

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

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

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

The strange case of Dr. Watson: liability implications of AI evidence-based decision support systems in health care / Francesca Lagioia; Giuseppe Contissa. - In: EUROPEAN JOURNAL OF LEGAL STUDIES. - ISSN 1973-2937. - ELETTRONICO. - 12:2(2020), pp. 245-289. [10.2924/EJLS.2019.028]

*Availability:*

This version is available at: <https://hdl.handle.net/11585/763572> since: 2021-01-07

*Published:*

DOI: <http://doi.org/10.2924/EJLS.2019.028>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

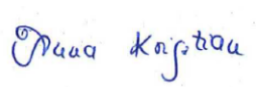
(Article begins on next page)

To whom it may concern,

Florence, 25 June 2020

This is to state that the attached paper titled 'The strange case of Dr Watson: liability implications of AI evidence-based decision support systems in health care' written by Francesca Lagioia and Giuseppe Contissa (doi:10.2924/EJLS.2019.028) has successfully passed the peer review process and is currently under formatting. The attached pre-formatted manuscript, which was prepared for your reference, while does not entirely correspond to the final version, comes very close to the version that will be published in a forthcoming issue of the European Journal of Legal Studies (ISSN 1973-2937).

Yours sincerely,

A handwritten signature in blue ink, appearing to read 'Anna Krisztian'.

Anna Krisztian  
Editor-in-Chief

# THE STRANGE CASE OF DR WATSON: LIABILITY IMPLICATIONS OF AI EVIDENCE-BASED DECISION SUPPORT SYSTEMS IN HEALTH CARE

Francesca Lagioia\*and Giuseppe Contissa\*\*

*This paper investigates the legal issues emerging from the adoption of clinical decision support systems (CDSS) based on artificial intelligence (AI). We explore a set of questions whose answers may affect the allocation of liability in misdiagnosis and/or improper treatment scenarios. The characteristic features of new-generation CDSS based on AI raise new challenges. In particular, the argument is made that a new shared decision-making authority model shall be adopted, in line with the analysis of the task–responsibility allocation. It is also suggested that the level of automation should be taken into account in classifying these systems under the European regulations on medical device software. This classification may indeed affect not only the certification procedures but also the allocation of liability. To this end, we finally design some scenarios providing variations on the possible causes of failure in the decision-making process and the consequent liability assessment.*

**Keywords:** Artificial Intelligence, Clinical Decision Support Systems, Liability and Automation

## TABLE OF CONTENTS

I. INTRODUCTION .....	3
2. DR WATSON VS TRADITIONAL CLINICAL DECISION SUPPORT SYSTEMS .....	7
1. The data-driven approach.....	8
2. Unpredictability by design .....	10
3. Impact on the decision-making process.....	12

\* Senior Research Fellow, European University Institute and Adjunct Professor, University of Bologna.

\*\* Junior Assistant Professor, University of Bologna and part time Professor, European University Institute.

3. THE EUROPEAN LEGAL FRAMEWORK ON MEDICAL DEVICE SOFTWARE: ITS LEGAL QUALIFICATION AND THE CONFORMITY-ASSESSMENT PROCEDURE.....	13
1. <i>The legal qualification</i> .....	13
2. <i>The conformity-assessment procedures</i> .....	16
4. THE LEVEL OF AUTOMATION AS A TASK-RESPONSIBILITY CRITERION .....	19
5. THE SOURCE OF DECISION-MAKING AUTHORITY AND THE ROLE OF WATSON IN HEALTH CARE ..	25
1. <i>Patients' peculiarities and the concept of evidence-based medicine</i> .....	27
2. <i>The role of explanation in decision-making</i> .....	28
3. <i>The role of trust in medical practice</i> .....	33
6. VARIATIONS ON A THEME: POSSIBLE FAILURES AND LIABILITY SCENARIOS.....	36
1. <i>Failures in the acquisition-of-information phase</i> .....	37
2. <i>Failure in the information-analysis phase</i> .....	38
3. <i>Failure in the decision-and-action-selection phase</i> .....	40
4. <i>Failure in the action-implementation phase</i> .....	43
7. CONCLUSION .....	43

## I. INTRODUCTION

The ageing of population is becoming one of the most significant phenomena of the 21st century. Over the past decades, life expectancy has significantly increased: 12 per cent of the world population is currently over the age of 60, and by 2050 this percentage is expected to rise to 21.<sup>1</sup> While this is a large triumph for modern science and medicine, it places a huge strain on the delivery of healthcare services, this owing to the increasing costs and the inexorable decrease in the number of medical personnel relative to the number of patients.<sup>2</sup> The advent of big data and the artificial intelligence (AI) era is usually considered part of the solution. The increased focus on preventing medical errors, coupled with the introduction of clinical decision

<sup>1</sup> Love Patrick, *OECD Insights Ageing Debate the Issues: Debate the Issues* (OECD Publishing 2015).

<sup>2</sup> Ibid.

support systems (CDSS), has been pointed out as key to the effort to improve healthcare quality and patient safety.<sup>3</sup> The adoption of CDSS for diagnosis and treatment should also facilitate evidence-based practice, which is regarded as the gold standard for decision-making in health care.<sup>4</sup>

In this context, the IBM Watson system is one of the most promising AI technologies developed in recent years. Initially designed to compete with human champions at the *Jeopardy!* quiz show,<sup>5</sup> Watson is currently being experimented as an evidence-based CDSS. It is based on the DeepQA technology, which exploits natural language processing and a variety of search techniques to analyse both unstructured information, for example natural language documents, and structured information, such as relational databases and knowledge bases.<sup>6</sup> DeepQA is trained on a set of documents on which human experts annotate all instances of pairs of questions and answers. The system learns how to identify and correlate questions and answers on the basis of the examples within the training set. It applies the acquired knowledge in analysing new input questions and generates 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, DeepQA tries to find evidence that either supports or refutes the hypothesis in question. The process outputs a ranked list of candidate answers – a potential diagnosis – with an associated confidence score.

This paper investigates some legal issues emerging from the adoption of Watson and similar AI CDSS in health care, especially as concerns medical practice and liability for accidents, calling for new models of allocating

<sup>3</sup> Linda T Kohn and others (eds), *To Err Is Human: Building a Safer Health System* (National Academies Press 2000).

<sup>4</sup> David L Sackett and others, *Evidence Based Medicine: What It Is and What It Isn't* (British Medical Journal Publishing Group 1996).

<sup>5</sup> Jeopardy! is an American television game show based on a quiz competition in which contestants are presented with general knowledge clues in the form of answers, and must phrase their responses in the form of questions. David Ferrucci et al., 'Building Watson: An overview of the DeepQA project' (2010) *AI magazine* 31(3) 59-79.

<sup>6</sup> David Ferrucci, Anthony Levas, Sugato Bagchi, David, Gondek, and Erik T. Mueller, 'Watson: beyond jeopardy!' (2013) *Artificial Intelligence* 199, 94

decision-making tasks between medical experts and AI systems. Even though the analysis is mainly focused on Watson, results can be extended to all AI CDSS systems sharing similar features.

The liability for damages caused by AI systems has been addressed in a number of studies with regard to civil<sup>7</sup> and criminal law,<sup>8</sup> and recently also in legal disputes and legislative initiatives, such as the report on Civil Law Rules on Robotics, issued by the legal affairs committee of the European Parliament,<sup>9</sup> the AI Strategy of the European Commission, and the High-Level Expert Groups on AI.<sup>10</sup> However, the liability resulting from the use of AI systems in the health domain has mainly focused, with some exceptions,<sup>11</sup> on robotic surgery,

- 7 See, among others, Ugo Pagallo, *The laws of robots* (Springer 2013); Paulius Čerka, Grigienė Jurgita, and Sirbikytė Gintarė, 'Liability for damages caused by artificial intelligence' (2015) *Computer Law & Security Review* 31(3) 376-389.
- 8 Ugo Pagallo, 'AI and bad robots: The criminology of automation' in *The Routledge Handbook of Technology, Crime and Justice* (Routledge 2017); Francesca Lagioia and Giovanni Sartor, 'AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective' (2019) *Philosophy & Technology* 1-33.
- 9 P8\_TA (2017)0051 Civil Law Rules on Robotics European Parliament resolution of 16 February 2017, with recommendations to the Commission on Civil Law Rules on Robotics 2015/2103 INL.
- 10 On 25 April 2018, the EU Commission set up three different groups of experts on (i) the ethics of AI; (ii) whether and to what extent to amend the directive on liability for defective products; and, (iii) liability and new technologies formation (<<https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>> accessed 27 June 2020. See also the Commission's document on Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines IP/18/3362; European Commission, White Paper On Artificial Intelligence - A European approach to excellence and trust (2020) (available at <[https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020\\_en.pdf](https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf)> accessed 27 June 2020; European Commission, Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics (2020) COM(2020) 64 final. For an extensive literature analysis of the foreseeable threats of AI crimes see Thomas C King, Nikita Aggarwal, Mariarosaria Taddeo, and Luciano Floridi 'Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions' (2018) 26(1) *Science and engineering ethics* 89-120.
- 11 Andreas Holzinger, 'Interactive machine learning for health informatics: when do we need the human-in- the-loop?' (2016) 3(2) *Brain Informatics* 119–131; W. Nicholson

telemedicine and smart prosthetics.<sup>12</sup> Moreover, the literature is still fragmented and a comprehensive and unified approach is still missing. Indeed, in this context, the legal analysis requires a systemic approach in order to consider the functioning and goals of the health system, calling for a novel method for analysing the roles and tasks of the actors involved and the associated responsibilities. A socio-technical perspective<sup>13</sup> — resulting from the combination of technical artefacts (surgical robots, decision-support systems, robotic prosthetics, etc.), human operators and users (physicians, paramedics, clinicians, caregivers, patients, etc.), and social artefacts (including laws, medical procedures, technical manuals, and institutions, such as hospitals, national institutes of health, and regulatory agencies) — provides the means to investigate what activities are entrusted to AI CDSS and the role that such systems play in health care.

