this mimicking process remove human bias? What are the dangers of this process? How can a legal system safeguard the security and privacy of personal data needed to train such algorithms? Which is the current legal framework? Is the General Data Protection Regulation 2016/17 an adequate tool?

iii. *Explainability of AI produced results*. The fact that these processes are not always transparent and explainable raises new ethical and legal (regulatory) challenges. This is true in many fields, such as in scientific research, where at least a relevant part of success appears to be in the hands of the researcher rather than relying on automatic processes, or in medicine, where diagnostic and treatment decisions depend on the accountability of the single medical practitioner or of a professional team. And this is true even for law, where public (administrative and/or legislative) decisions have to be explainable, and ethics, where traditionally any decision process derives from an interindividual confrontation aimed at solving ethical dilemmas.

3.1.2. Neuroscience as an ideal case study on decisional processes

The Human Brain Project (HBP) is a notable example where new technologies have been applied for both, research on new models of brain functioning and clinics through the implementation of the so-called “Medical Informatics Platform”. Moreover, clinical neuroscience now requires an ensemble of high-technology tools and devices to support diagnosis (from brain images to functional tests) and therapy (including tele-health and brain stimulation).

These technologies are generating "big data" that one may exploit to support hypothesis generation and decision-making. Machines already take a large range of decisions (from elementary programmed consequences of established premises to more complex outcomes as a result of algorithms applied to datasets), together with human decisions. Human choices, however, are susceptible to the scenario of presentation and to emotional interference.

Robotics and AI-equipped devices can be profitably used to deal with neurodegenerative diseases, implying the collection of huge quantity of data and the exploration of such accumulation through algorithms which produce evidence for further decisions (data-driven decision making). They are already largely in use: *Telemedicine* (which allows for the possibility of immediate assistance to patients living in remote places far from hospitals), *Sensor technology* (e.g. the electronic nose device that has received attention for research and applied sciences), *Diagnostic Assistants* (software tools designed to support diagnostic decisions, relying on AI-methods and able to automatically classify, for example, images), *Surgical Assistants* (remote-controlled robots that assist surgeons in performing operations), *Rehabilitation Robots* (designed to improve motor functions including coordination, postural control, and mobility, also thanks to virtual reality systems), *Research on the human brain*. These advanced technologies facilitate diagnosis and treatment of patients with complex diseases that show a progressive reduction of autonomy, requiring progressively increasing assistance.

3.1.3. Research in neuroscience (the case of HBP)

Grounding on large data sets poses the problem of the access to them. Data derive from different sources and need to be categorized and selected. Difficulties of access, in fact, depend on the different styles of data collection and also on the diverse human styles (several classifications, multiple scales and tests).

Data access thus requires a systematic approach organized in large and multilayered projects involving clinical, scientific, informatics, ethical and legal components. A prototypical example of this is represented by the Human Brain Project (HBP), which also includes the Medical Informatics Platform (MIP). The MIP intends to act as a bridge between research in neuroscience and clinics and patient care. It implements data from different sources with the aim to provide the tools to improve our knowledge of the human brain and to identify the signature of diseases.

This novel approach is deeply interdisciplinary and involves clinical, applicative and research aspects. It requires continuous updating and also a specific training of different professional figures. The application of new technologies with the aim to find solutions to complex problems in complex environments requires the acquisition of knowledge to provide a global vision of such complexity.

3.1.4. … to clinics: the case of neurodegenerative diseases

Dementia affects over 47 million people in the world, and induces dependency and disability with huge social and economic impact *(World Health Organization, 2016*). There are different kinds of dementia, being Alzheimer's disease (AD) the most common. Vascular dementia (VaD) and dementia with Lewy bodies (DLB) are also frequent. Frontotemporal lobar degeneration (FTLD) is frequent as well and has an early onset. The difficult differential diagnosis among these conditions derives from the considerable overlapping of symptoms, although some diversity in the cognitive profiles also occurs. Furthermore, not only behavioral features, but also biomarkers and neuroimaging findings have become relevant for diagnosis, making the clinical process more and more complex. As a consequence, the definition of guidelines is problematic. Diagnosis of these diseases is typically multidimensional; furthermore, different approaches do not always converge on the same conclusions on the pathogenesis and the treatment. Clinical decision support systems (CDSS) could provide a systematic way for helping clinicians in this complex diagnostic process. CDSSs for differential diagnosis of dementia were proposed since the 90s (Plugge et al., 1990, 1991). The case of dementia represents a big challenge also for research as it is a complex disease with no clear pathogenesis, no effective therapy and also involving relevant problems from the epidemiological and social point of view. Due to the high incidence of dementia, new approaches such as *data mining and machine learning* (Neural Networks, Support Vector Machines, Random Forest) may well contribute to our knowledge of different aspects of the relevant conditions. These approaches are well suited to highlight hidden patterns in large data sets. As a matter of fact, a large variety of technologies are now tested in the area of home care and family/patient support, ranging from telemedicine (Marceglia et al, 2018, Piau et al, 2018) to assistive/social robots (Gongora Alonso et al, 2018). These technologies have "on-board" AI systems with some degree of autonomy that deals with many aspects of decision support, from monitoring to surveillance. The balance between knowledge-driven and data-driven algorithms is a crucial aspect of the design and implementation of such systems.

3.1.5. …and, finally, to Public Health

The health information systems must generate, analyze and disseminate data. Due to a number of intervening variables, (depending for example on the historical and economical scenario), this process is rarely systematic. This state of things determines important shortcomings in health management, with a general, diffuse inability to generate the data needed for a real progress in ameliorating the public health service. Solutions might come from the creation of a Health Information Network, in the frame of a EU and/or global collaboration. The case of neurodegenerative diseases is paradigmatic. Planning interventions is essential to: i) improve collective awareness; ii) improve diagnostic capabilities; ii) improve monitoring capabilities; iv) improve assistance.

At the Public health level, several issues need to be addressed: what are the best strategies to collect data really relevant for decision-making, given the introduction and implementation of new technologies? Which principles should guide the algorithm implementation to analyze the data and to take the intervention decisions? What kind of modeling can be useful for this purpose? Are simulation and agent-based modeling proper tools to support decision making?

3.2. Rule-based and data-driven decision making: understanding the point

A further line of research is focused on how the use of big data analytics may affect the inner nature of law: i.e. law-making and application.

Data-driven decision-making originally is a use of analytics in business for the purpose of taking decisions based on verifiable data and achieving productivity gains. Nowadays, data-driven decision-making has moved towards all social activities and is becoming a general model. The effects are sometimes positive and in other cases problematic.

All this calls for ethics and law as sources of regulation of these new technologies. However, ethics and law, while regulating big-data induced social phenomena, are, at the same time, challenged by the applications of Big data tools within their own areas. Namely, legal decision-making, as traditionally based on rules (even though their nature may be different), is challenged by decision-making systems, which are based on data analysis.

All this requires a deeper analysis and understanding of what is the real issue at stake.

Data-driven and rule-based decision-making systems have different pros and cons. The typical problem of data-driven decisions is the quality of the data gathered, its analysis and interpretation, so that whatever mistake and/ or bias in one of the steps can heavily affect the decision.

The problem of decision systems based on rules is that being based on rules simply means that rules have been followed, whatever the quality and/ or efficiency of the taken decision. In any case, if applied to rule-based decision-making systems, data-driven approach might produce some positive effects, such as a reduction of the arbitrariness of the starting point and of some critical decisional steps.

Points to be clarified are as follows:

- Rule-based decision-making is not synonymous of decision made in a legal way. Rule-based decision-making is a more complex field. In its proper and strict sense it is the way of deciding according to a written, clearly defined rule. Frederick Schauer, who is the author of a fundamental study on *playing by the rules* is very clear saying that "rule-governed decision-making is a subset of legal decision-making, rather than being congruent with it"[10]. According to his very demanding definition, a decision taken by a legitimately appointed court according to "best interest" of the child or of the patient, or the system of equity or the sentencing process (where an unlimited range of factors may play a role) are all legal decisions even though they are not rule-based decisions, because of the nature and intrinsic quality of the rule/ criterion to be applied.
- In the broader field of legal+rule based decisions we can find also an important divide/opposition, such as legal formalism and realism, where the crucial point is that of the role the rules play in the decision: if they are the master and guide to establishing/distributing rights and wrongs or they are an ex-post justification of a decision taken according to other (personal, political, social, emotional and more) reasons[11].

[10] Frederick Schauer, Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life, Clarendon Press, Oxford, 1991, p. 11.

[11] Well-known advocates of the American legal realism are Edward. H. Levi (E. H. Levi, An

- Data-driven is not equivalent of data-based decision-making, as if it were synonymous of evidence based in opposition to arbitrary decision.
- Data-driven decision-making means to decide accordingly to what emerges from the application of machine learning algorithms. Taking advantage of the "distinction between a forward and an inverse problem" (where the forward approach –from the *model* to the *observable*– is that used in experimental or quasi-experimental approaches), "the inverse approach is the heart of machine learning", where "one uses the *observables* to build the *model* rather than using the model to assign causal weight to those observables"[12].
- In theoretical terms the question of the nature and interpretation of the results of algorithms is at the forefront. Following Kevin Ashley's (2017) recent and detailed analysis of the situation, we can stress that "since a Machine Learning (ML) algorithm learns rules based on statistical regularities that may surprise humans, its rules may not necessarily seem reasonable to humans. ML predictions are data-driven. Sometimes the data contain features that, for spurious reasons such as coincidence or biased selection, happen to be associated with the outcomes of cases in a particular collection. Although the machine-induced rules may lead to accurate predictions, they do not refer to human expertise and may not be as intelligible to humans as an expert's manually constructed rules. Since the rules the ML algorithm infers do not necessarily reflect explicit legal knowledge or expertise, they may not correspond to a human expert's criteria of reasonableness.[13]"
- The question is how to combine the intrinsic nature of patterns emerging from legal analytics (and their limited explainability) and the right to explanation of public (and sometimes even private) decisions, which basic constitutional provisions recognize to humans.

Introduction to Legal Reasoning, The University of Chicago Press, Chicago-London, 1949) and Oliver W.Holmes (O.W. Holmes, The Path of Law, 10 Harvard Law Review 457, 1897).

[12] Daniel Martin Katz, Quantitative Legal Prediction – or – How I Learned to Stop Worrying and Start Preparing for the Data Driven Future of the Legal Services Industry, Emory Law Journal, Vol. 62, 2013.

[13] Kevin D. Ashley. "Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age", 2017, p. 111.

4. A path is open

Our aim was to give some examples of how many lines of research are possible in the freshly opened area of data-driven decision-making. It seems to us that a very preliminary exploration in the scientific and legal field shows the great potentiality in both areas with very promising interactions between different disciplines and approaches.

References

Ashley, K. D., *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*, Cambridge University Press, Cambridge, 2017.

Holmes, O.W., «The Path of Law», 10 *Harvard Law Review* 457, 1897.

Katz, D. M., «Quantitative Legal Prediction – or – How I Learned to Stop Worrying and Start Preparing for the Data Driven Future of the Legal Services Industry», *Emory Law Journal*, Vol. 62, 2013.

Levi, E. H., *An Introduction to Legal Reasoning*, The University of Chicago Press, Chicago-London, 1949.

Schauer, F., *Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*, Clarendon Press, Oxford, 1991.

Data Driven Decision Making in Neuroscience: Pros and Cons

Gabriella Bottini, Stefania Basilico, Valeria Peviani

1. Introduction

The amount of data produced and collected in many scientific environments is enormously increasing, posing problems related to their volume as well as to their extreme variety, which require a complex process of selection to be performed. Their volume and variety outstrip the limits imposed by canonical statistical analysis, that grounds on manual approaches and conventional databases. The cooperation among centres operating in different fields of science requires datasets to be shared and networking to be easily implemented. These processes are being facilitated by the exponential development of IT supplies. As a consequence, deeper statistical analyses can be performed on such large datasets. Importantly, this multifaceted approach to data requires an ethical control to ensure safe data exchange in respect of the rules concerning i.e., privacy. Furthermore, to make informed choices and navigate within these complex data, consumers need to be provided with easily available, accurate, and up-do-date information. However, the availability of such an abundance of data does not always result in a correct to information may represent a real dilemma for the consumers. In this regard, Big Data Mining and Analysis represents a new scientific field of research with numerous theoretical, applicative and ethical issues.

There are a few scientific projects aimed at promoting the start-up of research structures, at helping neuroscience development and at consolidating this new approach of data collection and analysis. One of the largest is the Human Brain Project (HBP), funded by the European Union (https://humanbrainproject.eu). Among the aims of the HBP, there is the improvement of brain data sharing. Such project provides Medical Informatics Platforms that are meant to allow the access to data collected from patients affected by several neurological diseases. Collecting data from multiple sources aims at identifying some crucial signatures of diseases affecting the central neural system. Considering the enormous amount of data (behavioural, neuroimaging, biological…), it becomes clear that the HBP should also provide a solid ethical control, since the big data analysis might eventually conduct to new theoretical models of diseases in a new societal context.

Generally, when accessing data from different sources, a major problem is the peculiar style of each data type (i.e. how the data are coded). Therefore, IT platforms need to overcome such diversities in order to provide clear information and to develop new perspectives to address diagnostic and therapeutic challenges. In other words, as H. Markram elegantly wrote, "Neuroscience is like the infant brain - it is flooded with data and theories but lacks the ability to bring them together in a unified view" (Kandel, Markram, Matthews, Yuste & Koch 2013).

Coming to research on the brain, thus on its structure and functioning, an extended range of new technological, molecular and computational tools allows neuroscientists to record efficiently and accurately the neural activity as well as to map neural connections in the brain. Such tools strongly facilitate the understanding of correlations and interactions between neural systems and functions, even when complex behaviours, such as decision making, comes into play. This research approach has progressively led to the identification of different neuronal sub-populations, paving the way for the investigation of the features of each neuronal sub-population (i.e. mono or polymodal, differently responding to diverse stimuli). Such specificity allows to fragment complex behaviours in their elementary cognitive components. As mentioned above, the multi-layered research approach, from genetics to systems, produces a large amount of data extracted from different sources. Inevitably, the canonical statistical approaches apply several methods to process these data, providing results concerning distant fields of research. However, in order to understand the complex structures of the brain as well as its functioning in normal and pathological conditions, convergence of these different levels of research becomes crucial. Big Data mining and analysis may help, when managing large sample size, in understanding heterogeneities and commonalities across different sub-populations. For instance, in case of a limited sample, diverging datapoints are more likely to be marked as outliers, whereas working on larger samples allows to extract common features across many sub-populations even in case of large inter-individual variations.

2. The emblematic example of dementia

Dementia is a disease that is progressively increasing in prevalence (Prince et al. 2013). Epidemiological surveys on dementia address the following elements: 1) the descriptive element, for which ratios are calculated considering the communities and populations enrolled; 2) the analytic element, which attempts to explain phenotypic variations observed by the identification of risk factors. Dementia rates are growing at alarming proportion worldwide and are related to population aging (Prince et al. 2013). Dementia was estimated by the Global Burden of Disease 2010 Study as the third leading cause of years lived with disability at global level together with other neurological diseases (GBD 2010, https://www.thelancet.com/gbd). So far, no cure is available for dementia, and its predicted increase will put the health systems as well as the caregivers of patients under an unprecedented pressure all over the world. Research on dementia is very active: the main purposes are diagnosing such disease in its very early stage before the brain is affected by a massive atrophy, as well as identifying specific therapeutic approaches. In this regard, the identification of biomarkers is a very promising field of research. Among these biomarkers, the protein tau, which is detected and quantified in the cerebrospinal fluid, is used as a marker of Mild Cognitive Impairment (MCI) that represents a very early phase of the Alzheimer disease (Vos et al. 2013a; 2013b; Petersen, Caracciolo, Brayne, Gauthier, Jelic & Fratiglioni 2014). Meanwhile, considering the unmanageable social burden of dementia, research has been trying to identify Disease Modifying Therapies (DMTs) as ways of slowing the progression of dementia. However, it is very clear that DMTs generally produce very disappointing results (https://clinicaltrials.gov in 2017; Canevelli, Bruno & Cesari 2017). Research has been also focusing on defining specific genotypes and phenotypes. In relation to the latter, neuropsychological research aims at identifying sensitive and specific cognitive tests, assessing memory and other functions, which might be able to detect early deterioration of such cognitive processes, as a predictor of the development of a multifactorial decline (Ismail et al. 2016). Taken together, these observations suggest that dementia diagnosis is a multidimensional construct and that collecting data from different sources may increase the power of research programs at different levels.

So far, the diagnostic approach to dementia is multidimensional. However, conclusions gained by different approaches are not always convergent. Furthermore, the available clinical classifications of diverse mental declines are unsatisfactory, any attempt for new classifications is disappointing, pharmacological and non-pharmacological treatments seem to be either ineffective or only partially effective in slowing the clinical progression of the disease.

Integrative approaches to such a complex disease appear to be ideal in order to trace clearer diagnostic pathways, to provide less dispersive classifications, and finally to suggest adequate therapeutic protocols (Ienca, Vayena & Blasimme 2018). Big Data refers to enormous amount of data that can be analysed through novel mining techniques for different purposes. Such approach is related to the possibility for a large number of clinical and research centres to access to these resources. Big data can be structured, such as cognitive tests' scores, or non-structured, such as ecological variables. Dementia can be represented with a complex model involving genetic, biological, clinical, behavioural and social components, all of them contributing to the diagnosis and the staging of the disease. This is the typical condition for which structured and unstructured data are both relevant to define the course of the disease. The combination of apparently divergent data may provide new interpretations of dementia. For instance, one of the typical problems when monitoring the course of dementia, which inevitably and progressively brings to a decreased autonomy of the patients, is the divergence between the close-to-normal scores at standardized tests and the deteriorated performance of patients within the ecological environment. As a matter of fact, caregivers frequently complain of the progressive worsening of who they are taking care of together with the parallel increase of their burden. Clearly, any chance of exploring correlations between ecological factors such as spatial orientation, strategic skills and cognitive functions such as memory, attention and language, would greatly facilitate the understanding of pathological behaviours in patients with mental decline, as well as the planning of specific neuropsychological rehabilitation protocols. To summarize, considering the issue of diagnosing and treating dementia, two points could be made. On the one hand, the integration and correlation of large amount of information across large groups of patients have been used to achieve early detection of the disease, typically during the MCI phase. To this purpose, polycentric investigations have accessed and analysed data from multiple sources including neuroimaging, by means of machine learning algorithms (Amoroso et al. 2017; Mathotaarachchi et al. 2017; Souillard-Mandar et al. 2016; Koronyo et al. 2017). In some cases these studies successfully detected significant predictors of the evolution of dementia (Mathotaarachchi et al. 2017; Souillard-Mandar et al. 2016). On the other hand, the combination of structured and unstructured data has been applied in order to achieve a better understanding of complex behaviours, such as decision making. However, there are a number of limits of such approach. Firstly, the use of biological markers has been criticized since a relevant proportion of MCI patients - estimated between 8 to 31% - does not evolve in dementia, rather returns to the pre-morbid cognitive state (Canevelli et al. 2016 for a review). Secondly, when complex cognitive and behavioural variables are considered, the emotional aspects are often disregarded. Finally, how personal data should be handled is still a matter of debate, as well as object of ethical discussion (Vayena, Gasser, Wood, O'Brien, Altman 2015). The prevalence of dementia seems to vary with clinical as well as cultural and socioeconomic factors. Interestingly, overall prevalence of dementia turned out to be higher in developed countries. This phenomenon may reflect the different level of exposure to cerebrovascular risk factors like hypertension, smoking habit, obesity, and diabetes, across nations. It is evident that dementia is represented by a complex pathological model including clinical and societal variables that must both be considered when exploring the causes of this neurodegenerative disease. Clearly, managing big data will greatly contribute in this investigation (Ienca, Vayena & Blasimme 2018; Khachaturian, Meranus, Kukull & Khachaturian 2013).

In summary, although Big Data Mining and Analysis represent powerful tools to clarify complex problems in Medicine and Neuroscience, its application needs rigorous scientific, clinical and ethical protocols to be defined.

From the theoretical point of view, many serious diseases, including Alzheimer's disease, schizophrenia, Attention Deficit Hyperactive Disorder, depression, and anxiety, have been shown to be related to dysfunctions of brain connectivity networks. The outstanding question is how these diseases modify the brain. We are still very far from having understood the hierarchical, complex, functional organization of the brain. Advanced neuroimaging has largely contributed to such issue, providing great potential for the study of functional brain networks. The hypothesis is that new, advanced statistical methods applying data mining and machine learning, such Neural Networks, Support Vector Machines and Random Forests, can contribute to the integration and the understanding of the large amount of data concerning the brain collected from different sources.

References

Amoroso N., M. La Rocca, S. Bruno, T. Maggipinto, A. Monaco, R. Bellotti, et al. (2017). "Brain Structural Connectivity Atrophy in Alzheimer's Disease". arXiv preprint arXiv:1709.02369.

Canevelli M., G. Bruno, M. Cesari. (2017). “The sterile controversy on the amyloid cascade hypothesis”. *Neurosci Biobehav Rev*, 83, pp. 472-473.

Canevelli M., G. Grande, E. Lacorte, E. Quarchioni, M. Cesari, C. Mariani, et al. (2016). “Spontaneous reversion of mild cognitive impairment to normal cognition: a systematic review of literature and meta-analysis”. *J Am Med Dir Assoc*, 17, pp. 943-948.

Ienca M, E. Vayena & A. Blasimme. (2018). “Big Data and Dementia: Charting the Route Ahead for Research, Ethics, and Policy”. *Front. Med*., 5, p. 13.

Ismail Z., E. E. Smith, Y. Geda, D. Sultzer, H. Brodaty, G. Smith, et al. (2016). “Neuropsychiatric symptoms as early manifestations of emergent dementia: provisional diagnostic criteria for mild behavioral impairment”. *Alzheimers Dement*, 12, 195–202.

