MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

University of Cyprus

MAI643 Artificial Intelligence in Medicine

Elpida Keravnou-Papailiou

January – May 2023

Co-financed by the European Union
Connecting Europe Facility

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

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Explainability in Medical AI

Co-financed by the European Union
Connecting Europe Facility

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

UNIT 2

Explainability in Medical AI

CONTENTS

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

INTENDED LEARNING OUTCOMES

Upon completion of this unit on Explainability in Medical AI, students will be able:

1. To enhance their understanding of the significance of explanation in relation to AI systems.
2. To discuss C.S. Peirce’s and P. Thagard’s general theories of explanation that have influenced the AI field, and the role of causality in the production of explanations.
3. To trace the history of explanations in symbolic AI, pointing out key milestones (rule-based explanations, strategic explanations, user-tailored explanations, case-based explanations).
4. To outline the recent resurgence of interest in explanation, in relation to connectionist AI, and the establishment of the research field referred to as XAI (eXplainable AI) aiming to ‘open’ the black box.
5. To point out explainability issues particular to medical AI and to present the key points of a recently coined manifesto on explainability for AI in medicine (definition of explainability, propositions, research directions).

Master programmes in Artificial Intelligence 4 Careers in Europe

The significance of explanation in AI

Co-financed by the European Union
Connecting Europe Facility

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

Why do AI systems need to give explanations?

There are many (critically) important reasons ……

- In general, decision support systems must be **interpretable** and not black-boxes – recall the roles of expert knowledge-based systems as consultants, critics or tutors.
- By and large AI systems are **interactive**, i.e., they do not just get an input, process it and give an output, but they engage in a dialogue with a human user, who needs to take a 'final' decision that could impact on another human (e.g., a patient) or an organization, the society, etc.
- Particularly **critical domains** are the medical/health care, legal, and defense domains.
- Explanations have at least a dual purpose: (i) **understanding the logic/model of the system**, also facilitating 'debugging' (e.g., revealing biases in logic/data and erasing them); (ii) **understanding the rationale of the recommended outcome of a specific consultation** and be convinced of its validity; recall that AI systems are complex software systems deploying algorithms and knowledge/data.

Co-financed by the European Union
Connecting Europe Facility

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

Why do AI systems need to give explanations?

- Traditionally, the central role of the explanation model is to reveal the system's reasoning; however, it has a subsidiary role in relation to **information-acquisition interactions**, that concerns individual items of information rather than the system reasoning processes:
  - The user needs to be able to ask, not only why the system is asking a particular question (i.e., how does it relate to the reasoning process), but also what the given question means.
- Nowadays the strive for **responsible, trustworthy and ethical AI**, emphasizes even more the need for AI systems to be bestowed with appropriate, user-tailored and hence **fit for purpose**, explanation models; different categories of users have different explanation needs.
- EU's General Data Protection Regulation (**GDPR**) and **ACM's Statement on Algorithmic Transparency and Accountability** make direct references to the need for explanation while the European Research Consortium for Informatics and Mathematics (ERCIM) devoted one of its special issues on **transparency in algorithmic decision making**.

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

GDPR

According to R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi and F. Giannotti, “A survey of methods for explaining black box models”, ACM Computing Surveys 51(5):93, 2018, DOI 10.1145/3236009:

An innovative aspect of the GDPR, which has been debated, are the clauses on automated (algorithmic) individual decision-making, including profiling, which for the first time introduce, to some extent, a right of explanation for all individuals to obtain **“meaningful explanations of the logic involved”** when automated decision making takes place. Despite divergent opinions among legal scholars regarding the real scope of these clauses, everybody agrees that the need for the implementation of such a principle is urgent and that it represents today a huge open scientific challenge. **Without an enabling technology capable of explaining the logic of black boxes, the right to an explanation will remain a “dead letter”.**

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

ACM Policy Council: Statement on Algorithmic Transparency and Accountability, 2017.
https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

usacm ACM US Public Policy Council

Principles for Algorithmic Transparency and Accountability

1. Awareness: Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.

2. Access and redress: Regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions.

3. Accountability: Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.

4. Explanation: Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made. This is particularly important in public policy contexts.