In this paper, this perspective is adopted in order to explore a set of questions whose answers may heavily affect the allocation of liability in misdiagnosis and/or improper treatment scenarios. In particular, section 0 explores the distinctive features of AI based CDSS by comparison with traditional ones. This analysis is meant to provide the necessary technological framework for evaluating how and to what extent these new AI technologies can change the medical practice and the potential risks associated with this transformation. At the same time, given the potential of such technologies to be transformative,

Price II, 'Regulating black-box medicine' (2017) Mich. L. Rev. 116, 421; Jason Millar and Ian Kerr, 'Delegation, Relinquishment and Responsibility: The Prospect of Expert Robots' (2013) Available at SSRN <<https://ssrn.com/abstract=2234645>> accessed 27 June 2020.

- 12 Andrea Bertolini, 'Robotic prostheses as products enhancing the rights of people with disabilities. Reconsidering the structure of liability rules' (2015) 29(2-3) International Review of Law, Computers & Technology 116-136; Shane O'Sullivan, Nathalie Nevejans, Colin Allen, Andrew Blyth, Simon Leonard, Ugo Pagallo, Katharina Holzinger, Andreas Holzinger, Mohammed Imran Sajid, and Hutan Ashrafian, 'Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery' (2019) The International Journal of Medical Robotics and Computer Assisted Surgery 15(1) e1968.
- 13 Pieter Vermaas, Peter Kroes, Ibo van de Poel, Maarten Franssen, and Wybo Houkes, 'A philosophy of technology: from technical artefacts to sociotechnical systems' (2011) in *Synthesis Lectures on Engineers, Technology, and Society*, VI(1) 1-134, 70.

there is the need to analyse how they are regulated by the existing legal framework and whether it is adequate or fail to provide appropriate solutions and guidance. Thus, section 0 deals with the legal qualification and the conformity-assessment procedure of AI based CDSS under the European Regulation on medical device software. This analysis is meant to evaluate whether additional criteria for classifying these systems are needed and how they can influence the certification procedures and medical liability as well.

Once the analysis of the specific technological features of AI CDSS and the regulatory framework governing their classification and certification is completed, the focus will fall on the allocation of tasks and activities and on the interaction between medical experts and AI CDSS. In particular, section 0 explores how and to what extent the level of automation may affect the allocation of liability. The analysis shall consider what activities are being delegated to the Watson system, as an example of AI CDSS, and what changes this introduces into interactions, and what new capacities and power relations are consequently engendered. This investigation is meant to address the connection between delegation and responsibilities and the relations of influence, leading to different legal responsibilities.

Section 5 investigates whether and to what extent the features of the Watson system raise questions with regard to the source of decision-making authority. Section 0 designs some scenarios, providing variations on the possible causes of failure in the decision-making process and the consequent liability assessment. It may be the case that, under the current legal regimes and without adequate adjustments, the allocation of liability will end up being unfair or inefficient. The adoption of a socio-technical perspective and the resulting liability analysis may be viewed as a governance mechanism<sup>14</sup> by which to enhance the functioning of the healthcare system.

## **2. DR WATSON VS TRADITIONAL CLINICAL DECISION SUPPORT SYSTEMS**

This section considers the CDSS as a technological component of the healthcare socio-technical system (STS). It focuses on the comparison between Watson,

<sup>14</sup> Gordon Baxter and Ian Sommerville, 'Socio-technical systems: From design methods to systems engineering' 23(1) in *Interacting with Computers* (2011) 4–17.



as an example of new-generation AI CDSS, and those based on the more traditional knowledge-based approach. As mentioned above, this analysis is meant to provide the necessary framework for assessing how and to what extent these new AI technologies can transform medical practice and pose new risks.

In particular, three main features are identified that distinguish Watson, and all the new AI CDSS, from traditional expert systems.<sup>15</sup> They are based on the formal representation of the specific domain knowledge: (1) the data-driven approach, (2) unpredictability by design, and (3) the possible stronger impact on the decision-making process. All these features pose new questions with regard to medical practice and the regulatory framework, under which current rules may fail to provide appropriate governance mechanisms.

### *1. The data-driven approach*

The first feature pertains the widespread adoption of data-driven methods in 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 structured and narrow problem domain.<sup>16</sup> As a result, such systems are suitable for dealing with, and providing advice on, repetitive problem areas, rather than with ad hoc and unique situations.

Human expertise has to be elicited and represented symbolically. In particular, symbolic reasoning is based on algorithms to make inferences grounded in the knowledge base using forward chaining (from data to conclusion) and backward chaining (from conclusion to data).<sup>17</sup> Such expert systems are typically based on classical procedural algorithms. The first examples were MYCIN and ONCOCIN, both developed at Stanford University in the early 1980s.

<sup>15</sup> For an overview on traditional expert systems see Jay E Aronson, Ting-Peng Liang, and Richard V. MacCarthy, *Decision Support Systems and Intelligent Systems* (Pearson Prentice-Hall 2005), ch. 3, 103ff.

<sup>16</sup> Jay E Aronson, Ting-Peng Liang, and Richard V. MacCarthy, *Decision Support Systems and Intelligent Systems* (Pearson Prentice-Hall 2005) 549.

<sup>17</sup> Ibid.

In particular, the MYCIN system was developed to identify bacteria causing blood infections to arrive at a probable diagnosis, based on reported symptoms and medical test results, and to recommend a course of treatment.<sup>18</sup> Similarly, ONCOCIN was an oncology-protocol management system designed to assist physicians in the treatment of cancer patients through a rule-based reasoner that encompasses the necessary knowledge of cancer chemotherapy. In generating its recommendation, the system combined initial data about the patient's diagnosis, results of laboratory tests, and the protocol-specific information in its knowledge base.<sup>19</sup>

Despite the great interests and appeal generated by these technologies and applications, they have not fundamentally transformed medical practice. This is mainly due to the so-called knowledge representation bottleneck: in order to build a successful application, the required information — including tacit and common-sense knowledge — had to be represented in advance using formalised languages. This proved to be very difficult, and in many cases impractical or impossible, also due to the endlessness evolution of medicine and new discoveries in medical science.

In the last decade, the focus of AI research has shifted to the possibility of applying machine-learning algorithms to vast amounts of data making an impressive leap forward. Data-driven AI systems, like Watson, use big-data analytics and data-mining techniques to discover patterns, with the help of machine-learning algorithms and statistics. Given the massive amount of processed structured and unstructured information, such systems are able to infer rules from data and develop models for making classifications, predictions, and decisions. It is important to note that these AI systems present a high level of complexity. First of all, they are not a single technology but rather a diverse set of different technologies.<sup>20</sup> For instance, the Watson

<sup>18</sup> Edward Hance Shortliffe, *MYCIN: A Rule-Based Computer Program for Advising Physicians regarding Antimicrobial Therapy Selection* (Stanford University Department of Computer Science 1974).

<sup>19</sup> Edward H Shortliffe and others, 'An Expert System for Oncology Protocol Management' in BG Buchanan and EH Shortliffe (eds) *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project* (1984) 656.

<sup>20</sup> The complexity AI systems is reflected in the multiplicity of components, software, parts, combined together. See, European Commission, *Report on the safety and*

system includes the Deep QA architecture, which goes from question analysis and answer type determination to search and then answer selection, and the Apache Unstructured Management Architecture (UIMA)<sup>21</sup> for content analytics. The latter provides a component software architecture for the development, discovery, composition, and deployment of multi-modal analytics for the analysis of unstructured information and integration with search technologies. Furthermore, these different technologies and components are in turn based on a combination of a variety of methods and algorithms performing their various functions. For instance, for the *Jeopardy* Challenge, computer scientists working on Watson used more than 100 different techniques for analysing natural language, identifying sources, generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses.<sup>22</sup>

A further dimension of this complexity concerns the internal complexity of the algorithms involved and the composition of the training sets used by such systems to learn methods for achieving their goals. It may be increasingly difficult to identify the source of possible problems and what ultimately caused harms and injuries.

## *2. Unpredictability by design*

The second feature, unpredictability by design, stems from the previous one. The reason is twofold: first of all, data-driven AI systems are able to learn and infer rules from data and make predictions on those data, rather than working on a set of predefined if-then rules, and secondly, they are trained on constantly changing datasets.<sup>23</sup> Algorithms may evolve through self-learning by developing new heuristics (problem-solving strategies) and modifying their

*liability implications of Artificial Intelligence, the Internet of Things and robotics* (2020) COM(2020) 64 final, 2.

<sup>21</sup> David Ferrucci, and Adam Lally, 'UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment' (2004) 10(3–4) *Natural Language Engineering* 327–348.

<sup>22</sup> David Ferrucci et al. 'Building Watson: An overview of the DeepQA project' (2010) 31(3) *AI magazine* 59–79.

<sup>23</sup> Jason Millar, and Ian Kerr, 'Delegation, relinquishment, and responsibility: The prospect of expert robots', in *Robot Law* (Edward Elgar Publishing, 2016) 107.

internal data and structure, or even by generating new algorithms.<sup>24</sup> Furthermore, due to their nature, such systems are open, since they often interact with other systems or data sources in order to function properly, thus allowing external input either via some hardware plug or through some wireless connection, and they come as hybrid combinations of hardware, software, continuous software updates, and various continuous services.<sup>25</sup>

Machine-learning-based (ML-based) systems present both advantages and disadvantages if compared to classical rule-based systems. The former are easier to develop and maintain, but the possible outputs are not fully predictable, and the systems' behaviour cannot be fully explained by reference to the source code. Indeed, such systems are designed to respond to, identify, and classify new and not necessarily predefined stimuli and to link them to a corresponding decision, selected among all the possible decisions. Moreover, they do not have the capability to explain the reasoning process behind the decision-making, a capability that is necessary for understanding why decisions are made in a certain way and providing explanations to their users (which are required by physicians).