Kandel E.R., H. Markram, P. M. Matthews, R. Yuste, & C. Koch. (2013). “Neuroscience thinks big (and collaboratively)”. *Nature Reviews Neuroscience*, 14 (9), p. 659.

Khachaturian A. S., D. H. Meranus, W. A. Kukull, Z. S. Khachaturian. (2013). “Big data, aging, and dementia: pathways for international harmonization on data sharing”. *Alzheimers Dement* 9, p. 61–62.

Koronyo Y., D. Biggs, E. Barron, D. S. Boyer, J. A. Pearlman, W. J. Au, et al. (2017). “Retinal amyloid pathology and proof-of-concept imaging trial in Alzheimer’s disease”. *JCI Insight* 2, 1–19.

Mathotaarachchi S., T. A. Pascoal, M. Shin, A. L. Benedet, M. S. Kang, T. Beaudry, et al. (2017). “Identifying incipient dementia individuals using machine learning and amyloid imaging”. *Neurobiol Aging*, 59, pp. 80-90.

Petersen R. C., B. Caracciolo, C. Brayne, S. Gauthier, V. Jelic, L. Fratiglioni. (2014). “Mild cognitive impairment: a concept in evolution”. *J Intern Med*, 275(3), pp. 214-228.

Prince M., R. Bryce, E. Albanese, A. Wimo, W. Ribeiro, C. P. Ferri. (2013). “The global prevalence of dementia: a systematic review and metaanalysis”. *Alzheimers Dement*, 9, pp. 63-75.e2.

Souillard-Mandar W., R. Davis, C. Rudin, R. Au, D. J. Libon, R. Swenson, et al. (2016). “Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test”. Mach Learn 102(3), pp. 393-441.

Vayena E., U. Gasser, A. Wood, D. R. O’Brien, M. Altman. (2015). “Elements of a new ethical framework for big data research”. *Wash Lee L Rev Online*, 72, 423.

Vos S.J., I. A. Van Rossum, F. Verhey, D. L. Knol, H. Soininen, L.-O- Wahlund, et al. (2013a). “Prediction of Alzheimer disease in subjects with amnestic and nonamnestic MCI”. *Neurology*, 80, pp. 1124-1132.

Vos S.J., C. Xiong, P. J. Visser, M. S. Jasielec, J. Hassenstab, E. A. Grant, et al. (2013b). “Preclinical Alzheimer’s disease and its outcome: a longitudinal cohort study”. *Lancet Neurol*, 12, pp. 957-965.

AI in Healthcare.
The avoidance of the infringements of rights and the interpretation issue

Frederike Seitz

1. Introduction

AI (Artificial Intelligence) and ML (Machine Learning) have begun to revolutionize entire disciplines and triggered leaps in performance of many digital technology thanks to the growth of big data and computational power. Society is being transformed by this technological revolution and the impact of innovations already outpace ethical and legal discourse. Specially, the Healthcare and the Life Sciences sector was highlighted as one that promises to be most benefited by the adoption of Machine Learning and AI technologies at scale.[1] The application of AI in Healthcare not only comes with great benefits for the patient but will also have a huge impact on the market. By 2021 it is expected that the growth of the AI health market will reach the $6.6 billion which is a compound annual growth rate of 40 percent[2].

Thus, the social and economic influence of the use of AI in Healthcare should not be underestimated and as the field is rapidly progressing, possible challenges need to be faced and discussed at an early stage.

2. The driving forces behind AI in Healthcare

To understand the challenges raised applying AI in Healthcare, it is inevitable to develop at least a basic understanding of the prime movers behind AI.

The current driving force behind AI is Machine Learning, i.e. the idea that – like children – AI can learn from experience and thus data. Data is

[1] https://www.economist.com/science-and-technology/2018/06/07/artificial-intelligence-will-improve-medical-treatments (last downloaded 18th November 2018).

[2] https://www.accenture.com/t20171215T032059Z__w__/us-en/_acnmedia/PDF-49/Accenture-Health-Artificial-Intelligence.pdf#zoom=50 (last downloaded 18th November 2018).

therefore an indispensable prerequisite for the use of Machine Learning; it is intimately linked to Big Data.

Broadly speaking, machine learning uses two main approaches to solve clinical problems: Firstly, "supervised learning", which e.g. a patient´s health state based on some data (e.g. picture of skin to diagnose skin cancer).[3] This type of approach is being very successfully implemented using current Deep Learning algorithms and can already outperform human clinicians in speed but more importantly in accuracy. Second, "reinforcement learning", an area of machine learning inspired by reward-based learning psychology, which concerns itself with how software agents should take actions in an environment so as to maximize some notion of cumulative reward.[4] These reinforcement learning algorithms have been deployed in healthcare to either learn treatment policies from direct interaction with patients, e.g. for learning dosage adjustments of pharmaceuticals during a treatment[5] but can also be used to learn from electronic healthcare records, so called "off -policy reinforcement learning", best practices across human clinicians' decisions.[6]

3. Big data and Healthcare – A cutout of possible risks and benefits

The use of AI will give rise to disruptive changes in Healthcare.[7] And, of course, there will not only be benefits, but also risks the people involved and also society will have to deal with.

The possible fields of application of AI in Healthcare range from "research" to "end of life care"[8]. In "drug research", AI might help to potentially cut the time to market for new drugs and their cost. In the context of "end of life care" it could help support clinical personnel, which opens up the possibility to spend more time with the patient.[9] Without doubt, the increasing number of possible (fields of) applications is accompanied by the further development of many products that can be of great benefit to the patients.

[3] Esteva & Thrun (2017).

[4] Aslund & Maurer (2018).

[5] E.g. Bothe & Faisal (2013).

[6] E.g. Komorowski & Faisal (2018).

[7] Jiang & Wang (2017) speak of a paradigm shift.

[8] https://www.pwc.com/gx/en/industries/healthcare/publications/ai-robotics-new-health/transforming-healthcare.html (last downloaded 18th November 2018).

[9] Beck (2018).

A major advantage of using AI is, that it can access a large amount of data and make decisions based on it. This promises to make it easier to diagnose rare conditions and diseases and identify the most promising treatments.[10]

The possible sources of data seem to be endless. In particular, patients can make a decisive contribution here. This might be illustrated by "wearables" such as the Apple Watch: Apple Watch series 4 is now capable to produce an ECG which can be sent to the treating cardiologist for analysis.[11] On the one hand, this has a decisive advantage for the patient himself. By monitoring his vital parameters and sending his data to the treating physician, possible illnesses can be prevented. On the other hand, given access to this data will help researchers analyze and draw conclusions from them. This shows that e.g. the Internet of things (IoT) allows that patients will play a major role in collecting clinically relevant data. The data might be used for prevention but also for the further development of possible treatments and improvement of the application of AI.

The application of AI in the healthcare sector might not only be a great advantage for the advancement and use of personalized medicine. Because of the availability of the various data, researchers will easily assemble even more data because it will get even easier and cheaper.[12] This inaugurates researchers from industry and academy to easily access the data and work with it. On the other hand, it is this easy access to the data that leads to problems. Due to the patients` possibility to collect and transfer data, he is getting more and more transparent which might provoke the misuse of data. One of the major challenges is therefore to take into account the right to informational self-determination and the privacy of the patient.

4. The patients' right to life and right to health

Of course, the application of AI in Healthcare poses many other issues in several areas, including privacy, regulation, commercialization, and other issues of intellectual property.[13]

[10] Ford & Price (2016).

[11] https://www.apple.com/apple-watch-series-4/health/ (last downloaded 18th November 2018).

[12] Ford & Price (2016).

[13] Ford & Price (2016).

But particularly in the Healthcare sector, questions have arisen regarding possible risks to patients´ legal autonomy as well as his rights to life and to health. Every action and decision in the healthcare area indirectly or directly affects the life and the health of the patient. Thus, this sector can be classified as the most vulnerable one applying AI. As they are fundamental rights, the right to life and the right to health must be protected and damage be avoided.

Therefore, more than in other fields of applications, the infringement of these legal and moral rights has to be prevented, in adequate balance with the interests of all other actors involved. For this reason, answers will have to be found with regard to liability and responsibility of all actors, including the question if an autonomous AI (Level 4 and above) is an actor.

As the field is rapidly progressing, it is necessary to discuss and work on such AI approaches now because, first, it is to expect that future developments will enable AI technology to learn and reach conclusions autonomously. And, as the basis for the decision of an AI can change with constant data growth, it is at least conceivable that it able to learn its own decision parameters. To sharpen the thought, in addition to learning legal rules, the AI might also learn to circumvent them.

Second, these challenges arise already now, in situations where a physician´s diagnosis is largely influenced by AI technology, so called hybrid decision-making[14] which also ask for legally and ethically conscious programming and training of AI.

5. Normative programming and the prevention of the infringement of the patients’ rights

While AI has many broad applications, AI in Healthcare exemplifies and amplifies all the challenges of human and AI interactions. Specifically, in Healthcare medical decisions are not only increasingly influenced by the application of technology, but also by the need for normative evaluations[15] and general adherence to existing legal requirements. But it is relevant for all other contexts in which AI will play a significant role.

One condition for responsible behavior is to design AI technology in such way that it incorporates but also learns mechanisms to prevent legal and ethi-

[14] Tzeng & Sheng (2017).

[15] Schramme (2016).

cal defendable decisions and be able to "explain" them and render them interpretable.[16] Another possibility could be seen in the use of "black box models". But, according to the current state of development, it is not justifiable to use those models as the patients right to life and right to health are potentially endangered.

This influence of adherence to legal requirements in the medical field can be illustrated by two examples:

> Example 1: Physicians are confronted by certain legal limits at diagnosis level. Within the context of a gene-analysis it might be discovered that a patient is suffering from Huntington`s disease. As this diagnosis is likely to have a serious impact in the patient´s mental state and social life, there must be an evaluation of the expected consequences of informing him or her. One of the questions to answered as part of this evaluation is whether or not the patient should remain ignorant of his or her diagnosis per his "right not to know" (§ 9 Abs. 2 Nr. 5 German GenDG).

> Example 2: In the medical sector one is permanently confronted with the patients´ "advance health care directive". Although this does not qualify as formal law, it must be part of any medical evaluation made. Should a patients´ advanced health care directive state that life-prolonging measures are not to be undertaken under certain circumstances, both the term "life-prolonging" and "certain circumstances" should be clearly and unequivocally defined in the specific case – often the patient would still accept such measures if they only are used for a very short time span to save his life during a surgery, e.g. the use of a heart-lung-machine.

The cases illustrate that for AI technology to be successfully applied in Healthcare, a true AI must be able to learn, interpret and apply legal requirements and limits on a case by case basis, instead of having them hardcoded by a programmer. Only then the patient´s right to life and health as well as the legal requirements regarding responsibility can be satisfied.

6. Normative programming and the interpretation issue

Of course, normative programming of AI in different context of decision-making is already discussed intensively: "Ethics by design" concerns the amalgamation of methods, algorithms and programming tools needed to endow AI with the capability to take into account the ethical aspects of their

[16] Explainable AI, Holzinger et al. (2017).

decision-making, as well as the methods, tools and formalisms to guarantee that its behavior remains within the given morals bounds.[17] The "Legal Tech" approach focusses primarily on an technology for the application of law and currently uses mostly supervised learning.[18]

These approaches have not yet been able to adequately solve the problems that are expected to arise. The main point of contention when comparing law and ethics is, that ethics mostly depend on philosophical concepts, whereas law is based solely on a rigid framework of law texts, case law and academic opinions. Laws, by definitions, try to exclude vagueness and at the very least offer a frame of reference to approach novel situations. In contrast, by now only "simple legal questions" can be answered using "legal tech".

This can be attributed to the fact, that legally correct decisions could be quite easily programmable if the analysis of an individual legal case can be reduced to a series of increasingly precise questions that can only be answerable by "yes" or "no". But, as is generally know and shown by the examples, this is not the case: Legal interpretation cannot be reduced to binary codes. To the contrary, the interpretation of laws is defined by several factors, such as the syntax, semantics of language and the person interpreting.[19] Also, it is influenced by underlying premises and principles, a certain cultural background and of course the legal system itself.[20]

Probably every legal system has its own interpretation methods.[21] But still there are cases which cannot be solved unequivocally. However, tracing this back to the major role language is playing in law and legal interpretation[22] would be in inadmissible simplification. Not only language but also law itself mainly depends, inter alia, on the person interpreting, his cultural background, his point of view and his interests. For example, if the "advance health care directives" states, that "life-prolonging measures" shouldn´t be undertaken, a Jehova´s witness might think of a blood transfusion as a "life-prolonging" measure, whereas an Atheist wouldn´t refuse such a remedy. Therefore, it is likely that the recourse to language-based AI approaches[23] will not be sufficient to solve the challenges adequately.

[17] Dignum (2018); Verbeek (2006).

[18] Buchholtz (2017); Fiedler (1980); Hähnchen & Bommel (2018); Remus & Frank (2016).

[19] Larenz (1991).

[20] Engisch (1963).

[21] For further information concerning the German interpretative methods, see Larenz (1991).

[22] Stamper (1991).

[23] Bengio & Jauvin (2003).

Instead, a possibility to face these challenges could be an integrative approach, converging towards a problem from different perspectives – such as medicine, law, social and computer sciences. This might also be an opportunity to provide a deeper understanding of law itself.

7. Perspective

The use of Artificial Intelligence has already changed the area of Healthcare and it is expected that it will undergo more fundamental changes.

At the same time, it has opened up innumerable possibilities. The collection and evaluation of data can provide new insights into diseases and new - personalized - treatment methods might be developed.

Whereas in many areas of application of AI the question of responsibility is discussed, the use of AI in Healthcare comes with special and even more pressing challenges. If the AI fails in the area of Healthcare, it might be life-threatening or at least a health-risk for the patient. Therefore, the primary aim of the scientific approach must be the avoidance of legal infringements. This not only raises the question of responsibility for a certain action: Even though AI in Healthcare may not yet replace physicians, there are human-in-the-loop-systems which are already applied. And even if the AI only supports, for example, the diagnosis of physician its decision is technically influenced.

One condition for responsible behavior might be to design AI technology in such way that it incorporates but also learns mechanisms to prevent legal and ethical defendable decisions and be able to "explain" them and render them interpretable. A major problem here, however, is that law is based, inter alia, on language which also needs to be interpreted. But as interpretation of law depends on many parameters, the question raised is whether the AI can be programmed or teached to make "one legal correct decision". It is more likely, that new, integrative, interdisciplinary approaches must be developed converging towards a problem from different perspectives.

References

Aslund, H., E. El Mhamdi, R. Guerraoui & A. Maurer. (2018). "Virtuously Safe Reinforcement Learning." *arXiv preprint arXiv*:1805.11447.

Beck, S. (2018). "Zum Einsatz von Robotern im Palliativ- und Hospizbereich". *MedR*, pp. 772-778.

Bengio, Y., R. Ducharme, P. Vincent & C. Jauvin. (2003). "A Neural Probabilistic Language Model". *Journal of Machine Learning Research*, pp. 1137-1155.

Bothe, M. K., L. Dickens, K. Reichel, A. Tellmann, B. Ellger, M. Westphal, & A. A. Faisal (2013). "The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas". *Expert review of medical devices*, 10 (5), pp. 661-673.

Buchholtz, G. (2017). "Legal Tech – Chancen und Risiken der digitalen Rechtsanwendung". *JuS*, pp. 955-960.

Dignum, V. (2018). "Ethics by Design: necessity or curse?". *Conference on Artificial Intelligence, Ethics and Society*.

Dignum, V. (2018). "Ethics in artificial intelligence: introduction to the special issue". *Ethics and Information Technology*, pp. 1-3.

Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau & S. Thrun. (2017). "Dermatologist-level classification of skin cancer with deep neural networks". *Nature*, 542(7639), p. 115.

Fiedler, H. (1980). "Functional Relations between Legal Regulations and Software". In: Niblett, B. (ed.) *Computer Science and Law*, pp. 137-146.

Ford, R.A. & N. Price. (2016). "Privacy and accountability in black box medicine". *Mich. Telecom. & Tech. L. Rev.*, pp. 1-43.

Holzinger, A., C. Biemann, C. Pattichis & D. Kell. (2017). "What do we need to build explainable AI systems for the medical domain?". *arXiv*: 1706.07979

Großfeld, B. (1990). *Unsere Sprache: Die Sicht des Juristen*, p. 52.

Hähnchen, S. & R. Bommel. (2018) "Digitalisierung und Rechtsanwendung". *JZ*, pp. 334-340.

Jiang, F. et al. (2017). "Artificial intelligence in healthcare: past, present and future". *Stroke and Vascular Neurology*, pp. 230-243.

Larenz, K. (1991). *Methodenlehre der Rechtswissenschaft*, *6. Aufl.*, pp. 271 ff.

Leike, J., M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq & S. Legg (2017). "AI safety gridworlds". *arXiv* preprint *arXiv*:1711.09883.

Lowery, C., & A. A. Faisal. (2013). "Towards efficient, personalized anesthesia using continuous reinforcement learning for propofol infusion control". *IEEE Neural Engineering* (*NER*), 6, pp. 1414-141?.

Remus, D., F. Levy. (2016). "Can Robots be Lawyers?". *Computers, Lawyers, and the Practice of Law (November 27, 2016)*. Available at SSRN: https://ssrn.com/abstract=2701092 or http://dx.doi.org/10.2139/ssrn.2701092).

Rissland, E. L. (1990). "Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning". *The Yale Law Journal*, pp. 1957-1981.

Schramme, T. (2016). "Gesundheit und Krankheit in der philosophischen Diskussion". In: Beck, S. (ed.), *Krankheit und Recht*, pp. 3-24.

Stamper, R. (1991). "The Role of Semantics in Legal Expert Systems and Legal Reasoning". *Ratio Juris*, pp. 219-244.

Tzeng, G. & Shen, K, (2003), *New Concepts and Trends of Hybrid Multiple Criteria Decision Making.*

Verbeek, P. P. (2006). "Materializing Morality – Design Ethics and Technological Mediation". *Science, Technology & Human Values*, pp. 361-380.

Data-driven decision making, AI and the Googlization of health research

Tamar Sharon

1. Introduction

The use of big data and artificial intelligence (AI) applications in the field of health and medicine is presenting exciting prospects for the future of disease detection, management of chronic conditions, drug discovery and even the delivery of health services. Machine learning in particular, one form of AI, has recently been shown to analyze images from radiology, dermatology and ophthalmology, to levels of accuracy that match clinicians' own abilities, thus successfully detecting diseases such as skin cancer, breast cancer and certain eye diseases (Esteva et al. 2017; Fogel and Kvedar 2018; Wise 2018). As is often the case with technological innovation, a good dose of hype surrounds the implementation of AI in health, and specialists are pointing out that these technologies are coming up against many practical limitations, while introducing a host of new challenges. For example, there is still a lack of standardization and interoperability in digital medical systems. Clinical practice, furthermore, often involves complex judgments that draw on contextual knowledge and the ability to read social cues that AI is unable, at least currently, to replicate. And, as has been increasingly researched in the fields of law enforcement and banking, AI is vulnerable to biases that are reproduced in decision-making (Agniel et al. 2018), that can have important consequences in medical treatment.

Yet even as the use of AI in health is at an earlier stage than the hyperbole surrounding it suggests, there is a strong desire among a number of stakeholder groups to see these tools developed and implemented in the healthcare and medical research setting, beginning with policy makers at the national and supranational levels (see for example EC 2016, 2018). Because of this, the development and implementation of big data and AI in health is advancing more rapidly than the process of finding answers to the thorny ethical, legal and societal questions that it raises. These include the potential of AI to make incorrect decisions and questions of accountability and responsibility in relation to them, effects on the role of healthcare professionals in practitioner-AI-patient relationships, the potential for dehumanization of medical care, privacy concerns, and others (see for example NCB 2018; EGE 2018). In this paper, I focus on one under-researched dimension of what we might call this ethical predicament of AI and big data in health: the fact that the experts in AI and big data are not so much biomedical researchers and clinicians, but large technology corporations. As I will argue, this adds an additional set of ethical and societal concerns to the already complex issues surrounding the implementation of big data and AI in health that need to be addressed before innovation in medical research and healthcare practices solidify.

2. The Googlization of health research (GHR)

The promise of big data and AI in health lies in the ability to capture many different types of data – messy, unstructured data that are also being generated outside of the traditional spaces of medicine – and to make sense of this data; to make it actionable. In this framing, healthcare and health science are reduced to a logistics process of data flows between scientists, practitioners and patients. And when health and medicine are framed as problems of effective management of complex data, experts in data management inevitably become experts in health. Thus, in the past few years, every major consumer technology corporation, from Alphabet to Apple, to Amazon, IBM, Amazon, even Facebook, has moved decisively into the health and biomedical sector. These are companies that for the most part have had little interest in health in the past, but that by virtue of their data expertise and the large amounts of data they already have access to, are becoming important facilitators, if not initiators, of data-driven health research and healthcare. I call this new model of research the "Googlization of health research" (GHR) (Sharon 2016).

GHR promises to advance health research by providing the technological means for collecting, managing and analyzing vast and heterogenous types of data. Apple's Health app, for example, which all iPhones come equipped with since 2014, allows users to store user-generated health data from various sources, including those generated by the iPhone. Since January 2018, it can now be used to integrate this data with users' electronic health records in collaborating hospitals across the United States, turning the iPhone into a "one-stop shop" for personal health data (Farr 2017). In the research domain, Apple's ResearchKit platform allows researchers to carry out medical studies on iPhones by collecting data using the phone's various sensors. Currently, more than twenty studies including over 100,000 participants are being carried out using the ResearchKit, in collaboration with some of the world's leading research institutes. Apple has also begun partnering with pharmaceutical companies who see potential in the iPhone and Apple Watch for virtual "at home" trials (Vincent 2016).