5. Data Provenance: A description of the way in which the training data was collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data-gathering process. Public scrutiny of the data provides maximum opportunity for corrections. However, concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to game the system can justify restricting access to qualified and authorized individuals.

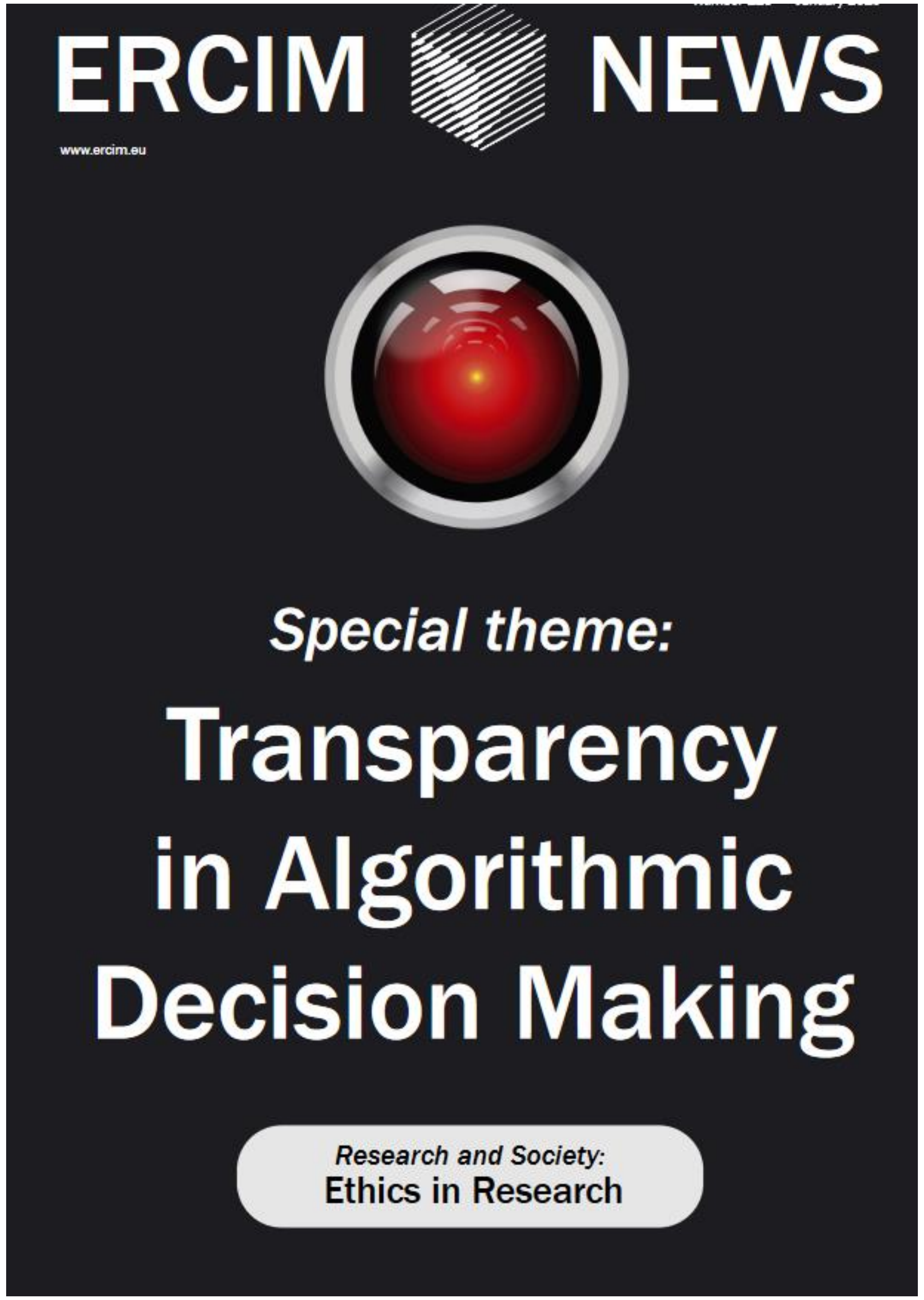
6. Auditability: Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.

7. Validation and Testing: Institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make the results of such tests public.