As a result, AI-based CDSS, opaque by their nature,<sup>26</sup> enable so-called black-box medicine, since grounds for decisions are at least partly unknown and

<sup>24</sup> For instance, genetic algorithms are the most widely used form of evolutionary computation for medical applications. They are a class of stochastic search and optimisation algorithms based on natural biological evolution. They work by creating many random solutions to the problem at hand. This population of many solutions will then evolve from one generation to the next, ultimately arriving at a satisfactory solution to the problem. The best solutions are added to the population while the inferior ones are eliminated. The process is repeated among the better elements, so that improvements will occur in the population, survive and generate new solutions. Genetic algorithms are applied to perform several types of tasks like diagnosis and prognosis, medical imaging and signal processing. See, for example, Ramesh, A.N., Kambhampati, C., Monson, J. RT and Drew, P.J. 'Artificial intelligence in medicine' (2004) 86(5) *Annals of the Royal College of Surgeons of England* 334.

<sup>25</sup> See the Report commissioned by the EU Commission: Expert Group on Liability and New Technologies – New Technologies Formation (2019) 'Liability for Artificial Intelligence and other emerging digital technologies' 33.

<sup>26</sup> W. Nicholson Price II, 'Black-Box Medicine', (2015) 28 *HARV. J.L. & TECH.* 433

unknowable.<sup>27</sup> As will be discussed in section 0, these characteristics raise new issues, in particular with regard to AI transparency, trustworthiness, and accountability,<sup>28</sup> complicating the possibility of discovering the reasons behind AI evaluations and decisions and thus establishing the causes of potential failures in the diagnosis and treatment process.

### *3. Impact on the decision-making process*

The third feature concerns the possible impact of AI technologies on the decision-making process. Experiments done at the Sloan-Kettering Hospital in the United States suggest that Watson diagnoses are better and more accurate than those of physicians.

According to Sloan-Kettering, only around 20 per cent 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 is 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 per cent, compared to 50 per cent for human doctors.<sup>29</sup>

As a result, three key factors can be identified that may strongly influence the decision-making process. The first factor is the ability of Watson and similar AI-based CDSS to overcome human cognitive limitations in collecting and processing information. The second one consists in their capacity to outperform human doctors in diagnosis. The last one pertains the adoption of

<sup>27</sup> W. Nicholson Price II, 'Describing Black-Box Medicine' (2015) 21 BUJ Sci. & Tech. L. 347; Alex John London, 'Artificial intelligence and black-box medical decisions: accuracy versus explainability' (2019) 49(1) Hastings Center Report 15-21.

<sup>28</sup> For a recent contribution on the importance of robotics transparency, interpretability and accountability see Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, 'Transparent, explainable, and accountable AI for robotics' (2017) 2(6) Science Robotics.

<sup>29</sup> Ian Steadman, 'IBM's Watson Is Better at Diagnosing Cancer Than Human Doctors' (Wired, 11 February 2013) <<https://www.wired.co.uk/article/ibm-watson-medical-doctor>> accessed 27 June 2020.

an evidence-based approach, focused on clinical trials, in making diagnoses and recommending treatment. The latter is often considered as a strong argument for justifying and trusting the decision-making of the system, as examined in the following sections.

Given the potential impact of these technologies on medical practice, there is the need to examine the existing regulatory framework in order to evaluate whether it is adequate or may fail to provide appropriate governance mechanisms.

### **3. THE EUROPEAN LEGAL FRAMEWORK ON MEDICAL DEVICE SOFTWARE: ITS LEGAL QUALIFICATION AND THE CONFORMITY-ASSESSMENT PROCEDURE**

This section deals with the social component of the healthcare STS. In particular, it analyses the legal qualification and the conformity-assessment procedure of AI CDSS like Watson under European Regulation 2017/745.<sup>30</sup>

The certification procedure sets the necessary requirement for obtaining the European Conformity (CE) mark, through which a medical device is certified as compliant with product-safety and performance requirements. The analysis is meant to assess whether additional criteria for classifying these systems are needed and how they can affect the certification procedures, in which lies the necessary requirement for placing a medical device on the market. We shall also examine how the mentioned criteria and the certification processes may impact on medical liability in case of technological failures and more generally in misdiagnosis and/or improper treatment scenarios.

#### *1. The legal qualification*

According to Article 2(1) of the Regulation, Watson can be classified as a medical device for diagnostic, prediction, and treatment purposes.<sup>31</sup> Under the

<sup>30</sup> Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. It will be applicable three years after its entry into force, i.e. 25 May 2017.

<sup>31</sup> Under Article 2(1) of the Regulation (EU) 2017/745, a medical device is defined as 'any instrument, apparatus, appliance, software, implant, reagent, material or other

Regulation, medical devices can be sorted into four different classes — class I (low risk), class IIa (moderate risk), class IIb (medium risk), and class III (high risk) — depending on the purpose of the device and its inherent risks. In particular, Annex VIII sets out three main classification criteria, which take into 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 a source of electrical energy or any source of power other than that directly generated by the human body or by gravity and works by converting this energy). For example, enema kits and elastic bandages fall under class I devices, because their potential for harm is minimal. Conversely, devices sustaining or supporting life, such as implantable pacemakers and breast implants, fall under class III, given their higher potential risks for patients' life and well-being.

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 class III (high risk).<sup>32</sup>

article intended by the manufacturer to be used, alone or in combination, for human beings for one or more of the following specific medical purposes: — diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease, — diagnosis, monitoring, treatment, alleviation of, or compensation for, an injury or disability [...].’ Rule 11 of Annex VIII of the Regulation (EU) 2017/745 reads: ‘Software intended to provide information which is used to take decisions with diagnosis or therapeutic purposes is classified as class IIa, except if such decisions have an impact that may cause: — death or an irreversible deterioration of a person's state of health, in which case it is in class III; or — a serious deterioration of a person's state of health or a surgical intervention, in which case it is classified as class IIb. Software intended to monitor physiological processes is classified as class IIa, except if it is intended for monitoring of vital physiological parameters, where the nature of variations of those parameters is such that it could result in immediate danger to the patient, in which case it is classified as class IIb. All other software is classified as class I.’

<sup>32</sup> See ch V, sec 1, art 51, of Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices.

In combination with the classification criteria in Annex VII, the definition provided in Rule 11 presents some challenges. First, under Rule 11, Watson cannot be clearly classified as a class III device. This classification appears to be predicated on an assessment as to whether patients can suffer irreversible damage to their health or a serious deterioration in their state of health. However, this assessment can only be made on a case-by-case basis, depending on the patient's specific clinical situation, and only once the design phase is completed. It may not always be possible to determine, for example, whether in the event of a patient's death, the latter is the consequence of a misdiagnosis and/or treatment or of the clinical course of the specific pathology.

Second, the level of risk posed by a device depends on its intended use, which is determined on the basis of the claims made by the manufacturer in labelling the device. In the case of AI CDSS, the risk associated with the device does not arise from physical interaction with the patient's body but rather from the way the AI recommendations are used by clinicians and from their influence on the decision-making process. Thus, in evaluating the risk level of AI CDSS, 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 criteria specified in Annex VII, it is important to note that the level of automation of a medical device in no way influences its risk class. However, as better specified and analysed in sections 0, 0, and 0 the level of automation deeply affects the division of tasks between humans and machines in performing different cognitive functions, including acquiring and analysing information, making decisions, and acting on them. Delegation is in fact a risk, since its rationality closely depends not only on the likelihood of properly achieving a certain objective but also on the costs associated with a possible failure.<sup>33</sup> No doubt, in the health context, a misdiagnosis, with the consequent failure to deliver the appropriate medical treatment, poses a high risk to the patient's health and safety.

AI CDSS are characterised by a high level of automation, particularly with regard to certain cognitive functions, such as the acquisition and analysis of information and the decision-making process (see section 0). These levels

<sup>33</sup> Cristiano Castelfranchi and Rino Falcone, 'Towards a Theory of Delegation for Agent-Based Systems' (1998) 24(3–4) *Robotics and Autonomous Systems* 141.



affect the degree of the associated risks, with regard to (i) the way AI CDSS affect the traditional decision-making process; (ii) transparency issues and medical awareness (as discussed in section 0); and (iii) 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'.<sup>34</sup> These devices pose moderate risks by comparison with the risks posed by systems like Watson, which do not simply provide additional information but also suggest and indicate a specific clinical decision to be made. Thus, AI CDSS for diagnosis and medical treatment should not be classified under the same risk class as former CDSS devices.

The level of automation in AI CDSS also affects the degree of risk with regard to transparency and medical awareness. This is especially the case given that AI lacks the ability to explain the internal reasoning process behind the decision-making, which should support diagnosis and treatment recommendations (see sections 2 and 0). In addition, there are risks associated with possible technological failures, misdiagnosis, or wrong-treatment scenarios, which may significantly affect patients' health and safety.

It clearly appears that the level of automation of a medical device should be considered an essential parameter for properly assessing the risk class of AI-based medical devices. This is even more so if it is considered that a different conformity-assessment procedure is defined for each class depending on the associated inherent risk, as discussed in the following section.

## *2. The conformity-assessment procedures*

According to Article 2 of EU Regulation 2017/745, conformity assessment means the process demonstrating whether the legal requirements relating to a device have been fulfilled.

<sup>34</sup> Sachin Kheterpal, Amy Shanks, and Kevin K Tremper, 'Impact of a Novel Multiparameter Decision Support System on Intraoperative Processes of care and Postoperative Outcomes' (2018) 128(2) *Anesthesiology: The Journal of the American Society of Anesthesiologists* 272.