Like Apple, Alphabet is also exploring a number of channels for tying personal health-related data to biomedical researchers, and is developing new tools to capture and organize unstructured health data. Verily, its life science branch, has recently launched an ambitious project to comprehensively "map human health" (Verily 2018). The "Project Baseline", in partnership with Duke and Stanford Universities, will collect and analyze a wide range of genetic, clinical and lifestyle data on some 10,000 healthy volunteers over a period of four years, with the aim of creating a vast baseline dataset of human health, and to develop preventive treatments. Concurrently, Alphabet's London-based AI offshoot, DeepMind, is eagerly seeking out medical applications for AI and deep learning, for the prediction of, amongst others, cardiovascular risks and eye diseases based on retinal scans, medical outcomes of patients based on hospital data, and breast cancer based on mammographies (Poplin et al. 2018; Ram 2018). Google, Microsoft, Amazon and IBM have also begun packaging their cloud infrastructures as centralized genomic databases, where genomic researchers can store and run queries on genomic data. And most recently, a number of these companies have begun moving into the domains of employee healthcare and independent care centers, and health insurance (Muoio 2018; Wingfield et al. 2018).

As mentioned above, many of these techniques still have not delivered on their promises, all the while introducing a host of new challenges and limitations. Yet their potential remains promising, and places these corporations in a privileged position in the move towards the data-driven medicine and healthcare of the future. First, because personalized medicine requires linking various types of data, including the lifestyle and environmental data that can be generated by mobile devices. This is exactly what the Apple ResearchKit software allows researchers to do (Savage, 2015; Shen, 2015). Similarly, the new data repositories and data analytics expertise offered by technology companies – from clouds to machine learning – promise to overcome the limitations of traditional tools used by universities and hospitals. Second, because this new model of research may become paradigmatic of the multi-stakeholder, or "quadruple helix", collaborations that are envisioned in both Europe and the United States as a basis for the future knowledge economy, that bring together academia, industry, government and civil society. The much anticipated "All of US" personalized medicine initiative in the US, for example, will include collaborations between the NIH, universities, private companies, and citizens. Google is already advising one of the pilot studies that is probing how to recruit volunteers.

3. New strains on informed consent and privacy protections

But this new model of research also raises a number of risks and challenges. The power of large-scale data analytics is the capacity to combine datasets from highly different contexts with relative ease. Data here becomes, and should be, perpetually available for repurposing. But this poses important challenges to the principle of informed consent, a pillar of medical ethics, which can no longer realistically capture the risks arising from unforeseen uses of a participant's data. At the same time, advanced computational techniques and data mining approaches developed in recent years make it possible to re-identify data that has been anonymized with increasing ease (Gymrek et al. 2013; Sweeney et al. 2013). Anonymity, in other words, which is often upheld as a means of ensuring privacy, can no longer be guaranteed in this context. This tension is further exacerbated in the context of research facilitated by commercial actors in several ways. First, consumer apps and devices that generate health-related data occupy somewhat of a gray area between the highly regulated medical domain and the less regulated consumer market (Lucivero and Prainsack 2015). Privacy and data protection legislation for health information in the EU and the USA does not necessarily apply to data shared in such devices, nor to personal health data shared by an individual voluntarily in a social network or with a third party. Thus, some commentators have suggested that the presence of tech firms in health-related data collection and research is fertile ground for new forms of corporate surveillance, whereby personal health data may be sold to third parties such as advertisers, insurers and employers (Olson 2014; Zang et al. 2015).

Contextual approaches to privacy are helpful for understanding how privacy is challenged here. The legal philosopher Helen Nissenbaum (2010) for example, argues that privacy expectations differ depending on contextual circumstances: on the nature of the information, the type of relationship in which information is transferred and the uses to which it is put. Different norms, Nissenbaum maintains, exist for regulating information in different contexts – be they medical, social or commercial. Thus, information shared with one's doctor is not governed by the same contextual norms as information shared with a colleague at work. Yet the ease of flow of information that is digital contributes to a transgression of contexts, in which individuals' expectations of data privacy may be violated and through which they may become exposed to a-contextual interpretation. In the framework of medical research using digital data generated in non-medical contexts, then, there is a relatively high likelihood of context transgression.

A recent controversy surrounding a data sharing partnership between Google DeepMind and the NHS illustrates how some of these issues are already playing out. Announced in 2016, the collaboration between DeepMind and the Royal Free London, a NHS Foundation Trust, granted DeepMind access to identifiable information on 1.6 million of its patients in order to develop an app to help medical professionals identify patients at risk of acute kidney injury (AKI). Following an investigation, the Information Commissioner's Office (2017) ruled that this transfer of data and its use for testing the app breached data protection law. Namely, patients were not at all aware that their data was being used. Under UK common law, patient data can be used without consent if it is for the treatment of the patient, a principle known as "direct care", which the Trust invoked in its defense. But as critics argue, insofar as only a small minority of the patients whose data were transferred to DeepMind had ever been tested or treated for AKI, appealing to direct care could not justify the breadth of the data transfer.

4. Innovating privacy protection

This is not to say that privacy breaches and ill-use of informed consent mechanisms will necessarily accompany a Googlization of health research. As these companies move into the highly sensitive domain of health data, they will need to adapt to the complex regulatory landscape of healthcare and medical research. What's more, getting privacy right – certainly on the backdrop of heightened public scrutiny and mistrust of how large tech corporations handle personal data, fueled by scandals like Cambridge Analytica – seems to be a priority for them moving forward in health.

Apple, for example, has been proactive about clarifying its commitment to protecting the privacy of participants in ResearchKit studies. The ResearchKit is designed so that data is collected on the iPhone but is not available to Apple. Instead, it is encrypted and sent to the researchers who are conducting the study. Thus, the first five ResearchKit studies were designed in collaboration with Sage Bionetworks (http://sagebionetworks.org), a non-profit research organization that gathers the data collected on a participant's phone, de-identifies and codes it, and hosts it on a platform where researchers can access it. Sage acts as a repository, or mediator here, between users' phones and medical researchers. And while Sage is very open about the technical limitations of making data completely anonymous, they are positioning themselves as a trustworthy data-sharing facilitator. Another example of the importance of privacy in GHR collaborations is one of Verily's current research projects on Parkinson's disease, in partnership with Radboud University Medical Center in the Netherlands. The Personalized Parkinson's Project (https://verily.com/projects/precision-medicine/personalized-parkinson-project/), which will track 650 patients with early Parkinson's disease over two years, is collecting a vast array of multidimensional data including brain images, DNA, spinal fluids, clinical data and data collected by a high-tech wrist watch developed by Verily that will gather physiologic and environmental information on subjects, including things like movement, pulse, and ECG. Here too, the importance of patient confidentiality as a condition of success has been foregrounded from the beginning of the project design. From the outset, the question of how to securely store all this data and share it with other projects that are exploring Parkinson's (one of the aims of the study), was entrusted to a group of digital security specialists at Radboud University who developed a novel privacy-protection-by-design framework for the project that works via encryption and pseudonymization (Verheul and Jacobs 2017).

5. Thinking beyond privacy

It is not unreasonable to expect, then, that GHR projects may implement better data protection mechanisms than traditional, public research using multidimensional data and advanced data analysis. But privacy and patient confidentiality may be only one of the challenges that a move towards this new model of research entails. Just as important are questions about the value of personal health data and publicly-generated datasets, and what market advantage is conferred to commercial entities who can access them and develop treatments and services based on this access (Sharon 2016). The DeepMind controversy, for example, and the critical questions and discussions it engendered (Powles and Hodson 2017; The Guardian view, 2017), are not limited to the issue of privacy breaches and mis-construal of the category of "implied consent". GHR also raises questions about the newfound role corporations that are already very powerful in other domains of human activity will begin to play in healthcare and research, and the new power asymmetries between corporations, public health institutions and citizens as data subjects, that may ensue. For example, how open will the datasets that these companies are compiling be? Will they be proprietary or publicly owned? What kind of mediating or gate-keeping role will corporations play in deciding who has access to these datasets and what criteria will this be based on? Also, what role will these companies begin to play in setting healthcare agendas? In the past, the monopolies given to drug companies via the patent system led to abuses. The same may become true of valuable insights derived from data that citizens, recast as participants in research, voluntarily give away, or datasets that public institutions like hospitals allow companies access to.

These are questions that concern collective and societal benefit, and that foreground a number of concerns that move beyond (just) privacy and informed consent, including accountability, social justice, democratic control and the common good. It is paramount that these values find their way into new forms of regulation and governance frameworks for GHR type collaborations if we are to reap the benefits that GHR can produce for advancing medical research and treatments in ways that secure the public interest. Such governance frameworks should be able to establish checks and balances in regard to the responsibilities and control given to involved parties, including commercial entities, public research institutions and patients. And further, they should be able to take into account who benefits from the use of health data and how, and to determine how value – financial or other – is shared and distributed between involved parties (including "society" at large). No one discipline is fully equipped to do this. Medical ethics, in its current form, is not broad enough to address many of these concerns. It may benefit from incorporating insights from non-medical legal frameworks like anti-trust law. And further, from social science disciplines like critical data studies that draw on a political economy critique to address the development of new big data divides based on access to and ownership of data, technological infrastructures and technical expertise. The first step towards this should be an open conversation will all parties involved about what is at stake in the move towards data-driven, technologically-enabled personalized medicine – what values are served by it, what trade-offs it may entail, what is morally relevant and which final goals and conceptions of the good are worth being pursued.

References

Agniel, D. et al. (2018). "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study". *British Medical Journal* 360: k1479 http://dx.doi.org/10.1136/k1479.

European Commission (EC). (2016). *Open Innovation 2.0*. Brussels: European Commission. https://ec.europa.eu/digital-single-market/en/open-innovation-20.

European Commission (EC) (2018) *Commission Staff Working Document: On Enabling the Digital Transformation of Health and Care in the Digital Single Market; Empowering Citizens and Building a Healthier Society*. https://ec.europa.eu/digital-single-market/en/news/communication-enabling-digital-transformation-health-and-care-digital-single-market-empowering.

European Group on Ethics in Science and New Technologies (EGE) (2018) *Artificial Intelligence, Robotics and 'Autonomous' Systems.* Luxembourg: Publications Office of the European Union.

Esteva, A. et al. (2017). "Dermatologist-level classification of skin cancer with deep neural networks". *Nature*, 542, pp. 115-118.

Farr, C. (2017). "Apple is quietly working on turning your iPhone into the one-stop shop for all your medical info". *CNBC*. Available at https://www.cnbc.com/2017/06/14/apple-iphone-medical-record-integration-plans.html.

Fogel, A. & J. Kvedar. (2018). "Artificial intelligence powers digital medicine". *Nature Digital Medicine*. DOI:10.1038/s41746-017-0012-2.

Gymrek, M., A. L. McGuire, D. Golan, et al. (2013). "Identifying personal genomes by surname inference". *Science*, 339 (6117), pp. 321-324. DOI:10.1126/science.1229566.

Lucivero, F. & B. Prainsack. (2015). "The lifestylisation of healthcare? "Consumer genomics" and mobile health as technologies for healthy lifestyle". *Applied &l Translational Genomics*, 4, pp. 44-49.

Muoio, D. (2018). "Apple dips its toe into employee health with independent care centers". *MobiHealthNews*, 27 February. Available at http://www.mobihealthnews.com/content/apple-dips-its-toe-employee-health-independent-care-centers.

Nissenbaum, H. (2010). *Privacy in Context.* Stanford: Stanford University Press.

Nuffield Council on Bioethics (NCB) (2018) *Artificial Intelligence in Healthcare and Research*. London: Nuffield Council on Bioethics.

Olson, P. (2014). "For Google Fit, our health data could be lucrative". *Forbes*. Available at: http://www.forbes.com/sites/parmyolson/2014/06/26/google-fit-health-data-lucrative/.

Poplin, R. et al. (2018). "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning". *Nature Biomedical Engineering*. DOI: 10.1038/s41551-018-0195-0.

Powles, J. & H. Hodson. (2017). "Google DeepMind and healthcare in an age of algorithms". *Health and Technology*, 7(4), pp. 351-367.

Ram, A. (2018). “DeepMind develops AI to diagnose eye disease”. *Financial Times*, 4 February. Available at https://www.ft.com/content/84fc-c16c-0787-11e8-9650-9c0ad2d7c5b5.

Savage, N. (2015). “Mobile data”. *Nature*, 527 (7576), pp. S12-3.

Sharon, T. (2016). “The Googlization of health research: From disruptive innovation to disruptive ethics”. *Personalized Medicine*, 13 (6), pp. 563-574.

Shen, H. (2015). “Smartphones set to boost large-scale health studies”. *Nature*, 10 March.

Sweeney, L., A. Abu, J. Winn. (2013). “Identifying Participants in the Personal Genome Project by Name”. *Data Privacy Lab, IQSS*, Harvard University. Available at https://privacytools.seas.harvard.edu/publications/identifying-participants-personal-genome-project-name.

The Guardian view on patient data: We need a better approach (2017, 5 July) *The Guardian*. Available at https://www.theguardian.com/commentisfree/2017/jul/05/the-guardian-view-on-patient-data-we-need-a-better-approach.

Verheul, E. & B. Jacobs. (2017). “Polymorphic encryption and pseudonymisation”. *NAW*, 5 (18), pp. 168-172. https://www.cs.ru.nl/B.Jacobs/PAPERS/naw5-2017-18-3-168.pdf.

Verily (2018) https://www.projectbaseline.com.

Vincent, J. (2016). “Apple partners with pharmaceutical giant GlaxoSmith-Kline for first clinical study”. *The Verge*. Available at: https://www.theverge.com/2016/7/18/12211970/apple-GlaxoSmithKline-researchkit-first-study.

Wingfield, N., K. Thomas and R. Abelson. (2018). “Amazon, Berkshire Hathaway and JPMorgan up to try to disrupt health care”. *New York Times*, 30 January 2018. Available at https://www.nytimes.com/2018/01/30/technology/amazon-berkshire-hathaway-jpmorgan-health-care.html.

Wise, J. (2018). “AI system interprets eye scans as accurately as top specialists”. *British Medical Journal*. DOI: https://doi.org/10.1136/bmj.k3484.

Zang, J., K. Dummit, J. Graves, P. Lisker & L. Sweeney. (2015). “Who knows what about me? A survey of behind the scenes personal data sharing to third parties by mobile apps”. *Technol Sci*. Available at: http://techscien

A “right to explanation” for algorithmic decisions?

Paul Vogel

1. Introduction

In discussions about the legal hurdles facing the increased use of artificial intelligence (AI), academic debate has focused on civil and criminal liability for damage caused by AI. At the same time, however, data protection law also creates challenges to the increased use of intelligent systems and machines, i.e. machines which are capable of learning, and these challenges should not be underestimated. As a means of protecting fundamental rights[1], data protection law is used to safeguard each individual’s general right of personality, and in particular, his right to determine what information concerning him is made available or known to parties in his surrounding environment.[2] To this end, the law has in its arsenal procedures, mechanisms and rights, which apply to every processing of personal data within its scope, even if the data is handled not by a human processor but by a self-learning system.

2. The starting point: entry into force of GDPR

An impetus to increased interest in the data protection challenges associated with the use of self-learning systems, was the entry into force of the European Data Protection General Regulation (GDPR) on 24 May 2016. This Regulation has been applicable law in all Member States of the European Union and the European Economic Area since 25 May 2018, following a two-year transitional period. Due to the wide range of organisational duties placed on processors of data, and the significantly increased sanctions available for infringements,

[1] Simitis, Spiros, in idem (Ed.), *Bundesdatenschutzgesetz*, 8th ed. 2014, introduction para. 30: “Datenschutz ist Grundrechtsschutz”.

[2] BVerfGE 65, 1 (43) – *Volkszählung Case*.

society in general, and companies in particular, as processors of data, have become much more conscious both of issues related to the protection of personal data, as well as compliance with statutory data protection duties.

From the perspective of data protection law, the question therefore arises regarding the use of artificial intelligence, of which new or possibly more stringent duties GDPR imposes on AI users.

3. Algorithmic transparency as a challenge for the users of artificial intelligence

3.1. Overview of the problem

Significant reservations about the increasing use of self-learning systems or machines have been directed against the opaqueness of their decision-making. In this regard, reference is made to the "black box nature" of machine-learning systems.[3]

Indeed, by looking at the source code, programmers and experts are able to understand the logic of the system, the algorithm architecture used, and the structure of the databases.[4] However, how individual, concrete decisions are arrived at, is hardly accessible even to experts. This is particularly the case where multi-layered neural networks are used.

Intelligent systems constantly adjust the internal weighting of their variables to the feedback on their decisions during their training processes and also later during regular operations. As a result, factors that were crucially important for a particular decision at time point to can produce a completely different decision at time point t_1.[5]

This may give the person concerned the impression that he is simply an "object" of machine decision-making.[6] If this were the case, these methodologies for processing information would massively infringe on data subjects' general right of personality in German law, as well as on the right to informa-

[3] Wischmeyer, Thomas, "Regulierung intelligenter Systeme" in *Archiv des Öffentlichen Rechts* 143 (2018), 1 (8).

[4] Cf Wischmeyer op.cit. 2018, 47.

[5] Ibid.

[6] Regarding the use of algoriths by the state, cf. Martini, Mario & Nink, David "Wenn Maschinen entscheiden… – vollautomatisierte Verwaltungsverfahren und der Persönlichkeitsschutz", *Neue Zeitschrift für Verwaltungsrecht-Extra* 10/2017, p. 3.

tional self-determination arising therefrom, and would need to be dealt with and appropriately resolved on the basis this law.

3.2. Application scenarios with conflict potential

It is conceivable that a data subject could feel like an "object" in various application scenarios involving artificial intelligence. For example, one could imagine algorithms that assisted the judge in determining the sentence for a convicted person. This could be done by calculating the risk of reoffending by the accused on the basis of large numbers of cases and then making recommendations for the length of the offender's sentence on that basis.[7] The convicted person would normally be very interested in finding out what the specific reasons were for the software suggesting the imposition of a sentence of nine months imprisonment instead of six months, or a fine of 120 days' wages rather than 80.

In contrast, a fully automated decision on a credit application is a much less dramatic event. Of course, where someone's credit application is rejected, it is understandable for the applicant to want to find out why the algorithm classified him as a bad credit risk. The same applies to algorithms that are used by the HR departments of large companies in the pre-selection of job applicants. Such algorithms, for example, independently sort out unsuitable candidates at an upstream level before a member of staff of the human resource department ever gets a chance to look at their written applications.

These examples make it clear that even scenarios common in everyday life are made much more opaque to human participants through the use of learning algorithms. These scenarios make understandable affected persons' unease at the risk of decisions being made "over their heads" that are no longer comprehensible for them.

3.3. A potential solution: a "right to explanation"?

In order to minimize this risk, legal instruments are being sought in academic legal literature to eliminate the dangers resulting from the opacity of algo-

[7] One software package used by American courts is COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). For discussion, cf., for example https://www.theatlantic.com/technology/archive/2016/06/when-algorithms-take-the-stand/489566/ (19/10/2018).

rithmic decision-making. Discussion focuses on the question of whether the current data protection law contains a “right to explanation”. If so, this would oblige operators of learning systems to inform persons affected by the decisions of their systems about how the data processing result was arrived at.[8]

3.3.1. A possible legal basis for a right to explanation

A legal basis for such a right to explanation has been found in various places in GDPR by proponents of such a right. Thus at the beginning of the Regulation, in recital 71 sentence 4, it is stated that persons affected by automated decision making, within the meaning of Art. 22 GDPR, are entitled to an explanation:

> Recital 71, sentence 4, GDPR
>
> “In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.”

Automated decision-making, within the meaning of Art. 22, exists where the data processing is exclusively done by machine, i.e. without any substantive intervention by a human being and this processing is immediately translated into a decision.[9] As an example of purely automated decision-making, recital 71 sentence 1 refers to the “automatic refusal of an online credit application or e-recruiting practices without any human intervention” (cf. also section 3.2 above).

In its section on the rights of data subjects, GDPR contains various points indicating the existence of a right to explanation where decision making has been carried out by an automated system. Thus, for example, Article 13 para. 2(f) of GDPR obliges the data controller to inform the person concerned of the existence of an automated decision-making process within the meaning

[8] For a strong argument in favour of the existence of such a right, cf. Goodman, Bryce & Flaxman, Seth, “European Union regulations on algorithmic decision-making and a ‘right to explanation’”, 2016, arXiv:1606.08813(v3), available at https://arxiv.org/pdf/1606.08813.pdf (19/10/2018).

[9] Herbst, Tobias, “Automatisierte Entscheidungen im Einzelfall einschließlich Profiling” in von Lewinski, Eßer & Kramer (Eds.), DSGVO/BDSG: Datenschutz-Grundverordnung, Bundesdatenschutzgesetz und Nebengesetze, 6th ed. 2018, Art. 22 DSGVO para. 5.

of Art. 22 GDPR. The Regulation also obliges the data controller to provide the person with information about the logic involved in the decision-making.

> **Art. 13 (2) (f) GDPR**
>
> "The controller shall […] provide the data subject with the following further information necessary to ensure fair and transparent processing:
>
> (f) the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."

Similarly worded provisions can also be found in Art. 14 para. 2(g) and Art. 15 para. 1(h) GDPR, which stipulate the duties imposed on data controllers to provide certain information. These provisions similarly state the informational rights of affected data subjects.

These provisions, the rather precise way recital 71(4) was drafted, and the generally very high level of protection of the rights of affected data subjects anchored in GDPR, lend support to arguments in favour of the existence of a comprehensive right to explanation for algorithmic decisions.

3.3.2. Arguments against a right to explanation

Although the wording of GDPR seems to indicate a clear slant in favour of the existence of such a right, cogent arguments do exist against a comprehensive right to explanation.