Co-financed by the European Union
Connecting Europe Facility

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

A. Rauber, R. Trasarti and F. Giannotti (eds.), Transparency in algorithmic decision making, Special theme, ERCIM News, Number 116, January 2019.
https://ercim-news.ercim.eu/images/stories/EN116/EN116-web.pdf



High-Level Expert Group on Artificial Intelligence

Sabine Theresia Köszegi, Professor of Labor Science and Organization Institute of Management Science, TU Wien, Chair of the Austrian Council on Robotics and Artificial Intelligence, BMVIT, Member of the High-Level Expert Group on Artificial Intelligence of the European Commission.

On 25 April 2018, the Europaen Commission published a Communication in which it announced an ambitious European Strategy for Artificial Intelligence (AI). The major advances in AI over the last decade revealed its capacity as a general-purpose technology and pushed inventions in areas of mobility, healthcare, home & service robotics, education and cyber security, to name just a few. These AI-enabled developments have the capability to generate tremendous benefits not only for individuals but also for the society as a whole. AI has also promising capabilities when it comes to address and resolve the grand challenges, such as climate change or global health and wellbeing, as expressed in the United Nations Sustainable Development goals. In competition with other key players, like the United States and China, Europe needs to leverage its current strengths, foster the enablers for innovation and technology uptake and find its unique selling proposition in AI to ensure a competitive advantage and a prosperous economic development in its Member States. At the same time, AI comes with risks and challenges associated to fundamental human rights and ethics. Europe therefore must ensure to craft a strategy that maximizes the benefits of AI while minimizing its risks.

The Commission has set out an interwoven strategy process between the development of a European AI Strategy and the development of a Coordinated Action Plan of Member States (hosted under the Digitising European Industry framework). The publication of the European policy and investment strategy on AI is envisaged for Summer 2019. To support this strategy development process and its implementation, the Commission has called for experts to establish a High-Level Expert Group on Artificial Intelligence (AI HLEG). Following an open selection process by DG Connect in spring 2018, the Commission has appointed 52 experts encompassing representatives from different disciplines of academia, including science and engineering disciplines and humanities alike, as well as representatives from industry and civil society. As an expert in labor science and with a research background in decision support systems, I was selected to join the exciting endeavor to lay the foundations for a human-centric, trustworthy AI in Europe that strengthens European competitiveness and addresses a citizen perspective to build an inclusive society.

Our mandate includes the elaboration of recommendations on the policy and investment strategy on ethical, legal and societal issues related to AI, including socio-economic challenges. Additionally, we serve as a steering group for the European AI Alliance to facilitate the Commission's outreach to the European society by engaging with multiple stakeholders, sharing information and gathering valuable stakeholder input to be reflected in our recommendations and work.

On 18 December 2018, we proposed a first draft on "Ethics Guidelines towards Trustworthy AI" to the Commission, setting out the fundamental rights, principles and values that AI has to comply with in order to ensure its ethical purpose. Additionally, we have listed and operationalized requirements for trustworthy AI as well as provided possible technical and non-technical implementation methods that should provide guidance on the realization of trustworthy AI. This draft on ethics guidelines is currently in a public consultation process in the European AI Alliance platform. Through this engagement with a broad and open multi-stakeholder & citizen forum across Europe and beyond, we aim to secure the open and inclusive discussion of all aspects of AI development and its impact on society. The finalised draft will be formally presented in the First Annual Assembly of the European AI Alliance in Spring 2019.

To advise the Commission with regards to the European policy and investment strategy, we are currently preparing a set of recommendations on how to create a valuable ecosystem for AI in Europe in order to strengthen Europe's competitiveness. The draft document of recommendations should be published in April 2019 and will undergo a public consultation process as well. The recommendations will primarily address European policy makers and regulators but also relevant stakeholders in Member States encompassing investors, researchers, public services and institutions. I would like to use the opportunity, to invite the readers of ERCIM News to engage in the European AI Alliance (see the link below) and to contribute your expertise and input to our policy and investment recommendations.

The complexity of AI-related challenges requires to set up a problem-solving process with highest information processing capacities that allows to consider different perspectives and to resolve conflicts of interest between different stakeholders. It can easily be imagined that our discussions as an inter-disciplinary expert and multi-stakeholder group are intense, difficult and at times emotional. In difficult situations, I remind myself of our commitment to the following statement in our ethics guidelines: "Trustworthy AI will be our north star, since human beings will only be able to confidently and fully reap the benefits of AI if they can trust the technology."