This process ranges from a basic conformity-assessment procedure for class I devices to a full quality assurance for class III devices (Article 52).<sup>35</sup> In the first case, the assessment of compliance with the Regulation can be carried out under the sole responsibility of the manufacturer, with regard to what the manufacturer claims in the EU declaration of conformity (Article 19 of the Directive). In the second case, however, 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, including its ability to achieve its intended purpose as claimed by the manufacturer through labels, instructions for use, and the assessment of benefits and risks. Indeed, as specified in Article 2(52) the clinical performance of a device refers to its ability 'to achieve its intended purpose as claimed by the manufacturer, thereby leading to a clinical benefit for patients, when used as intended by the manufacturer', as resulting 'from any direct or indirect

<sup>35</sup> Given the lower risk level in the first case, i.e. devices in class I, the conformity-assessment procedure can be carried out under the sole responsibility of the manufacturer (art 19). Under class IIa, the manufacturer is required to establish and implement a quality management system (annex IX ch. I and III), and provide technical documentation for representative devices, without expert review. The notified body must approve and periodically audit (surveillance assessment) the quality-management system and assess its conformity with the required standard (alternatively, a manufacturer may provide technical documentation aligned with annexes II and III and select a conformity-assessment avenue based on annex XI). The conformity-assessment procedure for a class IIb non-active and non-implantable device is identical to the procedure for a class IIa (chs I and III of annex IX). In the case of implantable devices, the technical documentation must be provided for every device without expert review. In the case of active devices, the technical documentation must be provided for every device with expert panel involvement. Generally, manufacturers of class III devices are subject to a conformity assessment as specified in annex IX, including full quality assurance audit and full technical documentation review. Additionally, for class III implantable devices, an expert panel is involved in the evaluation. While standards are voluntary, one way of presuming conformity to the GSPR (General Safety and Performance Requirements in annex I) and meeting the provisions of full quality assurance is to obtain a harmonized EN ISO 13485 standard certification (alternatively, the manufacturer may choose to apply a conformity assessment as specified in annex X (Type-Examination) coupled with a conformity quality management assessment focused on production and controls, as specified in annex XI).

medical effects which stem from its technical or functional characteristics, including diagnostic characteristics'.

The full quality-assessment procedure secures the highest level of security and safety guarantees, creating 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, for example through a technological failure.

In this scenario, the conformity-assessment procedure can affect the applicability of the legitimate expectation principle.<sup>36</sup> According to Article 6 of Council Directive 85/374/EEC, on liability for defective products, a legitimate expectation is determined by circumstances such as (a) the presentation of the product; (b) its intended use; and (c) the state of the art at the time it was put into circulation. Additionally, under Article 3 of Council Directive 85/374/EEC, the conformity of a product to the general safety requirement is to be assessed by taking account of multiple elements, including (a) national and European standards, (b) the Commission recommendations setting guidelines on product safety assessment, (c) product safety codes of good practice in force in the sector concerned; (d) the state of the art and technology; and (e) reasonable expectations about safety.

In particular, the CE mark may impact the applicability of the legitimate expectation principle in different ways depending on whether it assumes a merely formal or a substantive nature. If conformity is assessed under the sole responsibility of the manufacturer, then the CE mark should only have formal relevance. Conversely, whenever the procedure demands the involvement of both the notified body and the expert panel, under the full quality-assurance procedure the CE label should assume substantive relevance. The substantive nature of the certification is crucial to enabling the applicability of the legitimate expectation principle as a liability shield for physicians in the event

<sup>36</sup> For an application of the legitimate expectation principle in product liability, see the UK decision *A & Others v National Blood Authority* [2001] 3 All ER 289, and the advocate general's opinion in CJEU Joined Cases C-503/13 and C-504/13 *Boston Scientific Medizintechnik* [2015] ECLI:EU:C:2015:148.

of technological failure.<sup>37</sup> Since the class III classification of Watson raises some difficulties, the applicability of the legitimate expectation principle remains uncertain, simply in view of the high-risk class.

As noted, the conformity assessment procedure affects the expected level of product safety and quality. We believe that rather than focusing on the intended use of medical devices, the classification criterion should take into account the level of automation and how clinicians use the device in practice, including the extent to which they may impact, affect, and even guide their decisions. In conclusion, AI CDSS like Watson, which have high levels of automation related to different cognitive functions, should be classified under class III. The highest level — the level afforded by the full quality-assurance procedure — would act as a guarantee not only for physicians, enhancing the reliability of AI CDSS and allowing for the applicability of the legitimate expectation principle, but also for patients, ensuring a higher level of safety. In line with the above, the High Level Independent Expert Group on AI, set up by the European Commission, recently published a set of Guidelines for Trustworthy AI, highlighting the need for certification procedures that should apply standards developed for different application domains and AI techniques, appropriately aligned with industrial and societal standards in different contexts.<sup>38</sup>

#### **4. THE LEVEL OF AUTOMATION AS A TASK-RESPONSIBILITY CRITERION**

This section explores the interaction between AI systems and human operators to investigate how and to what extent the level of automation may affect the allocation of liability. As mentioned above, the healthcare system can be

<sup>37</sup> The Italian Supreme Court of Cassation, sez. IV, ruling no 18140/2012, stated that in the event of death caused by a defective medical device carrying a CE mark, it should be possible to apply the legitimate expectation doctrine, unless the defect is manifest and readily recognisable. In ruling no 40897/2011, the same court stated that with the full quality assurance procedure, the CE mark for class III devices would assume substantive relevance, since it provides the basis for legitimate expectation and the relationship of trust between the doctor-user and the notified body.

<sup>38</sup> HLEG, A. I. Ethics guidelines for trustworthy AI. B-1049 Brussels, 2019.

described as a complex STS, combining technological artefacts, social artefacts, and humans.<sup>39</sup>

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 automation of the technology at issue. Social artefacts, including norms and institutions, determine what should be done, governing tasks, obligations, goals, priorities, and institutional powers. Humans play an essential role in the functioning of STSs, including health care, providing them with governance and maintenance and sustaining their operation.<sup>40</sup>

In particular, the healthcare 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 are more often part of human-machine interaction processes.

In failure scenarios leading to patient injuries, a key aspect that should be considered in allocating liability is the level of automation of technological artefacts, 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.

On the one hand, the violation of such duties may result in personal liability for human experts.<sup>41</sup> Whenever there is a failure in a complex system, such a failure is usually connected with the non-execution or inadequate execution of a 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 and criminal law.

<sup>39</sup> Jan K.B. Olsen, Stig A. Pedersen, and Vincent F.A. Hendricks, *Companion to the Philosophy of Technology* (John Wiley & Sons 2012).

<sup>40</sup> Pieter Vermaas and others, 'A Philosophy of Technology: From Technical Artefacts to Sociotechnical Systems' (2011) 6(1) *Synthesis Lectures on Engineers, Technology, and Society* 1.

<sup>41</sup> Mark F. Grady, 'Why are people negligent? technology, nondurable precautions, and the medical malpractice explosion' (1987) (82) *Nw. UL Rev.* 293.

On the other side, it may be necessary to identify the task-responsibilities of AI systems, in other words the requirements they ought to meet. As task-responsibilities are progressively delegated to technology, the risk of liability for damage and injuries contextually shifts from humans to the organisations that designed and developed the technology and defined its context and uses, and are responsible for its deployment, integration, maintenance, and certification. Thus, responsibilities may change relative to the changing functionalities and automation levels that devices are taking on through the implementation of AI.

It is necessary to adopt a systematic approach<sup>42</sup> for matching automation levels to the different responsibilities of both human experts and AI systems.<sup>43</sup> Here, in order to determine how tasks ought to be allocated between human experts and AI CDSS, reliance is made on the Level Of Automation Taxonomy (LOAT),<sup>44</sup> based on the taxonomy developed by Endsley and Kaber,<sup>45</sup> and on the principles set out by Parasuraman, Sheridan, and Wickens.<sup>46</sup>

LOAT provides criteria for allocating tasks under four different cognitive functions: information acquisition (A), information analysis (B), decision-making (C), and action implementation (D). Figure 1 illustrates a simplified version of LOAT.<sup>47</sup> Each column starts with a 0-level of automation,

<sup>42</sup> Erik Hollnagel, 'The human in control: Modelling what goes right versus modelling what goes wrong' in *Human Modelling in Assisted Transportation* (Springer 2011) 3.

<sup>43</sup> Giuseppe Contissa and others, 'Liability and Automation: Issues and Challenges for Socio-Technical Systems' (2013) 2(1–2) *Journal of Aerospace Operations* 79.

<sup>44</sup> Luca Save and Beatrice Feuerberg, 'Designing Human-Automation Interaction: A New Level of Automation Taxonomy' (2012), In *Proc. Human Factors of Systems and Technology* 2012.

<sup>45</sup> David B Kaber and Mica R Endsley, 'The Effects of Level of Automation and Adaptive Automation on Human Performance, Situation Awareness and Workload in a Dynamic Control Task' (2004) 5(2) *Theoretical Issues in Ergonomics Science* 113.

<sup>46</sup> Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens, 'A Model for Types and Levels of Human Interaction with Automation' (2000) 30(3) *Systems, Man and Cybernetics, Part A: Systems and Humans*, *IEEE Transactions* 286.

<sup>47</sup> For a complete LOAT version see Luca Save and Beatrice Feuerberg, 'Designing Human-Automation Interaction: A New Level of Automation Taxonomy' (2012) in *Proc. Human Factors of Systems and Technology* 2012 (n. 44).

corresponding to a fully manual performance of a certain task, without any technical support.

**Figure 1: LOAT (simplified version)**

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

	Full Automatic Decision Making	Medium-Level Automat. of Action Seq. Execut.
		<b>D7</b> High-Level Automation of Action Seq. Execut.
		<b>D8</b> Full Automation of Action Sequence Exec

At level 1 the task is performed with 'primitive' technical tools, i.e. low-tech nondigital artefacts. From level 2 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 under the four cognitive functions, expressing varying levels of interaction between humans and technology.