3.3.2.1. A comparison with the previous law

Firstly, prior to the entry into force of GDPR, provisions with a similar wording already existed. An example of this is Art. 12 (a)(3) of Data Protection Directive 95/46/EC[10], which stated:

> **Article 12 Data Protection Directive – Right of access**
>
> Member States shall guarantee every data subject the right to obtain from the controller:

[10] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ 1995, No. L 281, 31.

> (a) without constraint at reasonable intervals and without excessive delay or expense: […]
>
> knowledge of the logic involved in any automatic processing of data concerning him at least in the case of the automated decisions referred to in Article 15 (1); […]

With regard to a right to explanation, the ECJ chose a cautious interpretation: According to its judgment in joined cases *YS v Minister voor Immigratie, Integratie en Asiel* (C-141/12) and *Minister voor Immigratie, Integratie en Asiel v M, S* (C-372/12), general information in an understandable form from the responsible entity would be sufficient to fulfill the legal duty to provide information.[11] Thus, in the case of algorithmic decisions, the information provided could be limited to a brief statement on the general decision-making structure; in exceptional cases, the logical decision tree would need to be disclosed.[12] A disclosure of the program code or the raw data, however, would not be required.[13] It remains to be seen whether the ECJ will position itself differently with GDPR in force. However, due to the similar drafting of the respective provisions, that may be doubted.

3.3.2.2. A countervailing interest in business secrets, trade secrets and intellectual property

Furthermore, it should be noted that the legislation adopted as GDPR specifically restricts the right to information. Recital 63 sentence 5 states: "That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software." Algorithms are often at a minimum the intellectual property of the programmer or the party contracting to have the code written, but they are often also business secrets.

> Example: *SCHUFA Decision* of the German Federal Supreme Court[14]
>
> In 2014 the German Federal Supreme Court heard a case in which an individual sought to obtain her credit score from the credit agency SCHUFA, after several purchases on credit fell through due to a negative credit rating. From the information provided, according to the plaintiff's view, it was not possible for her to determine why she had received such a low credit score. She therefore sued SCHUFA to obtain detailed information on the basis for the calculation of the credit score. The Court

[11] ECJ judgment of 17/07/2014 para. 50 et seq.

[12] Wischmeyer op.cit. 2018, 50.

[13] Wischmeyer, ibid.

[14] BGH, judgment of 28/01/2014, VI ZR 156/13 = BGHZ 200, 38.

> ruled that SCHUFA did not have to disclose its credit score algorithm because it was a protected trade secret. This precluded enforcement of the right to information under German data protection law.

According to this judgment, the interests of the owner of protected business and trade secrets must be weighed against the interests of the data subject on a case-by-case basis.

3.3.2.3. The factual benefit of comprehensive information

Furthermore, one could well ask the question of what benefit a comprehensive right to explanation would have for the person concerned. "Being able to inspect the actual source code does not generate added value for the vast majority of people."[15] For this reason, it has been proposed that the information, for which there is a duty of disclosure owed, be subject to a certain amount of recasting: the person responsible for providing the information ought to be able to describe the general mode of operation of the algorithmic decision-making process and at least name the factors contributing to the decision. This can be done in a way that this does not conflict with the protection of business or trade secrets.[16] However, the recasting and simplification of information related to the operation of a machine-learning system of inestimable complexity entails the risk that legal violations or discrimination committed by the algorithm will no longer be uncovered or recognized.[17] This would make a right to receive an explanation largely worthless.

3.3.2.4. Interim conclusions

Finally, as a counter-argument, it is often suggested that human decisions can also be like black boxes.[18] Not infrequently they are based on intuition or a gut feeling that even the person deciding either cannot explain or can explain only with difficulty. So, is it fair or expedient to place higher demands on machines than on human beings?

[15] Martini & Nink, op.cit. 2017, p. 11, fn. 108.

[16] Schwartmann, Rolf & Schneider, Adrian in Schwartmann, Rolf, Jaspers, Andreas, Thüsing, Gregor & Kugelmann, Dieter (Eds.), *DS-GVO/BDSG: Datenschutz-Grundverordnung Bundesdatenschutzgesetz* 2018, Art. 13 para. 58.

[17] Wischmeyer op.cit. 2018, 53.

[18] Tutt, Andrew, "An FDA for Algorithms", *Administrative Law Review*, vol. 69 (2017), 83 (103).

All in all, there are good reasons for not extending too far the scope of any right to explanation derived from GDPR. In any case, as a rule, the provision of detailed information about the algorithms used by a self-learning system is usually not required.

4. Regulatory alternatives to a right to explanation

Nevertheless, the problem of the lack of transparency of algorithmic decision-making has to be taken seriously. In addition to a right to explanation, mechanisms of self-regulation and external control may be considered, in order to encourage programmers of self-learning systems to create the highest possible levels of transparency.

4.1. Self-regulation

A promising approach to solving the transparency problem of machine-learning systems is the research area called *Explainable Artificial Intelligence* (XAI), which is currently in development. The goal of this approach to AI is to write algorithms in such a way that they themselves can provide information about the important factors contributing to their decisions. They are able to explain their internal processes in a way that is understandable to the technical layman - the black box should therefore become a "glass box".[19]

Another tool of self-regulation which could be considered would be a Code of Conduct for programmers of self-learning systems. It would oblige programmers subjecting themselves to the code to comply with certain ethical and legal standards. [20] Such a system of voluntary commitment to comply with a code of rules, could be combined with certification measures (which have already been established in the GDPR). As a reward it would be possible, for example, to couple membership in the scheme with reductions in civil liability for data protection breaches.

[19] Holzinger, Andreas, "Explainable AI (ex-AI)", *Informatik-Spektrum* vol. 41 (2018), 138 et seq, (in German).

[20] Cf. the exemplary "Asilomar AI Principles", https://futureoflife.org/ai-principles/ (19/10/2018).

4.2. External control and surveillance

State or state-controlled surveillance mechanisms for algorithms are conceivable as measures of external control. Not infrequently, the idea of an "algorithm TÜV" is raised (TÜV is an acronym which means roughly "technical inspection society". TUVs are private undertakings in Germany which provide safety certification and inspection services[21]). Such a system would constitute a system of independent review of algorithms for legal violations or for database-based discrimination with subsequent certification of compliance given[22]. The extent to which such an analysis could ever be possible for learning systems, given their black box character, ought to be examined more closely.

In order to best take account of the above-mentioned obstacles due to the potential competing legal interest in the protection of trade secrets and intellectual property, one could design an administrative control procedure which utilized an *in-camera* (behind closed doors) mechanism, which ensured that the source code and the specifics of the algorithm under review were only made available to the authorized inspectors and that respect for the duty of confidentiality by those persons was enforced by means of criminal penalties.[23]

5. Conclusions

This short overview of the problematic area of *algorithmic transparency* shows that the opacity of the decision making by self-learning systems stands in a not to be underestimated extent in conflict with currently applicable EU data protection law. This law is characterized by extensive data protection rights, all of which aim to guarantee individuals the greatest possible degree of sovereignty over their personal data. In the case of fully automated decisions - as often happens with the use of self-learning systems - this also includes the guarantee of being told the reasons why the algorithm came to its decision. However, given the levels of technical complexity which particularly exist in deep learning technologies and artificial neural networks, delivering such an

[21] Cf. https://en.wikipedia.org/wiki/Technischer_%C3%9Cberwachungsverein (19/10/2018).

[22] Cf. Martini, Mario, "Transformation der Verwaltung durch Digitalisierung" in *Die öffentliche Verwaltung* 2017, 443 (453).

[23] Wischmeyer, op.cit. 2018, 65.

explanation is anything but trivial. A so-called "right to explanation", as is – rightly – often read into GDPR, therefore quickly shows itself to be unworkable or nearly impossible in practice. There are good reasons why such a right should not be extended too far so as not to unduly hamper the future development of learning algorithms, thereby damaging Europe as a technology location. A number of regulatory alternatives, which to a certain extent would be able to tackle the phenomenon of the lack of transparency of algorithmic decision-making, have been addressed in this paper. We are indeed waiting with eager anticipation to see how the ECJ positions itself on the "right to explanation".

References

Goodman, Bryce/Flaxman, Seth, European Union regulations on algorithmic decision-making and a "right to explanation", 2016, arXiv:1606.08813(v3), available at https://arxiv.org/pdf/1606.08813.pdf.

Holzinger, Andreas, Explainable AI (ex-AI), in: Informatik-Spektrum 41 (2018), pp. 138-143.

Martini, Mario, Transformation der Verwaltung durch Digitalisierung, in: Die Öffentliche Verwaltung (DÖV) 2017, pp. 443-455.

Martini, Mario/Nink, David, Wenn Maschinen entscheiden … – vollautomatisierte Verwaltungsverfahren und der Persönlichkeitsschutz, in: Neue Zeitschrift für Verwaltungsrecht – Extra 10/2017, pp. 1-14.

Schwartmann, Rolf/Jaspers, Andreas/Thüsing, Gregor/Kugelmann, Dieter (Eds.), DS-GVO/BDSG, 2018.

Simitis, Spiros (Ed.), Bundesdatenschutzgesetz, 8th Edition 2014.

Tutt, Andrew, An FDA for Algorithms, Administrative Law Review 69 (2017), pp. 83-123.

Von Lewinski, Kai/Eßer, Martin/Kramer, Philipp (Eds.), Auernhammer, DSGVO/BDSG-Kommentar, 6th Edition 2018.

Wischmeyer, Thomas, Regulierung intelligenter Systeme, in: Archiv des Öffentlichen Rechts (AöR) 143 (2018), pp. 1-66.

Legal aspects of the use of artificial neural networks in diagnostic medical procedures

Nicolas Woltmann

1. Data driven decision making processes

Every human being makes countless decisions over the course of every day, every hour, even every single minute. At some level, absolutely every action we carry out, is, in mathematical terms, the net result of setting off differently weighted goals.[1] Certain environmental conditions must be recognized and calculated into the equation, which vary from situation to situation. This process takes place in an almost identical way in computer systems. Their operations, too, are sometimes the result of highly complex computational processes, which in turn have some overriding purpose and meaning. The more decisions about goals or ways of achieving those goals can be attributed to the actor, and the less they are determined by others, the more likely it is that we are willing to recognise that the author is acting autonomously.[2] Today, machines that are able to do this are referred to as autonomous, such as autonomous cars or robots.

Of course, from a philosophical perspective, there are certain terminological problems with using the term *autonomy* in this context, so that it is important to exercise caution.[3] However, all the way up to EU level, the view has prevailed that reference can be made at least to "technological autonomy", as long as a system possesses a not insignificantly wide scope of action and decision-making.[4] Ultimately, this seems to be justified if one considers

[1] On the importance of goals for the intelligence of an actor, cf. Erhardt & Mona, "Rechtsperson Roboter – Philosophische Grundlagen für den Umgang mit künstliche Intelligenz" in Gleß & Seelmann (eds.), *Intelligente Agenten und das Recht*, 2017, 61 (67).

[2] Christaller & Wehner, "Autonomie der Maschinen – Einführung in die Diskussion" in idem (eds.), *Autonome Maschinen*, 2003, 9 (18).

[3] On this subject, cf. Christaller & Wehner, op. cit 2003, 9 (12 et seq.), who discuss the risk of blurring categories.

[4] Report with recommendations to the Commission on Civil Law rules on Robotics (2015/2103 (INL)).

that the cognitive processes that result in decision-making within a computer system are, in their functioning, not unlike those of a human being. While humans obviously receive information about their environment via their five senses, machines first have to transform sensory data into electronic data in an intermediate step.[5] However, it is relatively easy to deal with this difference by maintaining a clear terminological and methodological differentiation between human and data-driven decision-making processes.

In contrast, the main difference is the sheer number of goals that come into play in human decision making, even when it comes to trivial things. Thus, for example, a person's response to a random event, like encountering someone on the street, may involve all sorts of contradictory motives and issues: Should I speak to them? Do I know the person well or is it a rather fleeting acquaintance? Do I like this person? Is this a person that my spouse possibly cannot stand? Nevertheless, would it be good for my career to be friendly to them? Am I late for an important appointment? Does it look like the person is in a hurry at the moment?

Against this background, the highest form of autonomy seems to me to involve the independent development of new goals. These are aims that did not exist previously in the decision maker's pool of reasons to act. Or to put it another way, it is the ability to learn that makes decisions, based on this process, distinguishable, assessable and therefore ultimately actionable.

2. Even machines can learn

We humans have in a sense a natural talent when it comes to working out new solutions to previously unknown problems. But even computers today are no longer limited to tackling tasks based on a previously defined pool of solutions. Through various different machine learning methodologies, machines are given the necessary flexibility to to solve problems. But how exactly does that work?

[5] According to the broadest definition of data in German criminal law, data are all coded or codable information, regardless of the degree of processing (cf. Lenckner & Eisele in Schönke & Schröder, *Strafgesetzbuch: Kommentar*, 2014, § 202a StGB, para. 3; Heger in Lackner & Kühl *Strafgesetzbuch*, 2018, § 263a StGB, para. 3). A narrower definition is used, for example, in the case of phishing, which is only possible with respect to information which is not publically available, that is stored or transmitted (cf. § 202b StGB read together with § 202a (2) StGB). For an examination of the concept of data from a philosophical point of view, which is worth reading, cf Voß, "Was sind eigentlich Daten?", *LIBREAS. Library Ideas*, 23 (2013), available at https://libreas.eu/ausgabe23/02voss/.

So-called Artificial Neural Networks (ANN) is an approach to machine learning that has actually been known about for some time, but in which great pioneering work has been done in recent years. They consist of a number of layers of individual, coded computing elements (called "neurons" as they are intended to work like nerve cells in the human brain) that interact with each other according to a specific algorithm[6]. The basic idea is to "train" the network with specific [sensory] patterns (symbols, sounds, images, etc.) until it is able to recognize previously unknown inputs of the same kind as "fitting" the learned pattern.

Using this technology, amazing results have been achieved to date, e.g. in face and speech recognition. The methodology is already bringing tremendous advances in modern medical diagnostics. Medical diagnosis is basically nothing other than a conclusion (really a probability statement) following from the recognition of a pattern or "picture" derived from a set of patient signs and symptoms. Various mobile health apps are already making use of this technology: when a patient presents with symptoms, they offer the option of communicating these symptoms in combination with patient characteristics such as age, gender, weight, etc., to a virtual "assistant physician".[7] The computer system in the background then performs a fully automated analysis of the data, and makes a "medical diagnosis". The assistant also states the probability that the diagnosis is correct and what further steps it would recommend to the person concerned (for example "can usually be self-treated" or "must go to physician").

The best example of the utility of neural networks in medicine, however, is tumour diagnostics. Mammography screening has always more or less been performed manually by the doctor (that is, visually examining the images with his own eyes). In contrast, a 2018 study conducted in Hungary concluded that the examination of ultrasound images for irregularities using neural networks provided results just as reliable as those of an experienced oncologist - and that in a fraction of the time necessary for manual analysis.[8]

[6] The way this works is initially determined by the neural network architecture of the neuron with a certain weighting and a stimulation threshold, which must be exceeded for the transmission of signals. The architecture of neural connections changes based on feedback, which the network receives on its performance (so-called backpropagation). For a more detailed explanation, cf. Mueller & Massaron, *Artificial Intelligence for Dummies*, 2018, p. 159 et seq.

[7] See, for example, https://ada.com.

[8] Ribli, Horváth, Unger, Pollner & Csabai, "Detecting and classifying lesions in mammograms with Deep Learning", *Scientific Reports* 2018, vol. 8 (1), p. 1.

According to the authors of the study, in the future the benefits of the technology, in terms of time savings and a reduced probability of errors, could probably be further improved.

Finally, one future scenario that brings the two areas discussed together, takes things to the extreme. There is another recent study on the use of neural networks, this time in the detection of skin cancer, which examined the reliability of the diagnostics performed by a particular software package, and concluded that the results from use of the technology were comparable to the breast cancer screening technology discussed above[9]. So work is now in progress on putting the technology into a mobile health app[10]. Imagine what it would mean if the engineers succeeded in this task. What a revolution it would be if medical laymen could someday be able to carry out skin cancer screening on the sofas in their living rooms using the camera of a smartphone?

3. Putting the concept of legal liability to the test

Revolutions, however, draw their strength from the radicalism of change, and from the uncompromising rejection of all the ways things have been done before. It is therefore part of their very nature to create states of great insecurity that compel us to question former certainties. From a legal point of view, therefore, there is a need to rethink not only detailed rules, but also major concepts.

One area that is particularly affected by this is tort law with its central concept of fault-based liability.[11] For example, in the extreme example of the tumour detection app discussed above, it would certainly be possible to ask the question of how liability for a diagnosis would be allocated among all the participants. Of course, the worst-case scenario would be that a malignant tumour was not recognized as such. Conversely, if a patient is wrongly diag-

[9] von Haenssle, Fink, Schneiderbauer, Toberer, Buhl, Blum, Kaloo, Hadj Hassen, Thomas, Enk & Uhlmann, "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists", *Annals of Oncology* 2018, vol. 29 (8), p. 1836 et seq.

[10] https://www.welt.de/gesundheit/article176835763/Kuenstliche-Intelligenz-erkennt-Hautkrebs-besser-als-Aerzte.html.

[11] Cf. on this Beck, "Technisierung des Menschen – Vermenschlichung der Technik. Neue Herausforderungen für das rechtliche Konzept „Verantwortung" in Gruber, Bung, & Ziemann, *Autonome Automaten: Künstliche Körper und artifizielle Agenten in der technisierten Gesellschaft*, 2015, 173 (174).

nosed as having cancer, his legal interests may also be injured. One need only think of the shock which could be suffered by a person informed that he had a very a serious illness, or the unnecessary further diagnostic procedures or the invasive treatments.[12] In any case, despite all the technological advances, it remains possible for such a system to make mistakes.[13] Infallible computers do not exist today - and probably will not exist in the future – just as infallible people do not exist. However, this alone - from a legal point of view - is no cause for alarm. As long as injury to people's legal interests can be dealt with by means of civil compensation and/or criminal punishment, the law as a whole will continue to retain its regulatory character and thus its legitimacy.

However, this premise is being seriously challenged by the use of artificial neural networks and their increased employment by technological lay persons. In essence, two major problem areas open up, which I will briefly discuss below.

3.1. Autonomy and unpredictability

The first obstacle to finding the cause of a mistake arises from the operation of ANNs. This involves a largely independent learning process which provides the necessary flexibility for recognizing previously unknown patterns of information. Such processes have a level of complexity that is sometimes hardly comprehensible in human terms. Even the designer of an ANN can no longer properly monitor or at all steer the interactions of the individual nodes of a system after a certain point in time. The slightest disturbance of the pattern recognition processes can even produce completely unexpected results.[14]

In case of malfunctions, it may therefore be very difficult to find the real cause. It can have its possible origin in any of the steps between designing the software and obtaining the results of the diagnostic procedure: it is possible that the manufacturer programmed the network incorrectly. On the other hand, it is not improbable that the patient caused the error by not complying with instructions for the new application. The most frequent cause of getting

[12] Gaßner & Strömer, "Mobile Health Applications- haftungsrechtlicher Standard und das Laisser-faire des Gesetzgebers" in *Versicherungsrecht* 2015, 1219 (1227).

[13] The software reported on by von Haenssle et al, op.cit. 2018 (cf. FN 9), correctly identified around 95% of the tumors in need of treatment and approx. 63% of the harmless moles.

[14] Monpetain, *Neuronale-Netze eine Einführung*, 1998, available at: http://www.weblearn.hs-bremen.de/risse/RST/SS98/NEURAL_N/NEURONET.HTM.

erroneous results from ANNs, however, is probably the partial unsuitability of the data set with which the network was previously trained. Data set quality is the decisive factor for the proper operation of an ANN![15] It is crucially important that patterns - in our case melanoma images – are used, from which the system can extract the information relevant for correct diagnoses. With sometimes millions of individual sample images that have to be fed into an ANN during the network's training, it is hardly possible to identify those which, in conjunction with the complex structure and workings of the network, have led to false diagnoses. This represents a huge obstacle to answering the question of what caused the error in respect of a particular diagnosis.

Rather - and this is perhaps even the more important point - for the reasons given above, even if the error could be traced, it is by no means always possible to establish any negligent or culpable behaviour by a relevant party. This is because it can be extremely difficult in individual cases to predict how the system will behave in the face of an array of unknown or unfamiliar images. However, such predictability is exactly one of the essential criteria that make it possible to attribute causation of damage to a person's behaviour in the first place.[16] In connection with such cases, courts may at some point come to the conclusion that the constructors of an ANN have attempted to the best of their knowledge and belief to prevent the concrete error that has taken place; that in view of the sheer mass of training data, they had no way of preventing an unfortunate misdiagnosis in an individual case; that the best was done which could be done given the current state of science and technology. So in the world of artificial intelligence, a determination of the facts of a case does not automatically mean that it is possible to allocate legal responsibility for a particular failure.

This is less of a problem from a civil law perspective than from a criminal law one. That is because for some time now there have been possible solutions that are not only under discussion but to a certain extent are also considered by many to be practicable. The ePerson, for example, is one such a legal construct that could be used to fairly allocate the financial risks of using autonomous systems among those involved.[17] Furthermore, there is also

[15] Mueller & Massaron, op.cit. 2018, p. 27 et seq.

[16] Fischer, *Strafgesetzbuch mit Nebengesetzen*, 2018, § 15 StGB, para. 17 et seq.