AI Alliance:
https://ec.europa.eu/digital-single-market/en/european-ai-alliance

Co-financed by the European Union
Connecting Europe Facility

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

Some theories of explanation that have influenced AI

- C.S. Peirce’s hypothesis of abduction – finding the most likely explanation of a set of observations
- P. Thagard’s theory of explanatory coherence

Co-financed by the European Union
Connecting Europe Facility

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

C.S. Peirce’s hypothesis of abduction

Has its origins on Peirce’s **architecture of theories** (https://arisbe.sitehost.iu.edu/menu/library/bycsp/arch/arch.htm)

The Architecture of Theories

By Charles S. Peirce

The Monist, v. I, n. 2, 1891 January, pp. 161–176. At Google Books. At Internet Archive.
Reprinted: *Writings* v. 8 (2010), 199–211; *The Essential Peirce* v. 1 (1992), 285–297; *Collected Papers* v. 6 (1931), paragraphs 7–34.
Also: *Logic of Interdisciplinarity* (2009), 58–69; *Values in a Universe of Chance* (1958), 142–159; *Philosophical Writings* (1940), 315–323; *Chance, Love and Logic* (1923), 157–178.

OF the fifty or hundred systems of philosophy that have been advanced at different times of the world's history, perhaps the larger number have been, not so much results of historical evolution, as happy thoughts which have accidently occurred to their authors. An idea which has been found interesting and fruitful has been adopted, developed, and forced to yield explanations of all sorts of phenomena. The English have been particularly given to this way of philosophising; witness, Hobbes, Hartley, Berkeley, James Mill. Nor has it been by any means useless labor; it shows us what the true nature and value of the ideas developed are, and in that way affords serviceable materials for philosophy. Just as if a man, being seized with the conviction that paper was a good material to make things of, were to go to work to build a *papier mâché* house, with roof of roofing-paper, foundations of pasteboard, windows of paraffined paper, chimneys, bath tubs, locks, etc., all of different forms of paper, his experiment would probably afford valuable lessons to builders, while it would certainly make a detestable house, so those one-idea'd philosophies are exceedingly interesting and instructive, and yet are quite unsound.

Co-financed by the European Union
Connecting Europe Facility

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

From the Stanford Encyclopedia of Philosophy

https://plato.stanford.edu/entries/abduction/index.html#DedIndAbd

Abduction

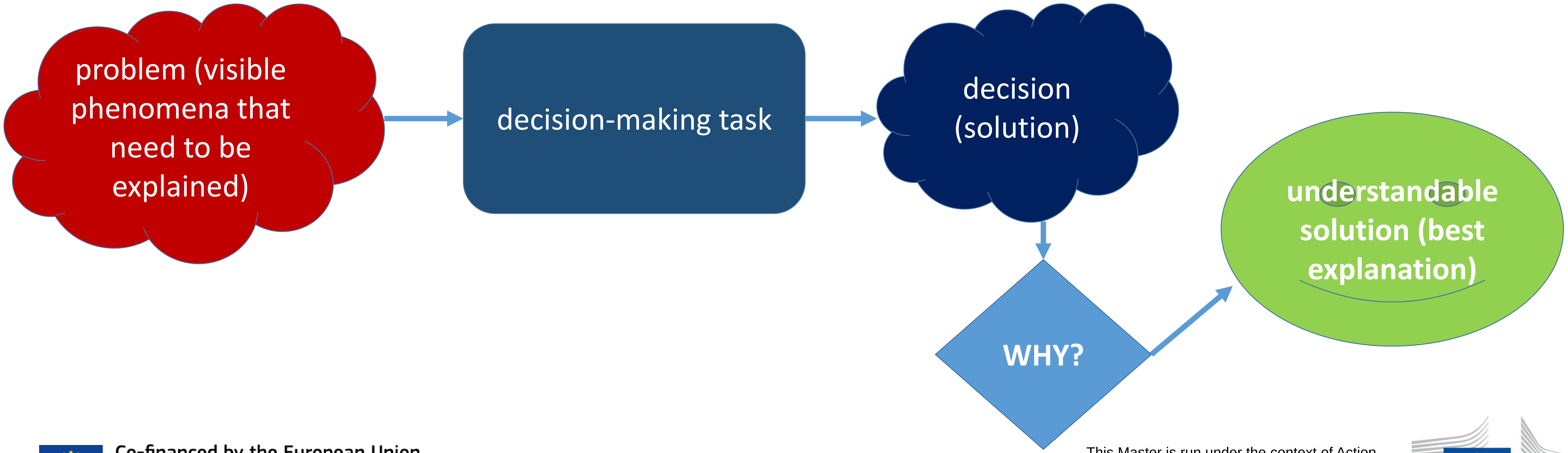
First published Wed Mar 9, 2011; substantive revision Tue May 18, 2021