In the following the IBM Watson system is considered as an example of AI CDSS, and its levels of automation are assessed. Even though this section is mainly focused on Watson, results can be extended to all AI CDSS systems sharing similar features and levels of automation. A complete technological analysis, especially with regard to the level of automation, should always be grounded in the technical specifications of the AI system in question and in its concept of operations. Watson was chosen as a focus of investigation because ample information is available about its functioning and architecture.<sup>48</sup> Additionally, Watson is a representative example of AI CDSS as reported in the literature.<sup>49</sup>

<sup>48</sup> For a general overview on Watson, see for instance Kevin D Ashley, *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age* (Cambridge University Press 2017)

<sup>49</sup> See, for example Alicja, Piotrkowicz, Johnson Owen, and Geoff Hall. 'Finding relevant free-text radiology reports at scale with IBM Watson Content Analytics: a feasibility study in the UK NHS.' (2019) 10(1) *Journal of biomedical semantics* 21. David Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T Mueller, 'Watson: beyond jeopardy!' (2013) *Artificial Intelligence* 199, 93–105; Marjorie Glass Zauderer, Ayca



As concerns information acquisition (A), Watson supports human experts in acquiring information on the process they are following. The system integrates data from different sources, such as personal health records, medical datasets containing domain-specific literature, and clinical trial reports. It then filters and/or highlights the relevant information items by selecting, for example, the results of clinical trials on cancer diseases rather than leukaemia. The criteria for integrating, filtering, and highlighting the relevant information are predefined at design level and not available to physicians. Thus, with regard to the first cognitive function, Watson reaches level A5 (full automation support of information acquisition).

As concerns the second cognitive function, namely the analysis of information (B), Watson compares and analyses the available data based on parameters defined at design level, reaching level B5 (full automation support of information analysis). In the LOAT classification, this level usually implies that the system triggers visual and/or sound alerts whenever a certain result requires human expert attention. Consider, for instance, an arrhythmia-detection alert generated by an electrocardiograph (ECG). Even though we can imagine a near future in which Watson will be connected to other kinds of medical devices, such as ECGs, the analysis of information lies in the internal process of the system, and it is not accessible to human experts.

With regard to 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 hypotheses (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 kinds of evidence extracted from a patient's electronic medical record (EMR) and other related sources of data. These kinds of evidence include symptoms, findings, patient history, family history, current medications, demographics, and so on. Each

Gucalp, Andrew S. Epstein, Andrew D. Seidman, Aryeh Caroline, Svetlana Granovsky, Julia Fu, Jeffrey Keesing, Scott Lewis, Heather Co, Het al. Piloting IBM Watson Oncology within Memorial Sloan Kettering's regional network, (2014) 32(15) Journal of Clinical Oncology 2014.

diagnosis links back to the original evidence that DeepQA uses to produce the associated confidence scores, and it supports the adoption of evidence-based medicine. Physicians can select any of the alternative diagnoses proposed by the system, or they can choose their own diagnosis, whenever, for example, they are aware of contextual circumstances (e.g. a certain medical condition, the patient's values, and others) unknown to or ignored by the system, as well as in cases where they have evidence of errors by the AI system. As a consequence, under the third cognitive function, the system reaches level C2 (automated decision support).

As it concerns action implementation (D), namely the administration of medical treatments, human experts (physicians, caregivers, etc) execute and control all actions without any kind of AI system intervention. Thus, Watson reaches level D0 (manual action and control).

It clearly appears that, even though Watson reaches full automation in information acquisition and analysis, physicians may play a central role with regard to the selection of decisions and actions, as well as to their implementation. This task allocation raises questions with regard to the source of decision-making authority, as analysed in the following section.

## **5. THE SOURCE OF DECISION-MAKING AUTHORITY AND THE ROLE OF WATSON IN HEALTH CARE**

In recent years, there has been an increased interest in examining the role of AI in decision-making and whether it should be used for supporting or augmenting human decision-making or rather for replacing and automating the whole process.<sup>50</sup> These technologies expand the scale of collected and processed evidence, broadening the questions about whether human experts

<sup>50</sup> See for example S Miller, 'AI: Augmentation, more so than automation' (2018) 5(1) Asian Management Insights 1–20. The author argues the imperative of a new human-machine symbiosis and calls for the rethink of 'how humans and machines need to work symbiotically to augment and enhance each other's capabilities'. See also J Wilson, and PR Daugherty, 'Collaborative Intelligence: Humans and AI are joining forces.' (2018) 96(4) Harvard Business Review 115–123; and Council of Europe, *Study on the Human Rights Dimensions of Automated Data Processing Techniques (in particular algorithms) and Possible Regulatory Implications* (2017) 3.

can still cope with the expertise and capacity of AI systems, and whether there is the need to rethink the role of humans in the decision-making process.

Given the characteristic features of Watson and those of new AI-CDSS sharing similar levels of automation, in particular with regard to information acquisition and analysis, this section investigates whether the source of decision-making authority should be attributed only to human experts (e.g. clinicians and physicians) or it should be completely shifted to AI systems or, finally, whether a shared decision-making model is possible and even preferable.

**Human decision-making authority.** In the first hypothesis, human decision-making authority, the AI system would be considered as a simple information-management tool supporting human experts. The standard of care would remain what is reasonable to expect from the average physician in the specific medical field in question.

However, AI technologies such as Watson are purposely designed to interfere with human-decision making:<sup>51</sup> they are used on the assumption that they can outperform humans, overcoming not only their cognitive limitations<sup>52</sup> but also time-sensitive ones in accessing, reading, understanding,<sup>53</sup> and incorporating evidence.<sup>54</sup> According to some scholars, this assumption would provide the basis for relinquishing control to AI CDSS, like Watson, as the better approach

<sup>51</sup> Andrew D Selbst, 'Negligence and AI's Human Users' (March 11, 2019) Boston University Law Review (forthcoming); UCLA School of Law, Public Law Research Paper No. 20-01, 16. Available at SSRN <<https://ssrn.com/abstract=3350508>> accessed 27 June 2020.

<sup>52</sup> See for example Deskus, C. 'Fifth Amendment Limitations on Criminal Algorithmic Decision-Making' (2018) NYUJ Legis. Pubs & Pubs. Pol'y 21, 237, 250, stating that human capacity for judgement is inferior to that of mathematical models when it comes to prognostic evaluations.

<sup>53</sup> Andrew D. Selbst, and Solon Barocas, 'The intuitive appeal of explainable machines' (2018) 87 Fordham L. Rev. 1085.

<sup>54</sup> Memorial Sloan-Kettering Cancer Center, 'IBM to Collaborate in Applying Watson Technology to Help Oncologists' press release <[press/us/en/pressrelease/37235.wss#resource](https://press/us/en/pressrelease/37235.wss#resource)> accessed 28 march 2019.

to reach the gold standard of evidence-based practice.<sup>55</sup> If there is strong evidence to suggest a particular diagnosis-and-treatment procedure, then that diagnosis and treatment is the most justifiable one.

**AI decision-making authority.** The second hypothesis, shifting the decision-making authority to AI CDSS, is indeed generally supported by two main arguments: (1) the normative pull of evidence-based practice,<sup>56</sup> which it would be questionable to ignore; and (2) the greater success rate over human experts. On this hypothesis, medical malpractice law would eventually require a superior ML-generated medical diagnosis as the standard of care in clinical settings.<sup>57</sup> As a consequence, medical experts, not being in a position to reach the same standard, would be bound by the decisions of AI systems, even in cases where such decisions go beyond their comprehension and control. In the event of failures resulting in patients being harmed or injured, any departure from the advice of an AI system may lead to the physician being held professionally liable for medical negligence.<sup>58</sup>

Relinquishing control to AI systems in medicine raises some legal issues with regard to (i) patients' peculiarities and the concept of evidence-based medicine; (ii) the role of explanation in decision-making; and (iii) the role of trust in medical practice.

### *1. Patients' peculiarities and the concept of evidence-based medicine*

The first issue has to do with patients' uniqueness and the concept of evidence-based medicine. Even though the latter is regarded as the gold standard, and

<sup>55</sup> Jason Millar and Ian R Kerr, 'Delegation, Relinquishment, and Responsibility: The Prospect of Expert Robots' (2013) SSRN <<http://dx.doi.org/10.2139/ssrn.2234645>> accessed 27 June 2020.

<sup>56</sup> Memorial Sloan-Kettering Cancer Center, 'IBM to Collaborate in Applying Watson Technology to Help Oncologists' (n. 54)

<sup>57</sup> A Michael Froomkin, Ian R Kerr, and Joëlle Pineau, 'When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning' (2019) 61(1) Arizona Law Review 33.

<sup>58</sup> Jeffrey M. Senger, and Patrik O'Leary, 'Big data and human medical judgment regulating next-generation clinical decision support' in Glenn Cohen, Holly F. Lynch, Effy Vayena, and Urs Gasser, *Big data, health law, and bioethics* (Cambridge University Press 2018)

is considered the best argument in favour of AI decision-making authority, a number of limitations and criticisms emerge when evidence-based medicine is applied to individual patients. These criticisms point to the occurrence of biological variations, the need to consider the individual patient's values, and the limits in describing evidence to patients in order to facilitate shared decision-making.<sup>59</sup> A broader understanding of evidence-based medicine 'requires a bottom up approach that integrates the best external evidence with individual clinical expertise and patients' choices.'<sup>60</sup>

Although it is true that the alternative to AI evidence-based diagnosis is not a perfect diagnosis but rather human diagnosis with all their flaws, the care process should be regarded as a complex and multidimensional concept. It can not only be based on the best external evidence supporting a specific diagnosis and treatment,<sup>61</sup> but should also consider the uniqueness of patients, their biological variations, and the diversity of individual values, moral attitudes, goals, and choices.

In this regard, medical experts cannot be reduced to that of mere executors of AI systems' advice or to that of intermediaries between AI CDSS and patients. In many cases, the best solution consists in integrating human and automated judgements by enabling physicians to review and eventually adapt the suggestions of AI to individual patients' goals and preferences. Moreover, the limitation in accessing and describing evidence is directly related to the second issue, namely the role of explanation in decision-making.