[17] On this cf. Schweighofer, *Auf dem Weg zur ePerson: aktuelle Fragestellungen der Rechtsinformatik*, Schriftenreihe Rechtsinformatik Wien 2001; Mayinger, *Die künstliche Person: Untersuchung rechtlicher Veränderungen durch die Installation von Softwareagenten im Rahmen von Industrie 4.0, unter besonderer Berücksichtigung des Datenschutzrechts*, 2017.

discussion about the introduction of strict liability for risks arising from the use of artificial intelligence.[18] In Germany, this would be relatively easy and straightforward for Parliament to do using the current law on civil liability in road traffic accidents as a model - provided one could agree on the conditions under which computers and machines would be covered by it. In any case, it would seem to be possible to avoid the evidential difficulties and the uncontrollability of autonomous technology by pre-defining areas of tort liability within which liability for damage would be independent of individual fault.

For a patient whose treatment for cancer was delayed because of a misdiagnosis by a computer system, receiving monetary compensation may be a very small consolation indeed. We live in a society where many people tend to judge such cases according to their own personal subjective feelings, and not according to objective standards. They sometimes get satisfaction for injustice suffered only by seeing the offender convicted and punished confusing this with justice. That includes - in extreme cases - identification of an "offender", or at least a full investigation of the facts by courts. But what if that is not possible? In the context of criminal liability, the possible solutions described above are completely unworkable, because in this legal system a person cannot be convicted of a crime if he did not possess the necessary mens rea at the moment of the offence: strict liability offences do not exist in Germany.

In my view, the biggest legal challenge facing the law from digitization is therefore to reconcile the (often legitimate) expectations of victims for criminal product liability sanctions to be enforced against wrongdoers with the fundamental principles of criminal liability in a state under the rule of law.[19] We will have to wait and see whether the realization that intelligent computer systems are likely to make our lives safer and more enjoyable is sufficient to reach a consensus on new norms.[20]

[18] Borges, "Haftung für selbstfahrende Autos", *Computer und Recht* 2016, 272 (279 et seq.); Bräutigam & Klindt, " Industrie 4.0, das Internet der Dinge und das Recht", *Neue Juristische Wochenzeitschrift* 2015, 1137 (1139); Spindler, "Roboter, Automation, künstliche Intelligenz, selbst-steuernde Kfz – Braucht das Recht neue Haftungskategorien", *Computer und Recht* 2015, 766 (775).

[19] Although the civil liability issue is probably more salient, given the sheer amount of litigation one would expect.

[20] In this context, reference is sometimes made to the legal concept of accepted risk, cf. for example Hilgendorf "Autonomes Fahren im Dilemma. Überlegungen zur moralischen und rechtlichen Behandlung von selbsttätigen Kollisionsvermeidessysteme" in idem (ed.), *Autonome Systeme und neue Mobilität: Ausgewählte Beiträge zur 3. und 4. Würzburger Tagung zum Technikrecht*, 143 (164).

4. A hierarchy of expertise?

The autonomy of an artificial neural network is a problem, especially from the point of view of the programmer, because it is he who nevertheless has to guarantee its proper functioning. Another, subordinate question is the weight of a medical diagnosis of an ANN compared with that of a physician. Ultimately, one must ask two contradictory but crucial questions:

(1) Can I rely on a diagnosis made by the system?

The essence of the problem may not be obvious from the way the question has been worded, but on closer inspection becomes increasingly apparent: If the results of the studies referred to above can be confirmed in the broad field of clinical practice, if the use of such systems becomes the medical standard, will an independent review by a human expert be necessary at all? Our first reaction (especially from the legal perspective of the professional sceptic): Yes. The system has up to now shown itself neither to have reached the necessary level of reliability in empirical scientific studies, nor has it attained a necessary level of acceptance by patients as a social reality. Having said that, it does not seem appropriate for us to place blind faith in a form of technology whose operations we ourselves are no longer able to understand even with the help of experts in the field.

Better always two diagnoses, one human and one data-driven? In fact, it is not so easy, because such an approach inevitably leads to having to give priority to one of two contradictory "medical" judgments from different sources, [should the two diagnoses not agree]. But when does the power of the computer force the doctor to cast aside his own diagnosis? Or asked from his point of view:

(2) Must I accept the diagnosis given by the system?

Does the Hippocratic Oath not oblige the physician to use such a device for the benefit of his patient - not because it has been shown to work better than this physician himself, but because on average it gives more reliable results when compared to the profession of physicians taken all together? I'm not talking about the fear that digitization could cost even highly qualified academics and thus all of us our jobs. Above all, this has nothing to do with injuring the pride of practicing physicians. The crucial point is a [possible] shift in who has the final say in relation to issues relating to life and death.

Of course, from the point of view of the doctor, this is accompanied by (at least a perceived) loss of status and power. The patient will see it very differently: he is now able to take care of himself far removed from over-filled waiting rooms and stressed-out doctors; he is getting new freedom. Some people may already be hailing it as the democratization of the health care system.[21] This development, however, will not be without problems.

We have been discussing at such length the autonomy of machines that we seem to have forgotten that patients are also supposed to have a certain autonomy. The well informed patient should be free to choose what type of treatment he receives. Medical law has had something to say about this principle far longer than about any (supposed) autonomy of computer systems.

Recently, however, it does not seem to be so easy to separate the one from the other. That is because in our hypothetical case, the owner of the tumour detection app depends almost entirely on the "diagnostic ability" of the software to give him the information he needs to exercise his decision-making power. At the same time, the patient bears the increased weight of responsibility for very important decisions about his own health: ought he better go to the doctor even though his cell phone app tells him that most likely everything is alright? Or should he again waste hours of his precious time in the doctor's waiting room, just like the last two times, when his phone recommended that he seek medical advice? Bringing an additional player into the diagnostic process does not always make decisions easier or simpler. From a legal perspective, exactly the opposite should be expected, because it increases the risk of blind trust in a computer application.

5. Conclusions

At the end of this short presentation, many uncertainties remain. By and large, they culminate in the question: how far do we want to give up our own decision making power and place it in the hands of machines, knowing that we are dealing with entities that cannot be completely controlled? What meaning and purpose does human judgment – regardless of whether it is based on hard-won professional expertise or a gut feeling – have in today's

[21] Uhlig, Ittel & Marx, "Die Zukunft der digitalen Gesundheitsversorgung in Deutschland" in *DIV Report Spezial: Digitale Gesundheit* 2017, 24 (25).

world? The current applicable law does not contain the answer. Engineers don't have it either. What is needed is an interdisciplinary dialogue, which moral philosophy and ethics also have parts in. It is only in this way that a concept can be developed which Parliament can enact into a binding framework of rules which establish a uniform quality standard for the integration of such applications into the provision of health care.

References

Beck, Susanne: "Technisierung des Menschen – Vermenschlichung der Technik. Neue Herausforderungen für das rechtliche Konzept „Verantwortung" In: Gruber, Malte, Bung, Jochen & Ziemann, Sascha (ed.): *Autonome Automaten: Künstliche Körper und artifizielle Agenten in der technisierten Gesellschaft*, Berlin 2015, pp. 173-188.

Borges, Georg: "Haftung für selbstfahrende Autos", *Computer und Recht* 2016, pp. 272-280.

Bräutigam, Peter & Klindt, Thomas: " Industrie 4.0, das Internet der Dinge und das Recht", *Neue Juristische Wochenzeitschrift* 2015, pp. 1137-1142.

Christaller, Thomas & Wehner, Herbert: "Autonomie der Maschinen – Einführung in die Diskussion" In: Christaller, Thomas & Wehner, Herbert (eds.), *Autonome Maschinen*, Wiesbaden 2003, pp. 9-XX.

Erhardt, Jonathan & Mona, Martino, "Rechtsperson Roboter – Philosophische Grundlagen für den Umgang mit künstlicher Intelligenz" In: Gleß, Sabine & Seelmann, Kurt (eds.), *Intelligente Agenten und das Recht*, Baden-Baden 2016, pp. 61-93.

European Parliament, Report from 01/27/2015 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). http://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html.

Fischer, Thomas: Strafgesetzbuch mit Nebengesetzen, 65th ed. Munich 2018.

Gaßner, Maximilian & Strömer, Jens M.: "Mobile Health Applications- haftungsrechtlicher Standard und das Laisser-faire des Gesetzgebers" in *Versicherungsrecht* 2015, pp. 1219-1228.

von Haenssle H.A., Fink C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum A., Kaloo, A., Hadj Hassen A.B., Thomas L., Enk, A. & Uhlmann, L., "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists", Annals of Oncology 2018, vol. 29 (8), pp. 1836-1842.

Hilgendorf, Eric: "Autonomes Fahren im Dilemma. Überlegungen zur moralischen und rechtlichen Behandlung von selbsttätigen Kollisionsvermeidessysteme" In: Hilgendorf, Eric (ed.), *Autonome Systeme und neue Mobilität: Ausgewählte Beiträge zur 3. und 4. Würzburger Tagung zum Technikrecht*, pp. 143-175.

Lackner, Karl & Kühl, Kristian: Strafgesetzbuch Kommentar, 29th ed. Munich 2018.

Mayinger, Samantha Maria: Die künstliche Person: Untersuchung rechtlicher Veränderungen durch die Installation von Softwareagenten im Rahmen von Industrie 4.0, unter besonderer Berücksichtigung des Datenschutzrechts, Frankfurt a.M. 2017.

Monpetain, Klaus: *Neuronale-Netze eine Einführung,* 1998. http://www.weblearn.hs-bremen.de/risse/RST/SS98/NEURAL_N/NEURONET.HTM.

Mueller, John Paul & Massaron, Luca: Artificial Intelligence for Dummies, Munich 2018.

Parsch, Stefan, "Wie Computer beim Kampf gegen Hautkrebs helfen". https://www.welt.de/gesundheit/article176835763/Kuenstliche-Intelligenz-erkennt-Hautkrebs-besser-als-Aerzte.html.

Ribli, D., Horváth, A., Unger, Z. et al. Detecting and classifying lesions in mammograms with Deep Learning. Sci Rep 8, 4165 (2018) doi:10.1038/s41598-018-22437-z.

Schönke, Adolf & Schröder, Horst: Strafgesetzbuch Kommentar, 29th ed. Munich 2018.

Schweighofer, Erich, Menzel, Thomas, & Kreuzbauer, Günther (ed.): Auf dem Weg zur ePerson. Aktuelle Fragestellungen der Rechtsinformatik, Wien 2001.

Spindler, Gerald: "Roboter, Automation, künstliche Intelligenz, selbst-steuernde Kfz – Braucht das Recht neue Haftungskategorien?", *Computer und Recht* 2015, pp. 766-776.

Uhlig, Stefan, Ittel, Thomas & Marx, Gernot: "Die Zukunft der digitalen Gesundheitsversorgung in Deutschland" in *DIV Report Spezial: Digitale Gesundheit* 2017. https://div-report.de/gesundheit-2017/die-zukunft-der-digitalen-gesundheitsversorgung-in-deutschland/.

Voß, Jakob: Was sind eigentlich Daten?, LIBREAS. Library Ideas, #23 (2013). https://libreas.eu/ausgabe23/02voss/.

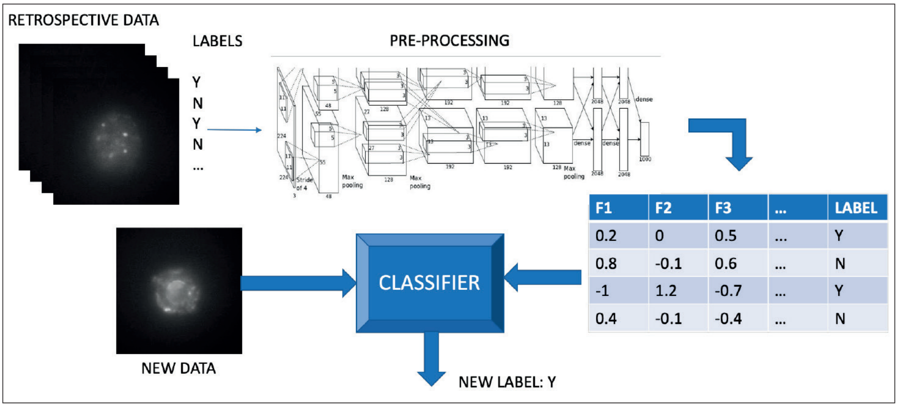
Data-driven and knowledge-driven decision-making in clinical medicine: the necessary approach

Riccardo Bellazzi, Francesca Bellazzi

1. Introduction

Recently, an increasing number of papers have shown that it is nowadays possible to automatically extract decision rules from large clinical data sets and that those decision rules may outperform experts' diagnostic or prognostic capabilities [1,2,3]. The analytical pipelines adopted by these research works are pretty similar. A large number of retrospective cases are collected. They usually contain clinical, textual and image data and they are labeled, i.e. each case is associated to an outcome, such as for example disease/no disease. The data belonging to each case are pre-processed and transformed into a suitable numeric representation, i.e. a feature vector. The feature vector is then used to associate the case to its corresponding label, i.e. the case is classified. Both the pre-processing step and the classification one are "learned" from the available data by resorting to a "so-called" machine learning algorithm. Such algorithms scrutinize the associations between the features and the labels of the retrospective cases, and they propose a suitable generalization that allows associating future unseen cases to a label. The algorithms are usually learned on a set of cases (training cases) and their performance is tested on unseen cases (test set). The implementation of such pipelines has been facilitated in the recent years by two concurring factors: i) the availability of large digital data collections; ii) the availability of algorithms able to properly, in a fast and smart way, pre-process images, text and data streams. Some of the pre-processing algorithms for dealing with images perform so well that the model induced on one data set (say a large data repository of generic images, like trees and houses) can be successfully used to pre-process other data sets of biomedical images without adaptation [4].

Those success stories, which have led FDA to start a discussion about regulatory aspects of AI-based software in medicine [5] has made the debate about the role of data-driven decision-making in medicine of fundamen-



F1	F2	F3	...	LABEL
0.2	0	0.5	...	Y
0.8	-0.1	0.6	...	N
-1	1.2	-0.7	...	Y
0.4	-0.1	-0.4	...	N

Figure 1. The machine learning pipeline.

tal importance. In this brief essay, we will discuss the limitations of current data-driven AI methods, and clarify that the coupling of data-driven and knowledge-driven approaches is not merely desirable but rather is the only possible approach to deal with automated decision-making in biomedicine.

2. Classification

Most of the approaches presented in the literature as example of data-driven decision-making actually deal with statistical classification. Statistical classification is "the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known" [6]. This type of decision problem is very common in practice, covering a large number of atomic decisions, including diagnostic ones. It is thus important to dissect the components of such data-driven decision-making algorithms. First of all, by definition, they depend on data. The basic assumptions made is that it is possible to infer a decision surface in the vector space of features that is able to discriminate between the labels. For this reason, these methods need: i) "enough" data; ii) "good" data; iii) the "right" features.

Data are "enough" if they are able to properly describe the statistical properties of the data for the purpose of classification. Figure 2 shows two examples with the same number of data: in case a) data are not enough to describe the classification boundary, while in case b) it is possible to infer the decision boundary from them.

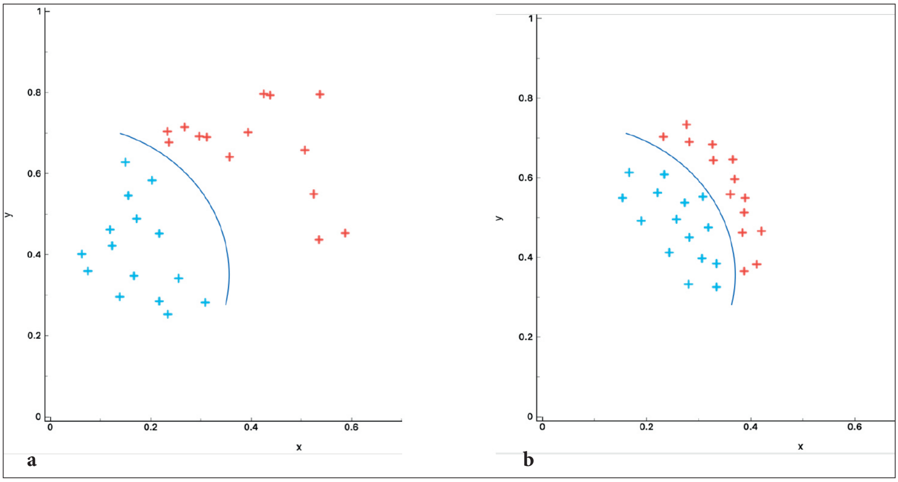


Figure 2. a) data are sparse and red points are not able to describe the decision boundary (blue line); b) data are collected along the decision boundary, and it is possible to infer the separation between the red and the blue area from the data.

Data are "good" if they are not noisy, i.e. have been correctly recorded, labeled, and they correspond to the statistical assumptions of the machine learning models. For example, the data sample can be assumed to be independent from each other. There are many confounders that may hamper a correct data collection, say for example shift of the personnel collecting the data (on Monday nurse A, on Tuesday nurse B, on Wednesday nurse A again, and each nurse annotating in a slightly different way eating habits of the patients).

Finally, features must be the "right" ones. "Right" features are the proper variables that can effectively help in discriminating between cases. This is not only related to the choice of the features, i.e. it is probably not useful to use Hemoglobin to diagnose hypothyroidism, but it also related to the correct interpretation of the meaning of the collected features, i.e. it is important to understand what is the real meaning of "sedentary" lifestyle if the analyzed data are self-reported by patients.

As it is shown is this simple example, data-driven decision-making programs have to deal with a first problem: data are never pure because they need a theoretical framework of evaluation. Indeed, their "being sufficient for statistical purposes", their "being good" and their "being the right ones" is necessarily evaluated in respect to the purpose of the research in question. The purpose of the scientists and data-collectors influence in a way that is unneglectable the collection of data themselves. Many studies from the 1980s have been directed towards the understanding and evaluation of the role of data in scientific theories and processes, and it is now wildly accepted that scientific processes that involve data presuppose theoretical frameworks [7]. These phenomena are called by the scholars the "theory-ladeness" of data, i.e. data always presuppose theoretical frameworks and the "theory-driven" of data, i.e. the collection of data is guided by the final purpose of the theoretical model, i.e. it has to confirm or disconfirm a particular hypothesis.

3. The nature of clinical data collection

Understanding the data that are used to "learn" a classification system is probably the most critical aspect of the entire machine learning pipeline. As reported by G. Hripcsak [8], clinical records are always the result of the "process of care", and, thus, they are not only theory-laden and theory-driven but they are also strongly influenced by the context in which they are collected. In particular, their collection, in terms of types of data, representation, frequency, depends on: i) the health care organization, ii) the reasoning process related to the specific patients' case. This has important implications in term of the classification rules we are learning with AI algorithms. It shows that data collection is driven by two factors: (i) the practical context, i.e. hospitals and health care organisations, (i) the whole theoretical models applied both accordingly to the reasoning processes used and the final purpose of the research, i.e. a specific diagnosis that should be confirmed or disconfirmed. In these cases, it seems legitimated to ask: are they really data-driven, or are they driven by the organizational model that is behind their collection? Is a classification model learned on a specific data set transferable to other contexts that use different data collection processes?

The natural conclusion is that before learning a classification system, it is necessary to deeply model the data collection process and to clarify what is the ultimate goal of the classification rules, i.e. what is the theory that is behind the collection of data and what is the theoretical purpose of it. In other words, it is essential to understand the context of the decision-making process, and to elicit the knowledge that grounds the entire decision problem.

4. Preferences

An often-neglected aspect of classification systems lies in the set of preferences they use when proposing decisions. Statistical learning algorithms usually estimate the probability of each class/label given the data collected on the case-at-hand. Formally, denoted with X the vector of the observed data, c the class/label, the classifiers compute for each c in the set of possible classes C the conditional probability P(c|X). If we have a two-class problem, with C = {c1, c2} the classifier computes P(c1|X) and P(c2|X). In order to provide a decision, a data-driven classification system needs to provide a rule that transforms these probabilities into a decision D. A straightforward rule is to assign the class with the highest probability to D, i.e. D=c1 iff P(c1|X) > P(c2|X) and D=c2 viceversa (in case of ties a random choice is provided). However, this assumes that the error has the same cost in all cases: taking D=c1 when the true class is c2 has the same cost/loss of taking D=c2 when the true class is c1. In this case, the probability decision threshold of a binary class decision problem is 0.5, i.e. D=c1 iff P(c1|X)>0.5. However, this case seldomly happens in practice. In Emergency rooms, clinicians (and certainly citizens) very often prefer to have false positives (people that undergo clinical exams that are actually healthy) than false negatives (people that go home without exams that are actually ill). This has impact on the decision algorithm, and the decision threshold is accordingly different (if c1 means “disease” then P(c1|X) will be lower than 0.5). The final impact is that not only data depends on the actual organization of care, but also the decision algorithm is inherently dependent on the preferences or values of the final decision maker.

5. Reasoning and taking decisions

One of the main fears of AI in medicine researchers is the risk of another “AI-winter”. AI-winter refers to two periods of disillusion that came after the rise of Neural Networks in the ‘70s and of expert systems in the ‘80s [9]. This will happen again in the future if there will be again a lack of understanding of what are the limits of current systems, and, referring to automated decision-making, the limits of data-driven decision-making. As it should seem clear from previous observations, in order to correctly posit data-driven methods it is useful to refer to a broader epistemological description of scientific inference.

As summarized in the so-called Generate-and-Test paradigm [10], experts involved in decision-making may use iteratively two inferential steps to formulate their decisions: i) a *hypotheses selection* phase, in which given the information available "at large" (clinical context, prior knowledge, actual observations) is used to generate a set of candidate hypotheses, and ii) a *hypotheses testing* phase, in which hypotheses selected in the previous phase are assessed by forecasting their expected consequences, which can be further matched with future observations. As firstly presented by Peirce [11] and as it has been developed in contemporary philosophy of science, this corresponds to the so-called "abductive reasoning", that reflects the "inference to the best explanation" method (IBE). In IBE a hypothesis is not simply confirmed by the data (the truth of the data does not imply necessarily the truth of the hypothesis), but the hypothesis is evaluated as the right one because of the explanatory power of it (the truth of the data is best explained in the light of the hypothesis) and this presupposes the theoretical framework developed in (i). This type of reasoning is used in science in order to avoid a part of the famous "inference-problem", that grounds on the logical fallacy known as "affirming the consequent": if it is known that A implies B, and B is observed to be true, then it can be hypothesized that A is true. Since B may be true because of other causes, this inference may be wrong and thus is not valid [12,13].