## *2. The role of explanation in decision-making*

The second issue concerns AI explainability and accountability, and the possibility of obtaining human-intelligible and human-actionable information. As noted in section 0, AI CDSS like Watson are essentially black-box systems, in

<sup>59</sup> Sharon E Straus and Finlay A McAlister, 'Evidence-Based Medicine: A Commentary on Common Criticisms' (2000) 163(7) CMAJ 837.

<sup>60</sup> David L Sackett and others, *Evidence Based Medicine: What It Is and What It Isn't* (British Medical Journal Publishing Group 1996).

<sup>61</sup> Alex John London, 'Artificial intelligence and black-box medical decisions: accuracy versus explainability' (2019) (n. 27).

other words opaque systems<sup>62</sup> that provide diagnosis and treatment recommendations without supporting explanations. They lack the capability to explain the internal process of reasoning behind the decision-making, or the reasons why decisions are made in a certain way and/or why they are recommended. These decisions, in other words, do not come with any supporting justifications. Medical experts make diagnoses by relying on multiple sources of knowledge, such as scientific literature, relevant past cases, and their trained common sense. They also use these sources of knowledge for generating explanations and ground their diagnoses and treatment decisions.

The question is whether and to what extent statistical evidence provided by AI CDSS like Watson – referring to probabilities or statistical relationships between certain symptoms and diagnoses or between specific treatments and recovery – is sufficient to provide an exhaustive explanation. The explainability of AI systems is required as well under Articles 13 and 14 GDPR, according to which 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing' is to be provided when decision-making is automated. Indeed, AI explainability has recently become central in the scientific debate as one of the core principles in developing AI systems, along with the principles of beneficence, non-maleficence, autonomy, and justice.<sup>63</sup> Some authors have raised the question whether the explanation should provide an account of (a) *all* the patterns and variables taken into account by the system (a model-centric explanation) or (b) only those that are relevant to the specific patient's case (subject-centric explanation).<sup>64</sup> Regardless of the ability to outperform human experts, the

<sup>62</sup> Jenna Burrell, 'How the machine "thinks": Understanding opacity in machine learning algorithms' (2016) 3(1) *Big Data & Society* 1, 3.

<sup>63</sup> Luciano Floridi, Josh Cows, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge et al. 'AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations' (2018) 28(4) *Minds and Machines* 689-707; Andrew D. Selbst, and Solon Barocas, 'The intuitive appeal of explainable machines' (2018) 87 *Fordham L. Rev.* 1085.

<sup>64</sup> Lilian Edwards and Michael Veale, 'Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For' (2017) 16 *Duke L & Tech Rev.* 18.

explanation plays an essential role in the medical decision-making process for both medical experts and patients.

To properly understand the concept of explanation and its role within the health domain, we need to focus on who the explanation is provided for. As noted by Miller,<sup>65</sup> explanation can also be seen as a communication problem. From this perspective, it is necessary to consider the interaction between two roles, explainer and explainee, recognising that there are certain 'rules' that govern this interaction. Indeed, the concept of explanation may assume different meanings, being subject to specific rules, depending on what perspective is adopted. Furthermore, different aspects may be relevant, depending on whether the explainee is the medical expert or the patient.

From a computer-science perspective the explanation needs to include three elements. First of all, it needs a model explanation, i.e. an interpretable and transparent model, capturing the whole logic of the obscure system. Secondly, it requires a model inspection, i.e. a human-comprehensible representation of the specific properties of an opaque system and its prediction, making it possible to understand how the black box behaves internally depending on the input values, namely its sensitivity to certain attributes (e.g. specific symptoms), up to and including, for instance, the connections in a neural network. Finally, it needs an outcome explanation, making it possible to understand the reasons for certain decisions, i.e. the causal chains leading to a certain outcome in a particular instance.<sup>66</sup>

While the first two models, the model explanation and the model inspection, seem to be mostly directed at computer scientists and IT experts, the outcome explanation is also relevant for medical experts, for a variety of reasons.

<sup>65</sup> Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences.' (2019) 267 *Artificial Intelligence* 1-38.

<sup>66</sup> See Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 'A survey of methods for explaining black box models.' 51(5) *ACM computing surveys (CSUR)* (2019) 93; Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad, 'Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission' (2015) *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721-1730. ACM

First, research in social science suggests that providing explanations for recommended actions deeply influences users' confidence in, and acceptance of, AI-based decisions and recommendations.<sup>67</sup> From this perspective, medical experts would benefit from causal explanation, providing the rationales behind AI decisions and facilitating further investigations. Physicians should be able to assess the coherence of the arguments supporting the suggestions of the system in relation to the medical literature, clinical practice, past cases similar to the one in question, and individual patients. The explanation would also enable physicians to determine the extent to which a particular input was determinative or influential in yielding the output<sup>68</sup> and to evaluate whether and to what extent they can rely on the AI CDSS recommendations.

For instance, it may prove necessary to determine whether a patient's interests were taken into account in recommending a certain diagnosis and treatment, as well as whether a certain factor (e.g. a certain symptom, the patient's age) was crucial in determining the diagnosis at issue and the suggested treatment. From this perspective, the role of medical experts remains central in considering factors which may affect decisions, such as symptoms that AI CDSS are unable to perceive (e.g. a specific body odor or the consistency of tissue to the touch) and the patient's values, attitudes, and preferences. As experts in the medical domain, physicians are the only ones who can integrate such factors with the evidence and suggestions provided by AI CDSS.<sup>69</sup> All these factors are necessary for eventually identifying possible counterarguments, which, if taken into account, may lead to different decisions. Thus, medical experts should play an oversight and monitoring role. This is even more

<sup>67</sup> L Richard Ye and Paul E Johnson, 'The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice' (1995) 19(2) MIS Quarterly 157.

<sup>68</sup> Finale Doshi-Velez and Mason Kortz, 'Accountability of AI under the Law: The Role of Explanation' [2017] Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper <<http://nrs.harvard.edu/urn-3:HUL.InstRepos:34372584>> accessed 27 June 2020.

<sup>69</sup> For an overview, see Eliza Strickland, 'Ibm watson, heal thyself: How IBM overpromised and underdelivered on ai health care' (2019) 56(4) IEEE Spectrum 24. See also Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR' (2017) 31 Harv. JL & Tech. 841.



relevant and necessary if we consider that current AI CDSS do not support a meaningful explanation function.

On the other hand, explanation is essential for patients as well, making it possible to ensure a patient-centered care process, informed decision-making with regard to care and treatment, and ultimately the acceptability of medical advice. Thus, if not only physicians but also patients are considered as addressees of the explanation, its dialectical dimension becomes crucial, in particular to make the explanation accessible and comprehensible to non-domain experts and to laypersons.

From this perspective, social scientists have focused on the communicative aspect of explanation, arguing for the following approaches<sup>70</sup>: (i) contrastive explanation; (ii) selective explanation; (iii) causal explanation; and (iv) social explanation. While contrastive explanation is used to specify what input values determined the adoption of a certain decision (e.g. treating the condition with certain drugs) rather than possible alternatives (e.g. recommending a different drug or a surgical procedure), selective explanation is based on those factors that are most relevant according to human judgments. The latter is the case since causal chains are often too large to comprehend,<sup>71</sup> especially for those who lack the specific domain competence, such as patients. Causal explanation focuses on causes, rather than on merely statistical correlations. If we consider patients as addressees, the most likely explanation is not always the best explanation. Referring to probabilities and statistical generalizations, provided by AI CDSS, is not as effective as referring to causes; for example, a certain diagnosis or medical treatment can be explained by the patient's clinical condition, rather than by the kind of symptoms that are common to patients affected by a certain disease. Finally, the explanation has a social nature. It is useful to adopt an interactive and conversational approach in which information is tailored to the recipient's beliefs and way of understanding. For instance, physicians may need to keep track of the state of the explanation by

<sup>70</sup> Tim Miller, 'Explanation in artificial intelligence: Insights from the social sciences.' (2019) *Artificial Intelligence* 267, 1-38.

<sup>71</sup> Kake M. Hofman, Amit Sharma, and Duncan J. Watts, 'Prediction and explanation in social systems' (2017) 355(6324) *Science* 486–488.

noting what has been already communicated to the patient and inferring what the patient has inferred him/herself.

In this dialectical sense, the role of medical experts would remain essential not only in making explanations accessible and meaningful to patients, but also in tailoring such explanations to individual patients, possibly considering their emotional state and reactions as well. Even if we imagine a future where AI systems will be able to provide human-understandable evidence and explanations, physicians would not be reduced to acting as mere intermediaries, for two reasons. First, only medical experts have the specific domain knowledge needed to interpret the pull of evidence and explanation — assuming AI explainability — and to evaluate its reliability and correctness. Secondly, in the ability to explain lies the keystone of the interaction and relationship of trust between doctors and patients across the entire care process as they cooperate in devising a treatment.

### *3. The role of trust in medical practice*

As a consequence, the third issue pertains to trust. Trust is traditionally considered a cornerstone of interpersonal relationships,<sup>72</sup> and in health care it is regarded as the effective foundation of the patient-doctor relationship. The need for interpersonal trust is owed to the patient's vulnerability, to the information asymmetry deriving from the specialistic nature of medical knowledge,<sup>73</sup> and to the uncertainty regarding the skills and intentions of the physician, on whom the patient is dependent. Where trust is concerned, arguing in favour of the decision-making authority of AI CDSS would necessarily undermine the patient-doctor relationship, which would be replaced with a patient-AI system relationship. This would ultimately lead to a concurrent transfer of the trustee role from medical experts to AI CDSS.

The patient-doctor trust relationship can, for different reasons, be argued to be still essential in the care process. First, medical competence encompasses

<sup>72</sup> Roger C Mayer, James H Davis, and F David Schoorman, 'An Integrative Model of Organizational Trust' (1995) 20(3) *Academy of Management Review* 709.