In [14], Ramoni and colleagues well described the process of abductive reasoning in the context of clinical diagnosis and formalized it in the light of automated reasoning in a model called ST (Select and Test) Model.

The ST-model is based on an iterative sequence of elementary inferential steps. Each step in the model can be seen as a specific inference type, as reported in Figure 3. First, an *abstraction step* is needed. Abstraction allows extracting high-level features from the initial data and information and corresponds to the definition of the feature sets and of one or more outcome variables that can be associated to the decision-making problem. Abstraction is followed by an *abduction* step, in which the features are used and potentially combined to specify more precisely the decision problem and make one or more hypotheses, each of which is a potential explanation for the observed data. This may correspond to one or more classification problems, which can be analyzed resorting to the machine learning algorithms here described. These algorithms can be used to *rank* the hypothesis by estimating their probability. Moreover, competing hypothesis space can be analyzed and ordered also taking into account users' preference criteria. The hypothesis needs to be further scrutinized by a *deduction* step that derives a set of consequences from the hypothesis, either by logical inference or simulation. This step uses existing models of cause-effect relationships. Finally, predictions are matched against the available data in an *induction* step: hypotheses whose consequenc-

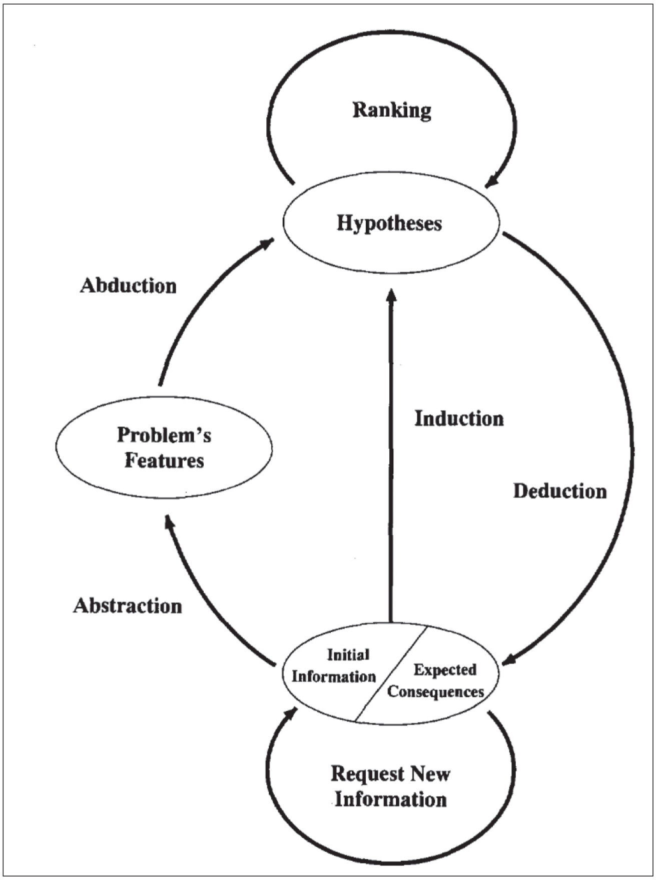


Figure 3. An epistemological model of diagnostic reasoning. Data driven decision-making can be easily mapped to the ranking phase (from [10]).

es match the available data are kept, while those that contradict the available data are eliminated. The sequence of steps can be iterated several times, involving humans and machines. This account of diagnostic reasoning clarifies that machine learning is a tool that can be helpful in some of the reasoning steps, but that it neither can run in a "stand-alone" nor it is sufficient to design, implement and deploy a (semi-) automated decision support system.

As a matter of fact, a fundamental problem of a "data-driven" system of classification is that it deals only with one of the elements of a far more complicated process that should be used in diagnosis systems. The insufficiency of these data driven systems leads us to consider a second problem: the so-called thesis of empirical equivalence (EE) [15]. EE states that two contradictory and different hypotheses can be supported by the same set of empirical data. This means that without the whole theoretical framework of an epistemological model, such as the ST-model, the same data can support two contradictory diagnoses and thus it can be not only wrong, but also potentially meaningless for the purposes expected. Without the implementation of knowledge-driven system, it can be possible to support all different diagnosis from the same set of empirical data.

Data representation frameworks, including ontologies and clear semantics, are needed, as well as models of health care organizations, models of cause-effect relationships, models for performing simulations and, finally, "blackboard" type systems to annotate the current state of reasoning.

6. Conclusions

The recent widespread diffusion of machine learning and artificial intelligence methods to deal with the analysis and classification of multi-modal data, i.e. images, text, coded information, speech and data streams, have increased the attention towards the implementation of "data-driven" decision support systems in medicine. Those systems are supposed to automatically and autonomously learn decision rules on the basis of the available data, in particular when the data collection become "large" enough to be able to represent the problem domain. Some papers have also shown that in specific contexts machine learning systems have decision-making performances that are comparable or superior to those of human experts. In this paper, starting from these results, we have focused our attention to the characteristics of the classification problem, which is actually the task that current machine learning methods are able to deal with. In particular, we have identified two main problems: the theory-ladeness of data, i.e. data are never "pure", but they are always collected and analyzed in theoretical and practical frameworks; the empirical equivalence of scientific hypothesis. First, we have shown that it is crucial to fully understand that machine learning algorithms require enough data, data of good quality and the right features to properly run. These three elements are only possible if there is a deep knowledge about the data collection and data interpretation processes because and they can only be present in the light of theoretical evaluations (data are theory-laden and theory-driven). Moreover, data generation is always related to a concrete health care organization model, which runs behind the scene in the everyday activities of clinicians and practitioners. Furthermore, any decision is related to preferences and values of users, and there is no system that do not use these values to translate predictions into decisions. Finally, classification is only a component of the larger picture of reasoning, which requires proper epistemological and ontological modeling in order to be translated into a computational system. This is due to the empirical equivalence thesis, for which the same set of data can support contradictory and different hypothesis. Thus, to drive relevant and useful conclusions from data, we need a broader and more complex model of decision making.

As a conclusion, we can state that a data-driven decision support system cannot exist on its own. The design of semi-automated systems for decision-making needs rather the merging of knowledge-driven methods with data-driven algorithms. The careful conjunction of these components is the only suitable way to learn from the past and to soundly and properly design the AI systems of the future, in particular in a safety-critical context as medicine and clinical care [16].

Acknowledgements: The research for this paper was partly supported by the European Research Council Project 'The Metaphysical Unity of Science' (ERC CoG), grant no. 771509.

References

[1] Gulshan, V., L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P.C. Nelson, J.L. Mega & D. R. Webster. (2016). "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs". *JAMA*. 2016 Dec 13, 316(22), pp. 2402-2410.

[2] Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau & S. Thrun. "Dermatologist-level classification of skin cancer with deep neural networks". *Nature*. 2017 Feb 2, 542 (7639), pp. 115-118.

[3] Hannun, A.Y., P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia & A.Y. Ng. (2019). "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network". *Nat Med*. 2019 Jan, 25(1), pp. 65-69.

[4] Kensert, A., P.J. Harrison & O. Spjuth. (2019). "Transfer Learning with Deep Convolutional Neural Networks for Classifying Cellular Morphological Changes". *SLAS Discov*. 2019 Apr, 24(4), pp. 466-475.

[5]https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/Softwareasa-MedicalDevice/UCM635052.pdf

[6] https://en.wikipedia.org/wiki/Statistical_classification

[7] Brewer, W.F. & B. L. Lambert. (2001). "The Theory-Ladenness of Observation and the Theory-Ladenness of the Rest of the Scientific Process". *Philosophy of Science, Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers (Sep., 2001)*, 68 (3), pp. S176-S186

[8] Hripcsak, G. (2015). "Physics of the Medical Record: Handling Time in Health Record Studies". In: Holmes J., R. Bellazzi, L. Sacchi & N. Peek (eds.), *Artificial Intelligence in Medicine. AIME 2015. Lecture Notes in Computer Science* (Vol. 9105). Springer.

[9] Crevier, D. (1993). *AI: The Tumultuous Search for Artificial Intelligence*. New York, NY: BasicBooks, ISBN 0-465-02997-3

[10] Simon, H.A. (1977). *Models of Discovery: And Other Topics in the Methods of Science*. Springer.

[11] Peirce, C. & B. Buchler. (1955). "Abduction and Induction". In: *Philosophical writings of Peirce*, pp. 150-156. Dover Publications.

[12] Douven, I. (2017). "Abduction". *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/abduction/

[13] Douven, I. (2017). "Pierce on Abduction". *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/abduction/peirce.html

[14] Ramoni, M., M. Stefanelli, L. Magnani, G. Barosi. (1992) "An epistemological framework for medical knowledge-based systems". *IEEE Transactions on Systems, Man and Cybernetics*, 22, 1375, 1361.

[15] Laudan L. & J. Leplin J. (1991). "Empirical Equivalence and Underdetermination". *The Journal of Philosophy*, 88 (9), pp. 449-472.

[16] Fox, J. & Thomson R. (2002). "Clinical decision support systems: a discussion of quality, safety and legal liability issues". *Proc AMIA Symp*., pp. 265-269.

The Strange case of Dr. Watson: liability implications of evidence-based decision support systems in the health care

Francesca Lagioia, Giuseppe Contissa

1. Introduction

The ageing of population is becoming one of the most significant phenomena of the XXI century. Over the past decades, life expectancy is significantly increased. The 12% of the world population is currently over the age of 60 and by 2050 this percentage should rise to 21% [17]. While this is a huge triumph for modern science and medicine, it poses huge pressures for the provision of health care services due to the increasing costs and the inexorably decrease of the medical personnel rate compared to the number of patients [17]. The advent of the big data and AI era is usually considered as part of the solution. The increased focus on the prevention of medical errors coupled with the introduction of clinical decision support systems have been proposed as key factors to improve the health care quality and patient safety [12]. The adoption of clinical decision support systems (CDSS) for medical diagnosis and treatment should also facilitate the evidence-based practice, which is regarded as the gold standard of decision making in health care [20].

In this context, the IBM Watson system is one of the most promising AI technology developed over the past years.

Initially designed for competing with human champions at the Jeopardy! quizz show, Watson is currently experimented in the health care as an evidence-based CDSS. It is based on the DeepQA technology, which exploits natural language processing and a variety of search techniques for analyzing both structured and unstructured information. The DeepQA is trained on a set of documents, where human experts annotate all the instances of pairs of questions and answers. Thus, the system learns how to identify and correlate questions and answers and applies this knowledge to analyse new input questions and generate new possible candidate answers through a broad search on massive volumes of information, that have never been annotated. For each candidate answer a new hypothesis is generated. Then, for each hypothesis, the DeepQA tries to find evidence supporting or refuting it. The process results in a ranked list of candidate answers, i.e. diagnosis, with a specific confidence score.

This paper investigates some legal issues emerging from the adoption of Watson and similar AI systems in the medical area. In so doing, we explore a set of questions whose answers may heavily affect the liability allocation in misdiagnosis and/or improper treatment scenarios. Section 2 explores what are the distinctive features of new generation CDSS compared to the traditional ones. Section 3 investigates whether and to what extent such features pose questions with regard to the source of the decision making authority. Section 4 deals with the legal qualification and the conformity assessment procedure of these new AI CDSS, under the European discipline of Medical Device Software. In particular, we shall evaluate whether additional criterium for the classification of those systems are needed and how they can influence the certification procedures and the medical liability. Finally, section 5 explores how and to what extent the level of automation may affect the liability allocation. To this end, in section 6 we design some scenarios, which provide variations on the possible causes of failure in the decision making process and the consequent liability assessment.

2. Dr. Watson vs traditional clinical decision support systems

We identify three main features that distinguish Watson from traditional expert systems: (1) the data driven approach, (2) the unpredictability by design, and (3) the stronger impact on the decision making process.

The first feature pertains the widespread adoption of data-driven methods in the AI research and development, which are gradually replacing the traditional knowledge-based approach in specific domains of application. Traditional decision support systems are computer based information systems that use expert knowledge to attain high-level decision performance, in a narrow problem domain. Human expertise has to be elicited and represented symbolically. In particular, symbolic reasoning mechanisms are based on algorithms to make inferences out of the knowledge base, using forward (from data to conclusion) and backward (from conclusion to data) chaining [1]. Such traditional expert-systems are typically based on classical procedural algorithms. Examples of these systems are MYCIN [23] and ONCOCIN [24], both developed at the Stanford University in the early 1980s. However, in the last decade, the focus has shifted on the possibility of applying machine learning algorithms to enormous amount of data.

The data driven AI systems, as well as Watson, are based on big data analytics and data mining techniques for discovering patterns, using machine learning algorithms and statistics. Given the massive amount of processed structured and unstructured information, such systems are able to infer rules from data, develop models to make classifications, predictions and take decisions.

The second feature, i.e. the unpredictability by design, stems from the previous one. The reason is twofold: (1) the system is able to infer rules from data and make predictions on those data, rather than working on a set of pre-defined if-then rules, and (2) it is trained on datasets constantly changing. As a result, the system is easier to develop and maintain, but the possible output is not fully predictable, and its behavior cannot be fully explained by reference to the source code. Thus, these kind of systems enable the so called black-box medicine, that is opaque by its nature because the grounds for decisions are unknown and unknowable [19].

The third feature pertains the possible stronger impact on the decision making process. After conducting some experiments at the Sloan-Kettering Hospital in the USA, results show that Watson diagnosis are better and more accurate than those of physicians. "According to Sloan-Kettering, only around 20 percent of the knowledge that human doctors use when diagnosing patients and deciding on treatments relies on trial-based evidence. It would take at least 160 hours of reading a week just to keep up with new medical knowledge as it's published, let alone consider its relevance or apply it practically. Watson's ability to absorb this information faster than any human should, in theory, fix a flaw in the current healthcare model. Wellpoint's Samuel Nessbaum has claimed that, in tests, Watson's successful diagnosis rate for lung cancer is 90 percent, compared to 50 percent for human doctors." [25]. Thus, we can identify three key factors that may strongly influence the decision making process, i.e. the Watson's ability (1) to overcome the human cognitive limitations in collecting and processing information; (2) to outperform human doctors in diagnosis; and (3) the evidence-based approach as a strong argument to justify and trust the system's decision. These key factors may question Watson's role with regard to the decision making process in the health care, traditionally centered on human judgment and expertise.

3. The source of the decision making authority and the role of Watson in the health-care

In the health care domain, advanced AI systems have opened up the possibility of integrating highly autonomous systems into human teams. As a result, such systems expand the scale of collected and processed evidence, widening up the question on whether the human experts can still cope with AI system's expertise.

In this context we investigate whether the source of the decision making authority should be attributed only to human expert (e.g. clinicians and physicians), or conversely should be completely shifted on the AI system, or whether a shared decision making model is preferable.

In the first hypothesis, the AI system would be considered as a simple information management tool supporting the human expert. Thus, the standard of care remains what is reasonable to expect from the average doctor, in the specific medical field.

However, AI technologies such as Watson are purposely designed to interfere with human decisions making [22]: they are used on the assumption that they can outperform humans in medical expert tasks, overcoming not only cognitive limitations, but also time-sensitive limitations suffered by humans in accessing, reading, understanding and incorporating evidence into their expert practice. It has been argued that, evidence suggesting AI-expert systems can perform better than human expert constitutes also evidence that relinquishing control to AI CDSS, like Watson, is the better approach for reaching the gold standard of evidence-based practice [15].

Indeed, the second hypothesis, i.e. shifting the decision making authority to AI CDSS, is generally supported by two main arguments: (1) the normative pull evidence-based practice [15], which would be at least questionable to ignore; and (2) the better success rate demonstrated by such systems over human experts. Under this hypothesis, medical malpractice law would eventually require superior ML-generated medical diagnosis as the standard of care in clinical settings [7]. Thus, the medical expert, who would not be in the position to reach the same standard of care, would be bound to AI system's decisions. In case of failures resulting in injuries for patients, any deviation from the AI system's advice, would lead to the professional liability of the physician for medical negligence.

However, relying on AI systems in medicine rises further legal and ethical issues. The first one is related to the concept of evidence-base medicine: even though it is regarded as the gold standard in clinical practice, and it is considered as the best argument in favour of the AI decision making authority, there are a number of limitations and criticism when it is applied to the care of individual patients. These criticisms are due to the occurrence of biological variations, the need to consider the individual patient's values, and the limits for clinicians to access evidence and describe such evidence to patients, in order to facilitate a shared decision making between patients and doctors in the care process [26]. A broader understanding of the evidence-based medicine "requires a bottom up approach that integrates the best external evidence with individual clinical expertise and patients' choice" [20]. Under this approach, clinicians should make health care decision taking into account not only their expertise and the best scientific evidence available, but also the patient's values, goals, and preferences.

Additionally, the limitation to access evidence is directly related to the second issue, i.e. the explanation role in decision making, the AI systems' accountability and more precisely the possibility of obtaining human-intelligible and human-actionable information. As noted in section 2, AI systems like Watson are essentially black boxes inference engines that provide diagnosis and treatment recommendations without supporting explanations. The question is whether and to what extent statistical evidences can substitute the explanation function. Yet, if an AI expert system outperforms human experts in making diagnoses and suggesting treatments, then we would expect to understand why, on the basis of a satisfactory explanation. AI systems explainability is also required by articles 13 and 14 of the GDPR, according to which "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing" shall be provided in case of automated decision-making. In this regard, some authors raise the question whether the explanation should provide an account of all the patterns and variables taken into account by the system (model-centric explanation), or only those that are relevant to the specific patient case (subject-centric explanation) [6].

The access to evidence and the explanation play an essential role in the medical decision making process for both medical experts and patients. Medical experts should be able to assess the consistency of system's arguments in relation not only to the medical literature and the clinical practice, but also with regard to individual patients. The explanation allows physicians to determine the extent to which a particular input was determinative or influential on the output [5]. Thus, we may want to verify whether a patient's interests were taken into account in the diagnosis and treatment determination, as well as whether a certain factor was determinative. Under this perspective, the role of medical experts is central in considering factors which may affect the individual decision, such as symptoms that AI systems are unable to perceive (e.g. altered body odor, tissue consistency perception, etc.), patient's values and preferences, as well as in integrating such factors with AI systems' evidence and suggestion. All those aspects are necessary for eventually identifying counter arguments which may lead to different decisions. Moreover, empirical evidence shows that, providing explanations for recommended actions deeply influence user confidence in, and acceptance of AI-based decisions and recommendations [28]. On the other hand, the explanation is essential also for patients, to whom such explanation should be provided. A physician is usually required to explain why he/she is recommending a course of treatment, and how he/she came to a certain diagnosis. The explanation function is indispensable to guarantee a patient-centered care process, as well as patients' informed decision about their care and treatment. Additionally, the explanation can make medical advises more acceptable to patients. As already noted, the expertise of AI CDSS like Watson is not primarily based on the ability to provide such explanations; their success rate in performing particular tasks would seem to substitute such ability. The decision reliability is based on statistical evidence rather than on the ability to provide an explanation. The role of medical experts would remain central in interpreting and providing a meaningful comprehension to patients of both their health status and the recommended treatment. Even if we imagine a future where AI systems will be able to provide human-understandable evidence and explanation, medical experts would not be merely reduced to being intermediaries between AI systems and patients. Firstly, only medical experts have the specific domain knowledge for interpreting the pull of evidence and explanation and evaluating their reliability and correctness. Furthermore, such competence constitutes the keystone of both the doctor-patient trust relationship and interaction, with the regard to the whole care process and cooperation in treatment.

As a consequence, the third issue pertains the dimension of trust. Trust is traditionally considered a cornerstone of interpersonal relationship [13] and in the health care as the effective foundation of patient-doctor relationship. The need for interpersonal trust is rooted in the patient's vulnerability for being ill with regard to the actions of clinicians and doctors, the information asymmetry deriving from the specialist nature of medical knowledge [8], and the uncertainty and element of risk regarding the competence and intentions of the physician on whom the patient is dependent. Under the trust dimension, arguing in favour of AI systems decision making authority would necessarily undermine the patient-doctor relationship, that would be substituted by a patient-AI system relationship. This would ultimately bring to a concurrent transfer of the trustee role from medical experts to AI expert-systems.

We would argue that patient-doctor trust relationship is still essential in the care process for different reasons. The first one is based on a full and deepest understanding of the medical competence, which includes more than knowledge, judgment, and skill in technical functions. Medical competence also relies on the ability to help the patient feel at ease, interviewing sensitively and effectively to elicit both relevant symptoms and patient's concerns, conveying a sense of interest in listening carefully and providing responsive and meaningful feedback [14]. Removing such interpersonal human skills from the trust relationship may lead to undermining the patient's trust in the AI system's competence, even leading to a distrust situation resulting in an unwillingness to follow the AI advice.

A further reason is related to the information asymmetry deriving from the specialist nature of medical knowledge. Even though such information asymmetry also revolves around the medical-expert and the AI-expert system relationship, it would be overstretched since patients, completely lacking specific domain information, would never be able to comprehend and interpret data and assess evidence and explanations. A meaningful data comprehension as well as the evidence and explanation assessment are essential to make an informed decision on whether opt-in or opt-out from the AI system's recommendations.

Given all the mentioned criticisms, we would argue in favour of a shared decision-making authority. A further argument in favour of this model relies on the concept of joint cognitive system. It has been observed than when humans and AI systems interact for the fulfilment of a goal, it would be better to describe humans and technology not as two interacting "components", but as constituting a joint cognitive system, where the control is accomplished by an ensemble of the human cognitive system and the AI system that exhibits goal-directed behaviour [9]. Thus, tasks traditionally associated to the role of physician, shall be attributed to the joint cognitive system, so that they are distributed between the human expert and the AI system. Under this perspective, the standard of care would result from a combination of the reference to the standard of care for medical practice, and the standard resulting from ML-generated medical diagnosis. The first dimension should be taken into account with regard to the tasks assigned to the human expert, while the second one to those assigned to the AI system. As a result, the human should maintain the capability to oversee the overall activity of the AI system (including its legal and ethical impact in the care process) and the ability to decide whether and how use the system and relay on its recommendations. In case of failure resulting in injuries for patients, the liability should be assessed taking into account the task allocation as specified in section 5. The shared model allows the physician to ground his/her decisions not only on the pool of literature and clinical evidence, but also on the individual patient's biological variation, values and preferences, as well as factors that the AI system is unable to perceive.

The decision reliability will be based both on statistical evidence and on the physician's ability to interpret such evidence, at least detecting whether or not there is good evidence contradicting the system's suggestion, or evidence of AI system's errors, as well as on providing meaningful explanations to patients.

This model brings to a three-dimensional trust relationship involving the AI system, the human expert and the patient. In the context of AI, control over the system is constitutive of trust [3]. As already noted, given the specialist nature of medical knowledge, such control can be only exercised, at least partially, by physicians, also avoiding the risk to overstretch the information asymmetry.

The patient-doctor trust relationship would remain unchanged, relying on the full and deep concept of medical competence. In conclusion, AI systems cannot substitute the human expert as the source of decision making authority, which remains central for interpreting evidence, detecting AI system errors, providing explanations to patients, as well as for considering patient's legal and ethical values and principles, preferences and morality, and other information not available to the system.

4. The European discipline of Medical Device Software: the legal qualification and the conformity assessment procedure

This section deals with the legal qualification and the conformity assessment procedure of AI CDSS like Watson, under the European Regulation 2017/745. Such procedure constitutes the necessary requirement to obtain the European Conformity (CE) mark, through which a medical device is certified as compliant with product safety and performance requirements. In particular we shall evaluate whether additional criterium for the classification of medical devices are needed and how such criterium can influence not only the certification procedures, which constitute the necessary requirement to place a medical device on the market, but also the medical liability in case of technological failures and more generally in misdiagnosis and/or improper treatment scenarios.

According to Article 2(1) of the Regulation Watson can be classified as a software as a medical device for diagnostic, prediction and treatment purposes.

Under the Regulation, medical devices can be divided into four different classes, i.e. class I (low risk), IIa (moderate risk), IIb (medium risk) and III (high risk), depending on the device purpose and its inherent risks (See Chapter V, Sec. 1, Article 51 of the EU Regulation 2017/745). In particular, Annex VIII sets out three main classification criterium which take in to account (1) the duration of use (e.g. transient, short term, long term); (2) whether the device is invasive (i.e. any device which, in whole or in part, penetrates inside the body, either through a body orifice or through the surface of the body); and (3) whether the device is active (i.e., whether a device depends on source of electrical energy or any source of power other than that directly generated by the human body or gravity and acts by converting this energy, including software). Thus, while for example enema kits and elastic bandages fall under class I devices, because they present minimal potential for harm to patients; devices sustaining or supporting life, such as implantable pacemakers and breast implants, fall under class III, since they present potential high risk of injury to patients.

According to Rule 11 of Annex VIII, decision support systems generally fall under class IIa devices (moderate risk), unless they may seriously affect the patient's state of health, in which case they may fall under class IIb (medium risk) or III (high risk).

Reading the definition provided by Rule 11 in combination with the classification criterium under Annex VII presents some difficulties. Firstly, Rule 11 does not allow to clearly classify Watson under Class III devices. This classification appears to be based on the evaluation of whether the patient can suffer irreversible or serious deterioration in the health state. However, this evaluation can be made only case-by-case depending on the specific clinical situation of the patient and can only be conducted subsequently to the end of the design phase. It might not always be possible to assert, for example, whether in case of patient's death, the latter is the consequence of a wrong diagnosis and/or treatment, or of the clinical course of the specific pathology.

Secondly, the level of risk posed by a device depends from its intended use, which is determined on the basis of the manufacturer's labelling claims for the device, rather than on how clinicians use the device in practice. In case of AI CDSS like Watson, this distinction became particularly relevant, since the risk associated to the device does not arise from the physical interaction with the patient's body, but rather from how the system advices are used by clinicians and its influence on the decision-making process. Thus, in evaluating the risk level of AI CDSS like Watson, the parameter should be based on the accuracy of the data provided and the intended impact on a physician's clinical decision-making.

Focusing on the classification criterium, as specified in Annex VII, it is important to note that the level of automation of a medical device in no way influence the device risky-class. However, the level of automation of an AI system deeply affects the division of tasks between humans and machines, in performing different cognitive functions (i.e., acquiring information, analysing information, making decisions, and acting on them), as better specified in section 5. Delegation is in fact a risk, since its rationality strictly depends not only on the probability to properly achieve a certain goal but also on the costs associated to a possible failure [2]. Sure enough, in the health context, a failure in properly diagnosing the correct disease and delivering the appropriate medical treatment constitute a high risk to the patient's health and safety.

Watson and, more generally, the AI CDSS are characterised by a high level of automation, in particular with regard to certain cognitive functions, such as the acquisition and analysis of information, and the decision-making process, as shown in section 5. These levels affect the degree of the associated risks, with regard to the AI CDSS influence on the traditional decision-making process, the transparency issues and the medical awareness (as shown in section 3), as well as to the possible technological failures, misdiagnosis or wrong treatment scenarios. Consider for instance a computer-aided detection device like the AlertWatch:OR, which is intended for "secondary monitoring of patients within operating rooms and by supervising anaesthesiologists outside of operating rooms" [11]. These types of devices pose moderate risks compared to those like Watson, which do not simply provide additional information, but direct a specific clinical decision. Thus, AI CDSS for diagnosis and medical treatment should not be classified under the same risky class of former CDSS devices.

The level of automation of AI CDSS also affects the degree of the associated risks with regard to the transparency issues and the medical awareness, as already noted in section 3, as well as with regard to possible technological failures, misdiagnosis or wrong treatment scenarios, which may significantly affect patient's health and safety.

It clearly appears that the level of automation of a medical device should be considered as an essential parameter to properly assess the risky-class.

This is even more important if we consider that a different conformity assessment procedure is defined for each class, depending on the associated inherent risk. In particular, the procedure ranges from a basic conformity assessment for class I devices to the full quality assurance for class III devices (art 52).

While in the first case, the compliance assessment with the Regulation requirements can be carried out under the sole responsibility of the manufacturer, the full quality assessment procedure demands the involvement of both a notified body and an expert panel in evaluating and verifying the performance and the clinical safety of a medical device, i.e. the ability to achieve its intended purpose.

The full quality assessment procedure determines the highest level of security and safety guarantees, allowing reasonable expectations regarding both the functioning and the trustworthiness of class III medical devices. This reasonable expectation as well as the role played by the notified body and the expert panel, may significantly affect the liability assessment in case of injuries suffered by patients as a consequence of the use of class III devices (e.g. a technological failure).

Under this scenario, the conformity assessment procedure can affect the applicability of the legitimate expectation principle, which is strictly related to the expected level of security and safety guarantees. In particular, the CE mark may have a different impact on the applicability of the legitimate expectation principle, depending on whether it assumes a merely formal or a substantial nature. If the conformity is assessed under the sole responsibility of the manufacturer, then the CE mark should only have a formal relevance. Conversely, whenever the procedures demands the involvement of both the notified body and the expert panel, under the full quality assurance procedure, the CE label should assume a substantial relevance.

The substantial nature of the certification is crucial to allow the applicability of the legitimate expectation principle as a liability shield for physicians in case of technological failure.

Since Watson's classification under Class III presents some difficulties, the applicability of the legitimate expectation principle remains uncertain, simply considering the risky class.

As already noted, the conformity procedure affects the expected level of products' safety and quality. We believe that, rather than focusing on the intended use of medical devices, the classification criterium should take in to account a level of automation taxonomy as well as how clinicians use certain devices in practice. In conclusion, AI CDSS like Watson, which present high levels of automation related to different cognitive functions, should be classified under Class III. The highest level of full quality assurance procedure would act as a guarantee not only for physicians, enhancing AI CDSS reliability and allowing for the applicability of the legitimate expectation principle, but also for patients, ensuring a higher level of safety.

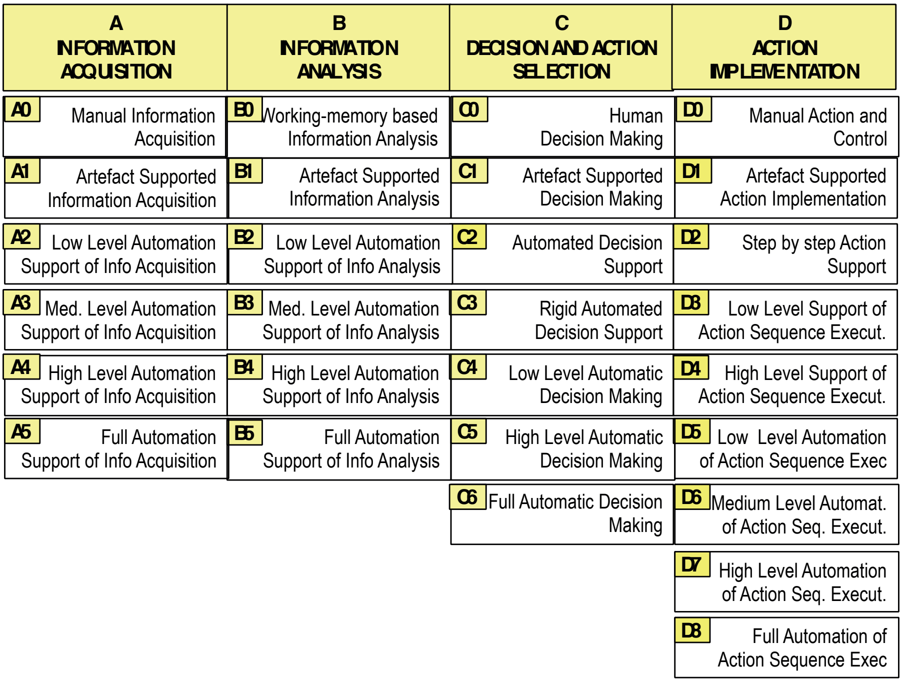
5. The level of Automation

Nowadays, the main productive, administrative and social organizations can be described as complex socio-technical systems (STSs), i.e. systems that combine technological artefacts, social artefacts and humans.

Technological artefacts, which to some extent involve the use of automated tools and machines, determine what can be done in and by an organization, amplifying and constraining opportunities for action according to the level of their automated technology. Social artefacts, like norms and institutions, determine what should be done, governing tasks, obligations, goals, priorities and institutional powers. However, norms need to be understood, interpreted, negotiated and actuated by humans. More generally, humans play an essential role in the functioning of STSs, providing them with governance and maintenance and sustaining their operation [27].

From this perspective, the health care system is the result of the interplay between technical artefacts (surgical robots, decision support systems, robotic pros- thesis, etc.), humans operators and users (physicians, paramedics, clinicians, care givers, patients, etc.), and social artefacts, which coordinate behaviours (including norms, such as laws, medical procedures, technical manuals, and institutions, such as hospitals, national institutes of health, regulatory agencies, etc.). The health care system is increasingly reliant on AI technologies, and it operates by interconnecting information systems, as well as by employing AI technologies, which sometimes replace humans, though they more often are part of human- machine interaction processes.

In failure scenarios, a further aspect that should be considered for allocating the liability is the level of automation of technological artefact, since they may affect how the decision-making process is split between human experts (e.g. physicians) and AI systems. This is strictly related to the allocation of task-responsibilities, namely the allocation of duties pertaining to the correct performance of a certain task or role.



A INFORMATION ACQUISITION	B INFORMATION ANALYSIS	C DECISION AND ACTION SELECTION	D ACTION IMPLEMENTATION
A0 Manual Information Acquisition	B0 Working-memory based Information Analysis	C0 Human Decision Making	D0 Manual Action and Control
A1 Artefact Supported Information Acquisition	B1 Artefact Supported Information Analysis	C1 Artefact Supported Decision Making	D1 Artefact Supported Action Implementation
A2 Low Level Automation Support of Info Acquisition	B2 Low Level Automation Support of Info Analysis	C2 Automated Decision Support	D2 Step by step Action Support
A3 Med. Level Automation Support of Info Acquisition	B3 Med. Level Automation Support of Info Analysis	C3 Rigid Automated Decision Support	D3 Low Level Support of Action Sequence Execut.
A4 High Level Automation Support of Info Acquisition	B4 High Level Automation Support of Info Analysis	C4 Low Level Automatic Decision Making	D4 High Level Support of Action Sequence Execut.
A5 Full Automation Support of Info Acquisition	B5 Full Automation Support of Info Analysis	C5 High Level Automatic Decision Making	D5 Low Level Automation of Action Sequence Exec
		C6 Full Automatic Decision Making	D6 Medium Level Automat. of Action Seq. Execut.
			D7 High Level Automation of Action Seq. Execut.
			D8 Full Automation of Action Sequence Exec

Figure 1: The LOAT (simplified version)

First of all, the violation of such duties may result in personal liability for human experts. Whenever there is a failure in a complex system, such failure is usually connected with the missing or inadequate execution of a certain task, and with the (natural or legal) person responsible for that task. As a consequence of the failure to comply with their task-responsibilities, such persons may be subject to liability under civil, criminal, and tort law.

Secondly, it may be necessary to identify task-responsibilities of AI systems, i.e. the requirements they should comply with. As task-responsibilities are progressively delegated to technology, the liability for damages and injuries shifts from human operators to the organisations, which designed and developed the technology, defined its context and uses, and are responsible for its deployment, integration, maintenance, and certification.

It is necessary to adopt a systematic approach for matching the levels of automation to different responsibilities of both human-experts and AI systems [4]. To determine the tasks allocation between human experts and AI CDSS like Watson, we consider the Level Of Automation Taxonomy

(LOAT) [21], based on the taxonomy developed by Endsley and Kaber [10] and the principles set out in Parasuraman et al. [16].

The LOAT provides criterium for allocating tasks with regard to four different cognitive functions, i.e. the information acquisition (A), the information analysis (B), the decision-making (C), and the action implementation (D). Figure 1 shows a simplified version of the LOAT. Each column starts with a 0 level of automation, corresponding to a fully manual accomplishment of a certain task, without any technical support. At level 1 the task is accomplished with "primitive" technical tools, i.e., low-tech non-digital artefacts. From level 2 on upwards, "real" automation is involved, and the role of the machine becomes increasingly significant up to the level where the task is fully automated. A certain technology may have different levels of automation with regard to the four cognitive functions, expressing varying levels of interaction between humans and technology.

In the following we consider the IBM Watson system and present the assessment results of its levels of automation.

Concerning the Information Acquisition (A), Watson supports the human expert in acquiring information on the process s/he is following. The system integrates data coming from different sources, such as Personal Health Records, medical datasets containing specific-domain literature and clinical trial reports. Then, it filters and/or highlights the relevant information items, for example selecting results of clinical trials concerning cancer diseases, rather than leukemia. The criterium for integrating, filtering and highlighting relevant information are predefined at design level and not available to physicians. Thus, with regard to the first cognitive function, Watson reaches a level A5 (Full Automation Support of Information Acquisition).

Concerning, the second cognitive function, namely the Analysis of Information (B), Watson performs comparisons and analyses of the available data, based on parameters defined at design level, reaching a level B5 (Full Automation Support of Information Analysis). In the LOAT classification, this level usually implies that the system triggers visual and/or aural alerts whenever a certain result requires the human expert attention. Consider, for instance, an arrhythmia detection alert generated by an electrocardiograph. Even though, we can imagine a near future in which Watson will be connected to other kind of medical devices, such as electrocardiographs, actually the analysis of information is a system's internal process, not accessible to human experts.

With regard to the Decision and Action Selection (C), Watson generates a ranked list of diagnoses (differential diagnosis) with an associated confidence score. It proposes one or more alternative decisions to clinicians, leaving them the possibility and freedom to generate alternative options. The ability to explore alternative hypothesis (diagnoses), along with the confidence score and the associated supporting evidence, is a key feature of the DeepQA technology. Physicians can evaluate these diagnoses along different dimensions of evidence, extracted from a patient's electronic medical record (EMR) and other related content sources. These dimensions include symptoms, findings, patient history, family history, current medications, demographics, etc.. Each diagnosis links back to the original evidence used by DeepQA for producing the associated confidence scores and supports the adoption of evidence-based medicine. Physicians can select either one of the alternative diagnosis proposed by the system, or her/his own one, for instance whenever he/she is aware of contextual circumstances (e.g. certain medical condition, patient's values, etc.) unknown to or ignored by the system, as well as in case he/she has evidence of AI system's errors. As a consequence, with regard to the third cognitive function, the system reaches a level C2 (Automated Decision Support).

Concerning the Action Implementation (D), namely the administration of medical treatments, human experts (physicians, care givers, etc.) execute and control all actions without any kind of AI system intervention. Thus, Watson reaches a level D0 (Manual Action and Control).

It clearly appears that, even though Watson reaches the full automation level in the information acquisition and analysis, physicians remain central in the decision-making process, in particular with regard to the decision and action selection, as well as to the action implementation.

6. Variations on a theme: possible failures and liability scenarios

On the basis of the assessed levels of automation, this section provides variations on the possible failures in the decision-making process and the related liability assessment, in case of injuries suffered by a patient as a consequence of misdiagnosis and/or improper treatments.

As already noted, Watson is used to analyse symptoms, make a diagnosis and elaborate the more appropriate treatment for specific diseases. In particular, it acquires the relevant information, integrating data coming from different sources, and analyses the available data. The system generates a number of hypothesis and then goes through a process of evidence testing.

Watson collects and classifies all potentially emerging diagnoses and the respective therapeutic plans, assigning them a specific confidence scores, i.e.

ranking answers according to their probabilities of being correct. In this way, the system supports the adoption of evidence-base medicine, applying the best available evidence obtained from the scientific method to the medical decision-making, through an abductive reasoning process, in the form of inference to the best explanation [18].

As an example, let us consider the case where a patient dies as a consequence of misdiagnosis or improper medical treatment. In order to assess the liability allocation, we shall consider variations of possible failures in the diagnosis process. To this end, we design four main scenarios. Each scenario is related to a failure occurring in the execution of a specific cognitive function in the decision-making process.

1. Failures in the acquisition of information phase

In a first scenario, the patient's death is causally related to a failure in the acquisition of information phase. Under this scenario, we may consider two different hypothesis:

Hypothesis 1: missing, incorrect, and/or incomplete source information.

Let us consider the case where some information, for instance those in the personal health record, the literature dataset, or the clinical trial reports, are missing, incorrect or incomplete. We are not observing an error in the acquisition phase, but rather an error in the information source. Watson may not be able to detect such error, that might be attributed to different causes, such as a human error (e.g. by physicians, nurses, knowledge engineers, etc.) in collecting and recording the information, or a technical failure in the medical examination process (e.g. a malfunction of the electrocardiograph). Under this hypothesis, it seems that the liability cannot be attributed neither to the medical staff using Watson, nor to the actors involved in the certification process, or in developing the system.

Hypothesis 2: failure in retrieving and selecting the relevant information.

Let us consider the case where the failure is caused by an error in retrieving and selecting the relevant information, used to make the diagnosis and recommend the medical treatment. According to the classification carried out in section 5, Watson reaches a level A5 (Full Automation Support of Information Acquisition).As already noted, the criterium for integrating, filtering and highlighting the relevant information are predefined at design level and are not available to physicians. As a consequence, the liability may be attributed to the actors involved in the definition of such criterium, and in the design process. Actors involved in the certification process, such as the notified body and members of the expert panel, may be find liable only if they were involved in the evaluation and assessment process of the system's design. Under this hypothesis, the liability should not be attributed to users, i.e. the medical staff using Watson, since they usually do not intervene in retrieving, integrating, filtering and highlighting the relevant information.

We may wonder if the system interface should be designed so as to alert the human expert if some needed information are unavailable or unreadable. Consider for instance the case in which Watson, missing the pregnancy status of a certain patient, recommends drugs that cannot be dispensed to pregnant women, because they may cause serious problems in the fetus, such as kidney damage, birth defect, growth restriction, etc. In these cases, additional liabilities may attributed to the manufacturer, for the defective design of the interface (i.e. not providing the alert), and to the medical staff, for ignoring the missing information alert provided by the system.

It should be noted that, since the criterium for the acquisition of information are predefined at design level, if the system is certified under the full quality assurance procedure, the legitimate expectation principle should be applied as a liability shield with regard to human expert's choice of trusting the system and its capability of performing the delegated task. The only exception would be the case where the human expert is aware or should have been aware that some relevant information was missing, or there is evidence of his/her negligent behaviour for ignoring the missing information alert.

2. Failure in the information analysis phase

Let us consider the case of a failure occurring in the information analysis phase, involving the diagnosis generation, the evaluation of positive and negative evidence supporting or rejecting each diagnosis and possible treatments, and the assignment of the related confidence scores. According to the classification carried out in section 5, Watson reaches a level B5 (Full Automation Support of Information Analysis). As already noted, the parameters for comparing and analysing the available data are predefined at the design level (and may be not visible to physicians, and in any case they may not be human understandable). Under this hypothesis, the liability may be attributed to the manufacturer, where a design defect or a manufacturing defect occurs as a consequence of the selection and implementation of certain parameters in the design process, as well as to the notified body and members of the expert panel, if they were involved in the evaluation and assessment of the system's design and functioning.