<sup>73</sup> Monika Hengstler, Ellen Enkel, and Selina Duelli, 'Applied Artificial Intelligence and Trust—the Case of Autonomous Vehicles and Medical Assistance Devices' (2016) 105 *Technological Forecasting and Social Change* 105.

more than knowledge, judgment, and skill in technical functions; it also lies in the ability to help patients feel at ease, conversing with them sensitively and effectively to elicit relevant symptoms and patient's concerns, and providing responsive and meaningful feedback.<sup>74</sup> Removing such interpersonal human skills from the trust relationship may undermine the patient's trust in the competence of AI CDSS, even leading to a mistrust and unwillingness to follow the advice of AI.

A further reason has to do with the information asymmetry owed to the specialistic nature of medical knowledge. Even though this asymmetry also shapes the relationship between the medical expert and the AI CDSS, the imbalance would be even greater when it comes to patients, since they cannot be expected to have any domain-specific knowledge and would thus typically never be able to understand and interpret data and assess evidence and explanations. A meaningful understanding of the data, as well as the ability to access evidence and explanations, is essential to making informed decisions about whether to opt in or to opt out of AI recommendations.

**The shared decision-making model.** Given the criticisms just mentioned, neither the human decision-making authority model nor the AI decision-making authority model is supported here. Both models fail to fully explain the allocation of tasks and roles in the interaction between medical experts and AI CDSS in the healthcare STS. Thus, a shared decision-making authority model is here advocated. This model rests on the concept of a joint cognitive system. It has been observed that when humans and AI systems interact in working toward a goal, it would be better to describe humans and technology not as two interacting 'components' but as making up a joint cognitive system, where control is shared between the human cognitive system and the AI system.<sup>75</sup> Thus, tasks traditionally associated with the role of physician will be attributed to the joint cognitive system, so that they are distributed between the human expert and the AI CDSS. From this perspective, the standard of care would result from a combination of the standard of care for medical practice and the

<sup>74</sup> David Mechanic, 'The Functions and Limitations of Trust in the Provision of Medical Care' (1998) 23(4) *Journal of Health Politics, Policy and Law* 661.

<sup>75</sup> Erik Hollnagel and David D Woods, *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering* (CRC Press 2005).

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 CDSS.

As a result, the human should maintain the ability to oversee the AI CDSS overall activity (including its legal and ethical impact in the care process) and the ability to decide whether and how to use the system and rely on its recommendations. In case of failure resulting in injuries for patients, liability should be assessed taking account the task allocation as discussed in section 0. The shared model allows physicians grounding their decisions not only in the pool of literature and clinical evidence, but also in the individual patient's biological variation, values, and preferences, as well as in factors the AI CDSS is unable to properly perceive, including their emotional state and beliefs.

The reliability of a decision will be based on both statistical evidence and the physician's ability to interpret such evidence — at least when it comes to detecting whether or not there is good evidence contradicting the AI suggestion or evidence of errors by the AI CDSS — and to provide meaningful explanations to patients.

This model leads to a three-dimensional trust relationship involving the AI CDSS, the human expert, and the patient. In the context of AI, control over the system is constitutive of trust.<sup>76</sup> As noted, given the specialistic nature of medical knowledge, such control can be exercised only by medical experts, at least partly, while avoiding the risk of exacerbating the information asymmetry between AI and patients.

The patient-doctor trust relationship would remain unchanged, relying on the full and deep concept of medical competence.

In conclusion, AI CDSS cannot replace the human expert as the source of decision-making authority, which remains essential when interpreting evidence, detecting AI CDSS errors, and providing explanations to patients. Furthermore, the human expert is needed in order to take account of the

<sup>76</sup> Cristiano Castelfranchi and Rino Falcone, 'Trust and Control: A Dialectic Link' (2000) 14(8) *Applied Artificial Intelligence* 799.

patient's legal and ethical values and principles, preferences and morality, and other information not available or accessible to such systems.

## **6. VARIATIONS ON A THEME: POSSIBLE FAILURES AND LIABILITY SCENARIOS**

In the previous sections, the levels of automation of Watson and its influence and role in the decision-making process have been analysed. The findings provide the basis for assessing the connection between delegation and responsibilities. In particular, this section provides variations on some possible failures in the decision-making process and the related liability assessment in the event of injuries suffered by a patient as a consequence of misdiagnosis and/or improper treatment.

As previously noted, Watson is used to analyse symptoms, make a diagnosis, and find the most appropriate treatment for specific diseases. In particular, it acquires the relevant information, integrating data from different sources, and analyses the available data. The system generates a number of hypotheses, before going through a process of evidence-testing.

Watson collects and classifies all potentially emerging diagnoses and the respective therapeutic plans, assigning specific confidence scores to them and ranking answers according to the probability of their being correct. In this way, the system supports the adoption of evidence-based medicine, taking the best available evidence obtained from the scientific method and applying that evidence to medical decision-making through an abductive reasoning process in the form of inference to the best explanation.<sup>77</sup>

As an example, it will be helpful to consider a case where a patient dies as a consequence of misdiagnosis or improper medical treatment. In order to assess the allocation of liability, we have designed four main scenarios. Each scenario is related to a failure in the execution of a specific cognitive function in the decision-making process.

<sup>77</sup> Charles Sanders Peirce, 'Abduction and Induction' in Justus Buchler (ed), *Philosophical Writings of Peirce* (Dover 1955) 150.

### *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. In this scenario, two different hypotheses can be considered:

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

Here, some information—such as a personal health record, the literature dataset, or the clinical trial reports—is missing, incorrect, or incomplete. We are dealing with an error not in the acquisition phase but rather in the source information. Watson may not be able to detect such an error, which might be owed to different causes, such as a human error (by physicians, nurses, knowledge engineers and so on) in collecting and recording the information, or a technical failure in the medical examination process (for example an ECG malfunction). Under this hypothesis, it seems that liability cannot be attributed to the medical staff that is using Watson or to the actors involved in the system development and certification process.

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

In this scenario, the failure is caused by an error in retrieving and selecting relevant information in making a diagnosis and recommending a medical treatment. According to the classification laid out in section 0, Watson reaches level A5 (full automation support of information acquisition). As noted, the criteria for integrating, filtering, and highlighting the relevant information are defined in advance at design level and are not available to physicians. As a consequence, liability may be attributed to the actors involved in defining such criteria and in the design process. Actors involved in the certification process, such as the notified body and members of the expert panel, may be found liable only if they were involved in evaluating and assessing the system design. Under this hypothesis, 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.

It may be asked whether the system user interface should be designed so as to alert the human expert if some critical information is unavailable or unreadable. Consider, for instance, the case in which Watson, failing to detect that a certain patient is pregnant, recommends drugs that cannot be

administered to pregnant women, in that they may cause serious problems in the foetus. In these cases, additional liabilities may be attributed to the manufacturer for the defective design of the interface (not providing the alert) and to the medical staff for ignoring the missing-information alert.

It should be noted that since the criteria for the acquisition of information are defined at design level, if the system is certified under the full quality-assurance procedure, the legitimate expectation principle should shield the human expert from liability in choosing to trust the system and its ability to carry out the delegated task. The only exception lies in cases where the human expert is aware or should have been aware that some relevant information was missing, or when there is evidence that he or she was negligent in ignoring the missing-information alert.

## *2. Failure in the information-analysis phase*

Also worth considering are cases of failure in the information-analysis phase, involving the generation of a diagnosis, 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 laid out in section 0, Watson reaches level B5 (full automation support of information analysis). As noted, the parameters for comparing and analysing the available data are defined in advance at design level (and may not be visible to physicians, and in any case may not be meaningful for humans).<sup>78</sup> Under this hypothesis, liability may be attributed to the manufacturer, where a design defect or a manufacturing defect occurs as a consequence of selecting and implementing certain parameters in the design process, as well as to the notified body and members of the expert panel, if they were involved in evaluating and assessing the system design and functioning.<sup>79</sup>

<sup>78</sup> Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, 'A survey of methods for explaining black box models.' (2019) *ACM computing surveys (CSUR)* 51(5) 93:2.

<sup>79</sup> Following the PIP Breast Implant Case (C-219/15) scandal, in which the EUCJ states that the Medical Device Directive does not create a right to patients to obtain damages from notified body, a number of additional measures have been taken by the European Commission, including the new Medical Device Regulation 2017/745 (MDR), in force from 2020. The latter has indeed tightened the regulatory framework

Also worth considering is the case in which the system may trigger visual and/or sound alerts, requiring attention by medical staff, as in the previously introduced ECG example. If the failure is causally linked to such a functionality (because it is defective or missing), liability will be attributed to the manufacturer, possibly for product defect. Conversely, members of the medical staff may be found liable if the failure is attributable to their behaviour, consisting, for instance, in negligently ignoring an alert.

in which notified bodies operate. First of all, the public law supervision of the activities of notified bodies has been intensified, so that the Member States retain the ultimate control on the market. Article 6 of Annex XI provides for the possibility that the Member State in which the notified body is based assumes liability for the actions of notified bodies. With regard to the State liability, see for example Carola Glinski and Peter Rott. 'The role and liability of certification organisations in transnational value chains.' (2018) *Deakin L. Rev.*, 23: 83. This suggestion at the same time indicates that a privately organised scheme may fail in serving the public interest. See also Rob Van Gestel, and Hans-W. Micklitz, 'European integration through standardization: how judicial review is breaking down the club house of private standardization bodies' (2013) *Common Market L. Rev.*, 50: 145. While a tighter Member States' control improves public supervision, in no way suggests that notified bodies themselves are not responsible. In this regard it is important to highlight that, the new MDR includes a number of important changes with regard to the obligations of notified bodies. Interestingly, all the obligations claimed in Schmitt (C-219/15) under the previous Medical Device Directive, e.g. to carry out unannounced inspections at least once in every five year (Article 3.4 of Annex IX, taking samples of the certified products to assess whether they comply with the design dossier (Article 3.4 of Annex IX), have now been expressly incorporated in the Regulation. Moreover, notified bodies have to verify that the amount of raw materials used by the manufacturer is consistent with the number of products which have been manufactured (Article 3.5 of Annex IX). Beyond the conformity assessment procedure, notified bodies will have to satisfy new requirements as regards inter alia their organisational structure, independence and impartiality, qualifications of personnel and contracted experts (Articles 1.1, 1.2, 3.2 and 3.4 of Annex VII). For a general overview, see Peter Rott, 'Certification of Medical Devices: Lessons from the PIP Scandal' in *Certification—Trust, Accountability, Liability* (Springer 2019), 189-211; and Paul Verbruggen and Barend Van Leeuwen, 'The liability of notified bodies under the EU's new approach: The implications of the PIP breast implants case.' (2018) 43(3) *European Law Review* 394-409.