We can also consider the case in which the system may trigger visual and/or aural alerts, requiring attention by the medical staff, as in the electrocardiograph example described above. If the failure is causally linked to such functionality (because it is defective or missing), the liability would be attributed to the manufacturer, possibly for product defect. Conversely, members of the medical staff may be found liable, if the failure is the consequence of their behaviour, consisting, for instance, in negligently ignoring the alert.

As in the previous scenario, the parameters for the analysis of information are predefined at design level. Thus, if the system is certified under the full quality assurance procedure, the legitimate expectation principle should be applied as a liability shield for the human expert's choice of trusting the system and its capability of performing the delegated task. The only exception would be the case where the human expert, i.e. member(s) of the medical staff, negligently ignored the alert.

Additionally, since AI CDSS like Watson are capable of analysing and processing massive amount of information in a way that would be impossible for any human expert, and their output is not fully predictable, it is not reasonable to assign to such expert the legal duty of being in control of the internal processing activity of the system.

3. Failure in the decision and action selection phase

On the basis of the results emerged from the information analysis, Watson generates a ranked list of diagnoses with associated confidence scores, proposing alternative diagnosis and the associated treatments, leaving clinicians the possibility and freedom to select the best hypothesis, and/or to generate alternative options. According to the classification carried out in section 5, Watson reaches a level C2 (Automated Decision Support). Under this scenario, we may consider different hypothesis:

Hypothesis 1: Watson generates a correct diagnosis, and an associated treatment. In the following we consider four different sub-hypothesis:

1. Both the diagnosis and the associated treatment generated by Watson are correct, and the human expert follows the system's suggestion. This case is relatively unproblematic, since no controversy emerges between the human expert and the AI system, and no failures can be traced at the decision and action selection stage.
2. Both the diagnosis and the associated treatment are correct, but the human expert does not follow the system's suggestion, for instance generating a new diagnosis or a different treatment. Under this sub-hypothesis, a failure may emerge from the diverging human expert's decision. From the liability perspective, some authors [15] noted that the outcome depends on which expert judgment shall be considered as the source of the decision making authority. In particular, if Watson is considered as such source, then the liability can be attributed to human experts (e.g. the liability of physicians) under a specific duty of following the system's advices. Any divergent decision should be considered as a violation of such duty. However, as noted in section 3, given the trust relationship between patients and doctors, it is questionable that Watson should be considered as the decision making authority. Conversely, if human experts are still considered as the source of decision making authority, then their liability should be connected to cases of medical negligence and/or malpractice. In this case, the full quality assurance certification process may work as a guarantee of the system trustworthiness, and be considered as the effective cornerstone for the applicability of the legitimate expectation principle.
3. The diagnosis is correct but the associated treatment is wrong, and the human expert follows the system's suggestion. Let us consider the case where the wrong treatment derives from an internal system failure in generating the medical treatment. In this case, the manufacturer may be found liable for the defective technology, as well as the notified body and the members of the expert panel, if during the full quality assurance procedure some anomalies emerged in the clinical testing phase. Conversely, it is doubtful that the physicians' liability can be grounded solely on following the system's suggestion, with the exception of cases where they had good evidence contradicting the system's advice, or evidence-based reasons for not trusting such advice, e.g. on the basis of wrong results in similar previous cases. Thus, under the shared deci-

sion-making authority model, the liability shield can be grounded on the application of the legitimate expectation principle, whenever the systems has been certified under the full quality assurance procedure and the former relies on the correct performance of the delegated task. The wrong treatment may also result from the negligent behaviour of the human medical experts who neglect specific contextual circumstances such as patient's medical condition unknown to or ignored by Watson, as in the example of drugs dispensed to pregnant women.

4. The diagnosis is correct, but the associated treatment is wrong, and the human expert does not follows the system's suggestion. This case is relatively unproblematic, with regard to a possible controversy between the human expert and the AI system. In case of possible undesirable outcomes, the human expert liability may derive only from his/her negligent behaviour and/or medical malpractice.

Hypothesis 2: Watson generates a wrong diagnosis, and an associated treatment. In the following we consider two relevant sub-hypothesis:

1. Both the diagnosis and the associated treatment generated by Watson are wrong, and the human expert follows the system's suggestion. In this case, the manufacturer may be found liable for the defective technology, as well as the notified body and the members of the expert panel, if they were involved in the assurance procedure and some anomalies emerged in the clinical testing phase. It is doubtful that the human expert liability may be grounded solely on following the system's advice, with the exception of cases where he had good evidence contradicting the system's suggestion, or evidence-based reasons for not trusting such advice, e.g. on the basis of wrong results in similar previous cases. As noted above, under the full quality assurance procedure, the liability shield should not be grounded on the delegation of such authority from the human expert to the AI system, but rather on the application of the legitimate expectation principle.
2. Both the diagnosis and the associated treatment are wrong, but the human expert does not follow the system's suggestion. Even though a controversy between the human expert and the AI system emerged, this case remains unproblematic since possible undesirable outcomes may only result from clinicians negligent behaviour and/or medical malpractice.

4. Failure in the action implementation phase

In this scenario, a possible failure may only result from the human expert's behaviour, such as in cases where care givers dispensed overdose drugs. As noted in section 5, under the LOAT Watson reaches a level D0 (Manual Action and Control), since the human expert executes and controls all actions without any kind of AI system intervention. Therefore, liability may only be attributed to human experts, i.e. clinicians, care givers, etc., as a result of a negligent behaviour and/or medical malpractice.

7. Conclusion

In this contribution, we explored the distinctive features of new generation AI CDSS compared to the traditional ones, and the main legal issues emerging from the adoption of such AI systems in the health care domain.

New AI CDSS are going to improve the health care quality and patient safety. However, since they outperform medical experts in some activities, it might be questionable whether human experts can still cope with their expertise, and whether such systems should be considered as the source of decision making authorities.

However, as noted in section 3, relinquishing control to AI systems presents some difficulties. Medical experts cannot be reduced neither to mere executors of the AI system's advices nor to the role of intermediaries between AI CDSS and patients. In the care decision making process, medical experts remains central for integrating the best external evidence with individual clinical expertise and patients' biological variations, values, goals and preferences.

We argued in favour of a shared decision-making authority model. Medical experts should maintain their authority to oversee and assess the overall activity of AI CDSS (including their legal and ethical impact on the care process) and the possibility of evaluating whether and how make use of such AI systems and relay on their recommendations.

Firstly, a shared decision-making authority model relies on a broader understanding of the evidence-based medicine, that integrate the best external pull of evidence with individual clinical expertise and the patient's preferences and choices.

Secondly, under this model, the decision reliability will be based not only on the statistical evidence generated by AI CDSS, but also on physicians' ability to interpret such evidence, at least detecting whether or not there are good evidences contradicting the system's advices, or evidence of AI system's errors, as well as on providing meaningful explanations to patients. Under this perspective, the standard of care would result from a combination of the reference to the standard of care for human-expert medical practice, and the standard resulting from ML-generated evidence-based diagnosis.

Thirdly, such model brings to a three-dimensional trust relationship that involves the AI system, the human expert and the patient. The patient-doctor trust relationship would remain unchanged, relying on a full and deep understanding of medical competence, avoiding (a) the risk to overstretch the information asymmetry, deriving from the specialist nature of medical knowledge, between patients and medical experts, and (b) a distrust situation resulting in an unwillingness to follow the AI advice.

Finally, a shared model is consistent with the concept of joint cognitive system and the task allocation between humans and AI systems, where the control is accomplished by an ensemble of the human cognitive system and the AI system that exhibits goal- directed behavior. This perspective is also confirmed from the assessment of the Watson levels of automation in section 5.

In section 4 we have also shown how highest level of automation in performing different cognitive tasks can have a strong impact on the inherent risk of a medical device. Under the European Regulation 2017/745, the legal qualification of AI CDSS under a certain risky-class affects the associated conformity assessment procedure, determining the respective degree of security and safety guarantees. In this context, the full quality assurance certification process may function as a guarantee of the system trustworthiness, and be considered as the effective cornerstone for the applicability of the legitimate expectation principle. For these reason, we argue that, the level of automation should be considered as a classification criterium to determine the risky class of medical devices.

Moreover, the level of automation of technological artefact, may affect how the decision-making process is split between human experts (e.g. physicians) and AI systems. This is strictly related to the allocation of task-responsibilities, namely the allocation of duties pertaining to the correct performance of a certain task or role, and the consequent liability allocation, as shown through the assessment of the Watson level of automation, in section 5, and the failure scenarios provided in section 6.

References

[1] Aronson, J. E., T.-P. Liang & E. Turban. 2005. *Decision support systems and intelligent systems*, volume 4. Pearson Prentice-Hall.

[2] Castelfranchi, C. & R. Falcone. 1998. "Towards a theory of delegation for agent-based systems". *Robotics and Autonomous Systems*, 24(3-4), pp. 141-157.

[3] Castelfranchi, C. & R. Falcone. 2000. "Trust and control: A dialectic link". *Applied Artificial Intelligence*, 14(8). pp. 799-823.

[4] Contissa, G., M. Laukyte, G. Sartor, H. Schebesta, A. Masutti, P. Lanzi, P. Marti, & P. Tomasello. 2013. "Liability and automation: Issues and challenges for socio-technical systems". *Journal of Aerospace Operations*, 2 (1-2), pp. 79-98.

[5] Doshi-Velez, F., M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger & A. Wood. 2017. "Accountability of ai under the law: The role of explanation". *arXiv preprint arXiv*:1711.01134.

[6] Edwards, L. & M. Veale. 2017. "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for". *Duke L. & Tech. Rev.*, 16, p. 18.

[7] Froomkin, A. M., I. R. Kerr & J. Pineau. Forthcoming. "When ais outperform doctors: Confronting the challenges of a tort-induced over-reliance on machine learning". *Arizona Law Review*, pp. 18-23.

[8] Hengstler, M., E. Enkel & S. Duelli. 2016. "Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices". *Technological Forecasting and Social Change*, 105, pp. 105-120.

[9] Hollnagel, E. & D. D. Woods. 2005. *Joint cognitive systems: Foundations of cognitive systems engineering*. CRC Press.

[10] Kaber, D. B. & M. R. Endsley. 2004. "The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task". *Theoretical Issues in Ergonomics Science*, 5(2), pp. 113-153.

[11] Kheterpal, S., A. Shanks & K. K. Tremper. 2018. "Impact of a novel multiparameter decision support system on intraoperative processes of care and postoperative outcomes". *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 128(2), pp. 272-282.

[12] Kohn, L. T., J. Corrigan, M. S. Donaldson, et al. 2002. *To err is human: building a safer health system*, volume 6. National academy press Washington, DC.

[13] Mayer, R. C., J. H. Davis & F. D. Schoorman. 1995. "An integrative model of organizational trust". *Academy of management review*, 20(3), 709–734.

[14] Mechanic, D. 1998. "The functions and limitations of trust in the provision of medical care". *Journal of Health politics, policy and Law*, 23(4), pp. 661-686.

[15] Millar, J. & I. R. Kerr. 2013. "Delegation, relinquishment and responsibility: The prospect of expert robots". Available at SSRN: https://ssrn.comabstract=2234645.

[16] Parasuraman, R., T. B. Sheridan & C. D. Wickens. 2000. "A model for types and levels of human interaction with automation". *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 30(3), pp. 286-297.

[17] Patrick, L. 2015. *OECD Insights Ageing Debate the Issues: Debate the Issues*. OECD Publishing.

[18] Peirce, C. S. 1955. "Abduction and induction". In: *Philosophical writings of Peirce*, J. Buchler (ed.). pp. 150-156.

[19] Price, W. & I. Nicholson. 2015. "Describing black-box medicine". *BUJ Sci. & Tech. L.*, 21, 347.

[20] Sackett, D. L., W. M. Rosenberg, J. M. Gray, R. B. Haynes & W. S. Richardson. 1996. "Evidence based medicine: what it is and what it isn't".

[21] Save, L. & B. Feuerberg. 2012. "Designing human-automation interaction: a new level of automation taxonomy". *Proc. Human Factors of Systems and Technology* 2012.

[22] Selbst, A. D. Forthcoming. "Negligence and ai's human users". *Boston University Law Review*.

[23] Shortliffe, E. H. 1974. "Mycin: a rule-based computer program for advising physicians regarding antimicrobial therapy selection". *Technical report, Stanford University- Department of Computer Science*.

[24] Shortliffe, E. H., A. C. Scott, M. B. Bischoff, A. B. Campbell, W. Van Melle & C. D. Jacobs. 1984. "An expert system for oncology protocol management". In: *Rule-Based Expert Systems*, B. G. Buchanan & E.H. Shortiffe (eds.), pp. 653-665.

[25] Steadman, I. 2017. "Ibm's watson is better at diagnosing cancer than human doctors". *Wired*.

[26] Straus, S. E. & F. A. McAlister. 2000 "Evidence-based medicine: a commentary on common criticisms". *Cmaj*, 163(7), pp. 837-841.

[27] Vermaas, P., P. Kroes, I. van de Poel, M. Franssen & W. Houkes. 2011. "A philosophy of technology: from technical artefacts to sociotechnical systems". *Synthesis Lectures on Engineers, Technology, and Society*, 6(1), pp. 1-134.

[28] Ye, L. R. & P. E. Johnson. 1995. "The impact of explanation facilities on user acceptance of expert systems advice". *Mis Quarterly*, pp. 157-172.

The Authors

Stefania Basilico Stefania Basilico is a clinical and experimental psychologist and neuropsychologist as well as a psychotherapist. Since 1998 she has been working as a clinical neuropsychologist in the Cognitive Neuropsychology Center of Niguarda Cà Granda Hospital – Milan, directed by Prof. Gabriella Bottini. She has been involved in several research projects about the study of neuropsychology functions in patients affected by cognitive deterioration with publications in peer-reviewed and indexed journals. She also took part in multicenter clinical experimental trials on degenerative disorders. She is a lecturer at the University of Torino. Her research interests include the study of cognitive functions in different dementia profiles. She is member of several scientific associations.

Riccardo Bellazzi Riccardo Bellazzi, is Full Professor of Bioengineering and Biomedical Informatics at the University of Pavia. He is the director of the Laboratory of Medical Informatics "Mario Stefanelli". Moreover, he leads Laboratory of biomedical informatics at the hospital ICS Maugeri in Pavia. He is currently the Chair of the Department of Electrical, Computer and Biomedical Engineering of the University of Pavia. The scientific interests of Prof. Bellazzi are highly interdisciplinary and are aimed at applications of informatics to medicine and life sciences, comprising data mining, temporal data analysis, decision support, clinical research informatics. Prof. Bellazzi has a wide and internationally recognized research activity. He was involved in several national and international research projects in biomedical informatics and, currently, he is the technical project manager of the H2020 project PULSE. In 2009 he became a Fellow of the American College of Medical Informatics. In 2017 he became Fellow of the International Academy of Health Sciences Informatics. He has been Vice-President of the International Medical Informatics Association in the period 2011-2014. He is Associate Editor of the "Journal of Biomedical Informatics" and member of the editorial board of the journals "Methods of Information in Medicine", "Journal of the American Medical Informatics Association", "International Journal of Medical Informatics", "Journal of Diabetes Science and Technology". Finally, he is co-founder of the academic spin-offs Biomeris, which implements software to support clinical research, and Engenome, which is specialized on the analysis of Next Generation Sequencing data.

Francesca Bellazzi Francesca Bellazzi is a PhD student at the University of Bristol in the ERC Project "The Metaphysical Unity of Science" (ERC CoG) grant no. 771509. She holds a BA in Philosophy from Università Cattolica of Milan (2017) and an MSc in Philosophy of Science from the LSE (2018). Francesca is specialising in philosophy of science, philosophy of biology and metaphysics. She is currently investigating the role of Aristotelian essences in scientific explanation and unification, in particular in evolutionary-developmental biology.

Gabriella Bottini Gabriella Bottini is a neurologist, neuropsychologist, full professor of Cognitive Neuroscience at the University of Pavia and directs the Cognitive Neuropsychology Centre of the Niguarda Hospital in Milan. Her fields of research are the normal and pathological representation of the body and the cognitive impairments of neurodegenerative diseases and epilepsy. She is interested in the interaction between Neuroscience and other disciplines such as Law and Art. She is a member of a number of scientific associations, and author of about one hundred articles published in indexed, international scientific journals, of numerous chapters and books, focused on the issue of Cognitive Neuroscience.

Giuseppe Contissa Giuseppe Contissa is professor in Legal Informatics at the University of Bologna, and professor in Legal Informatics and in Legal Theory at LUISS University – Rome. He obtained the Italian National Scientific Qualification for the role of Associate Professor in Philosophy of Law (12/H3). He received his PhD in legal informatics and computer law from the University of Bologna, where he is currently a researcher at CIRSFID Research Center. He has been a Max Weber fellow and a research associate at the European University Institute (EUI), Florence, and resident fellow at the Stanford Center for Computers and the Law (CodeX), at Stanford University. His research interests include artificial intelligence and law, computable models of legal reasoning and knowledge, legal theory, legislative drafting, legal and ethical issues of artificial intelligence and robotics, liability and automation in socio-technical systems. He has published widely on these topics and has worked in several national and European projects, while also speaking at national and international conferences.

Francesca Lagioia Dr. Francesca Lagioia is postdoctoral research fellow at the University of Bologna and research associate at the EUI in the Claudette project. She is teaching assistant and tutor in the courses of Legal Informatics and Computer law and Advanced Legal Informatics, at the University of Bologna (2013-to date). She has been Max Weber Postdoctoral Fellow (1st September 2017-until 31st August 2018) at the European University Institute (EUI), Florence (Italy). In March 2016, she earned her Ph.D. in Law Science and Technology from the University of Bologna with a thesis on Criminal Liability and Automation in E-health. Her research interests include: artificial intelligence and law, computable models of legal reasoning and knowledge, legal theory, computer law and internet law, in particular privacy and data protection law, and consumer law; law and automation in socio-technical systems, with a specific focus on normative and deliberative agents, and the liability issues arising in connection with the use of autonomous systems. She has published widely on these topics and has worked in several national and European projects, while also speaking at national and international conferences.

Valeria Peviani After having achieved a summa con laude Master Degree in Cognitive Psychology in 2015 at the University of Pavia, Italy, Valeria Peviani completed a Ph.D program in "Psychology, Neuroscience and Medical Statistics". She conducted her Ph.D between the University of Pavia, Niguarda Hospital (Milan, Italy) and the Department of Neuroscience of the Max Planck Institute for Empirical Aesthetics in Frankfurt am Main, Germany, where she is now a postdoctoral researcher. She currently teaches courses on clinical and research tools in neuropsychology, as part of the Bachelor program in Psychology at the University of Pavia, and of the Master program in Psychology, Neuroscience and Human Sciences jointly offered by the University of Pavia and the IUSS University.

Giulia Pinotti Giulia Pinotti graduated cum laude in Law from the University of Pavia in 2016 and the same year she obtained her final degree at Institute for Advanced Study of Pavia (IUSS). She was student of the Collegio Ghislieri, Pavia. Since September 2016 Giulia is a PhD candidate in administrative law at the University of Milan, in co-tutorship with the University Paris I. Her research is on the automation of the Public Administration's decisions.

Amedeo Santosuosso Professor of Law, Science and New Technologies at the Department of Law, University of Pavia, and ICTs, Artificial Intelligence and Law at the Institute for Advanced Study of Pavia (IUSS). He served as President of the First Chamber at the Court of Appeal of Milan till March 2019. He is one of the founders and current Scientific Director of the European Center for Law, Science and new Technologies (ECLT), which is an Interdepartmental Research Center at the University of Pavia (I). He is promoter and main organizer of the Technological Innovation and Law (TIL 2019) intensive course. He is member of the World Commission on the Ethics of Scientific Knowledge and Technology (COMEST – UNESCO). He has widely published in the field of law and technology.

Frederike Seitz Frederike Seitz studied law and political science at the Julius Maximilian University of Würzburg and completed her legal clerkship in 2012. Since 2014 she has been a research assistant at the Chair of Prof. Dr. Susanne Beck LL.M. (LSE) for Criminal Law, Criminal Procedure Law, Comparative Criminal Law and Philosophy of Law at the Leibniz University of Hanover. She wrote her doctoral thesis on the influence of deep brain stimulation on criminal guilt. Currently she is working on the project "GEENGOV - Governance of Biomedical Genome Editing" funded by the Federal Ministry of Education and Research of Germany.

Tamar Sharon Tamar Sharon is an Associate Professor in philosophy of technology and co-director of the Interdisciplinary Hub for Security, Privacy and Data Governance (iHub) at Radboud University. Tamar has published on human enhancement, self-tracking for health, privacy, citizen science and the political economy of personal health data. She is PI of the ERC-funded project "Digital Good", which looks at the increasing role large consumer tech companies, like Google and Apple, are beginning to play in health research, and what this means for the common good. She is a member of the WHO European Advisory Committee on Health Research.

Paul Vogel Paul Vogel works as a Research Assistant at the research unit "RobotRecht" and the chair of Professor Eric Hilgendorf at the University of Würzburg, where he studied law and obtained a degree in European law. In his research, he focuses on the legal challenges of emerging technologies with a special emphasis on data protection law. In his doctorate studies, he deals with privacy and transparency issues of Artificial Intelligence (AI) applications.

Nicolas Woltmann Nicolas Woltmann passed the First State Examination in Law at the Julius Maximilian University of Würzburg (Germany) in 2016. In addition to his law studies, he successfully completed an "Accompanying Studies in European Law". Since 2017 he has been a research associate and doctoral candidate at the Würzburg Chair of Prof. Dr. Dr. Eric Hilgendorf, where he focuses on criminal and technology law. At present, he also devotes himself to these areas of law in his dissertation, which examines special liability issues in the field of machine learning.