As in the previous scenario, the parameters for analysing information are defined at design level. Thus, if the system is certified under the full quality-assurance procedure, the legitimate expectation principle should shield the human expert from liability in choosing to trust the system and its ability to carry out the delegated task. The only exception would be the case where the human expert negligently ignored an alert.

Additionally, since AI CDSS like Watson are capable of analysing and processing massive amounts of information<sup>80</sup> 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 experts the legal duty to be 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 that have emerged from information analysis, Watson generates a ranked list of diagnoses with associated confidence scores, proposing alternative diagnoses and the associated treatments. It thus leaves clinicians the possibility and freedom to select the best hypothesis and/or to generate alternative options. According to the classification set out in section 0, Watson reaches level C2 (automated decision support). In this scenario, different hypotheses may be considered.

**Hypothesis 1:** Watson generates a correct diagnosis, along with an associated treatment. In the following, four different sub-hypotheses are considered:

- a) The diagnosis and the associated treatment generated by Watson are both correct, and the human expert follows its suggestion. This case is relatively unproblematic, since no conflict emerges between the human expert and the AI system, and no failures can be detected at the decision-and-action-selection stage.
- b) The diagnosis and the associated treatment are both correct, but the human expert does not follow the system suggestion; (s)he may, for instance, generate a new diagnosis or a different treatment. Under this sub-hypothesis, a failure may emerge from the divergent human

<sup>80</sup> Millar, J., and Kerr, I., 'Delegation, relinquishment, and responsibility: The prospect of expert robots' in *Robot Law* (Edward Elgar Publishing 2016) 105.

expert's decision. From a liability perspective, some authors have noted that the outcome depends on which expert judgment will be considered as the source of the decision-making authority.<sup>81</sup> In particular, if Watson is considered as such a source, then liability can be attributed to human experts (e.g. the liability of physicians) under a specific duty to follow the advice of the system. Any divergent decision should be considered a violation of such a duty. However, as noted in section 0, given the trust relationship<sup>82</sup> between patients and doctors, it is debatable whether Watson should be considered a decision-making authority. Conversely, both on human-expert and shared decision-making authority models, 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,<sup>83</sup> and may be considered the effective cornerstone for the applicability of the legitimate expectation principle.

- c) The diagnosis is correct, but the associated treatment is wrong, and the human expert follows the suggestion of the system. One might want to consider here the case where the wrong treatment derives from an internal failure of the system in generating the medical treatment. In this case, the manufacturer may be found liable for the defective technology, and so may 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 based solely on following the suggestion of the system, with the exception of cases where they had good evidence contradicting the advice of the system or had evidence-based reasons for not trusting such advice, e.g. on the basis of wrong results in similar previous cases. Thus, on the shared decision-

<sup>81</sup> Millar, J., and Kerr, I. *Delegation, relinquishment, and responsibility: The prospect of expert robots in Robot Law* (Edward Elgar Publishing 2016) 118; Selbst AD, 'Negligence and AI's Human Users' Boston University Law Review [Forthcoming] 16.

<sup>82</sup> Thom, D. H., and Campbell, B. 'Patient-Physician Trust: An exploratory study' (1997) 44( 2) The Journal of family practice 169.

<sup>83</sup> European Commission, *White Paper On Artificial Intelligence - A European approach to excellence and trust* (2020) 23.

making-authority model, the liability shield can be grounded in the application of the legitimate expectation principle whenever the system has been certified under the full quality-assurance procedure and the former relies on a 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 a medical condition of the patient unknown to or ignored by Watson, as in the example of drugs administered to pregnant women.

- d) The diagnosis is correct, but the associated treatment is wrong, and the human expert does not follow the suggestion of the system. This case is relatively unproblematic with regard to a possible conflict between the human expert and the AI system. In the event of undesirable outcomes, the liability of human experts may derive only from their negligent behaviour and/or medical malpractice.

**Hypothesis 2:** Watson generates a wrong diagnosis and an associated treatment. In the following, two relevant sub-hypotheses are considered:

- a) Both the diagnosis and the associated treatment generated by Watson are wrong, and the human expert follows the suggestion of the system. In this case, the manufacturer may be found liable for the defective technology, and so may 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 debatable whether the liability of human experts may be based solely on their having followed the advice of the system, with the exception of cases where they had good evidence contradicting the suggestion of the system or evidence-based reasons for not trusting such advice, e.g. on the basis of wrong results in similar previous cases. As noted, under the full quality-assurance procedure, the liability shield should be grounded not in the human expert's delegation of such authority to the AI system but rather in the application of the legitimate expectation principle.
- b) Both the diagnosis and the associated treatment are wrong, but the human expert does not follow the suggestion of the system. Even though a conflict between the human expert and the AI system emerged, this case remains unproblematic, since undesirable outcomes

may only result from the negligent behaviour of clinicians and/or their medical malpractice.

#### *4. Failure in the action-implementation phase*

In this scenario, a possible failure may only result from the human expert's behaviour, as in cases where caregivers overdose the drugs to be administered. As noted in section 0, under LOAT, Watson reaches 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, for example clinicians and caregivers, as a result of negligent behaviour and/or medical malpractice.

### **7. CONCLUSION**

In this contribution, the liability issues emerging from the adoption of AI CDSS in healthcare was explored from a socio-technical perspective by analysing the technological features of new-generation AI CDSS compared to traditional ones; the regulatory framework in place, especially with regard to the legal qualification of AI CDSS and the certification procedures; and the allocation of decision-making tasks between medical experts and AI systems. The adopted systemic approach shed light on the functioning of the healthcare system, making it possible to assign liability by analyzing the human-machine interaction.

With regard to the technological component of the healthcare STS, the specific features of new AI CDSS are going to improve the quality of health care and patients' safety, given their ability to outperform medical experts in certain activities, such as clinical diagnosis and treatment recommendations.

However, we showed how such features coupled with and the highest level of automation in performing different cognitive tasks can have a stronger impact on both the decision-making process and the inherent risk posed by AI medical devices.

From the social-component perspective, the regulatory framework in place, and in particular the criteria for assessing the risk class of medical devices and the related conformity-assessment procedures, does not consider the level of

automation of a medical device as a risk factor. However, automation may affect medical practice, influencing or even directing clinicians' decisions. Thus, rather than focusing on the intended use of medical devices, the classification criterion should take account of the level of automation. Indeed, the latter may affect how the decision-making process is split between human experts (e.g. physicians) and AI systems, also becoming a criterion by which to assess possible liabilities in case of failure.

With regard to the interaction between medical experts and AI CDSS, some scholars considered their ability to outperform humans in diagnosis and recommendations as one of the main reasons to doubt that humans can still be considered as the source of decision-making authority. In fact, AI systems have demonstrated an ability to successfully act in a domain traditionally entrusted to the trained intuition and analysis of humans.

However, relinquishing control to AI systems presents some challenges. Although it is true that the alternative to AI diagnosis is not a perfect diagnosis, but rather human diagnoses with all their flaws, the care process should be regarded as a complex and multidimensional concept. It cannot only be based on the best external evidence supporting a specific diagnosis and treatment, but should also consider the uniqueness of patients, their biological variations and the diversity of individual values, moral attitudes, goals and choices. Medical experts cannot be reduced to mere executors of AI systems' advice or to intermediaries between AI CDSS and patients. In many cases, the best solution consists in integrating human and automated judgments by enabling physicians to review AI suggestions and patients to request a meaningful explanation of the diagnosis and the recommended treatment, taking account of its communicative and dialectical dimension. If the trustworthiness and explainability of AI are to be promoted, there will need to be an emphasis on transparency, while developing methods and technologies that enable human experts to analyse and review automated decision-making.

The future challenge will consist in finding the best combination between human intelligence and AI intelligence, taking into account the capacities and the limitations of both. On these grounds, an argument was made here in favour of a shared decision-making authority model, which relies on a broader understanding of evidence-based medicine and the care process. On this

model, the reliability of a decision will depend not only on the statistical evidence generated by the AI system but also on physicians' ability to interpret such evidence. From this perspective, the standard of care would be determined by combining the standard of care for human-expert medical practice and the standard resulting from ML-generated evidence-based diagnosis. This model also leads to a three-dimensional trust relationship involving the AI system, the human expert, and the patient. Finally, a shared model is consistent with the concept of a joint cognitive system and the allocation of tasks between humans and AI, where control is accomplished by coupling the human cognitive system with an AI system that exhibits goal-directed behaviour.

All these elements were taken into account in analysing liability under the existing regulatory framework, given the technological features of AI CDSS, their level of automation, and their interaction with medical experts. The ways in which activities and the related liabilities are attributed and distributed between humans and AI systems should also be taken into account in a proactive way, during the design phase of a new operational concept/system, to address possible legal issues arising from future potential accidents or malfunctions. The adoption of a socio-technical perspective also makes it possible to assess and improve the existing regulatory framework by analysing legal risks that AI technology introduces in complex STS.

In conclusion, if valuable practices surrounding the use of AI in the healthcare domain are to be promoted, it needs to be ensured that the development and deployment of AI tools takes place in a socio-technical framework — inclusive of technologies, human skills, organisational structures, and norms — where individual interests and the social good are both preserved and enhanced.