In the philosophical literature, the term "abduction" is used in two related but different senses. In both senses, the term refers to some form of explanatory reasoning. However, in the historically first sense, it refers to the place of explanatory reasoning in *generating* hypotheses, while in the sense in which it is used most frequently in the modern literature it refers to the place of explanatory reasoning in *justifying* hypotheses. In the latter sense, abduction is also often called "Inference to the Best Explanation."

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

Relevance to AI decision-making tasks

- Any decision-making task strives to reach a decision that constitutes the best solution and hence **best explanation** of the problem at hand, whether it refers to a classification, prediction, plan of action, etc.
- Explanations are essential when decisions are critical, unclear or not easily understandable.



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

The basic reasoning methods

- **Abduction**: Formulates hypotheses, making a combined space to create new ideas
- **Deduction**: Tests hypotheses to narrow down existing choices
- Recall **hypothetico-deductive model** of reasoning leading to best explanation, where deduction is a sub-process of abduction: Contextualized versus unconstrained deductions
- **Induction**: Reaches conclusions and generalizes existing ideas
- **Explanations arise as rational connections between hypotheses and observations**

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

H.E. People's Mechanization of Abductive Logic

https://www.ijcai.org/Proceedings/73/Papers/017.pdf

- In a **deduction**, the objective is to determine **whether** some statement is true
- In an **abduction**, the objective is to determine **why** something is true (i.e., why the observed abnormalities hold)
- In answering the why question, it is obviously important to be able to determine whether, thus deduction may be considered a process subordinate to deduction

Co-financed by the European Union
Connecting Europe Facility

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

Abduction and Deduction

- Abduction is far more complicated than deduction.
- A queried statement may be deduced (derived) in a multitude of ways, and any of these suffices; effective deductive systems are able to follow the simplest derivation paths, but this is an implementation rather than a conceptual issue.
- In abduction, it is not sufficient just to generate one **plausible explanation** of the observed situation; instead, all plausible explanations need to be compared and contrasted.
- **An explanation is usually not deducible**, and so once an explanation is hypothesized, it is not possible to deduce it.

Co-financed by the European Union
Connecting Europe Facility

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

(i) What are the plausible explanations and (ii) How is the best explanation selected?

- Peirce has not specified any criteria …
- A trend in abductive diagnosis has been to explore how much can be achieved with somewhat restrictive and thus nonpragmatic criteria
- **Explanation plausibility**: complete accounting (coverage) of all observations of abnormality irrespective of their relative importance, say, for therapy
- Two celebrated theories of **abductive diagnosis** are based on this restricted notion of explanation plausibility:
  - Peng and Reggia's parsimonious covering theory
  - Poole's logic-based theory
  - The principle used to select the best explanation from the plausible ones is that of **simplicity**

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

Peng and Reggia's parsimonious covering theory

Parsimonious criteria based on:

- **Relevancy** – every disorder hypothesis included in an explanation is **causally related** to some observation of abnormality
- **Irredundancy** – none of the proper subsets of an explanation is itself an explanation
- **Minimality** – prefer the explanation with the minimum cardinality

Co-financed by the European Union
Connecting Europe Facility

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

Pool’s logic-based theory

Criteria based on:

- **Minimality** – prefer the explanation that makes the fewest, in terms of set inclusion, assumptions
- **Least presumption** – prefer the explanation that makes the fewest, in terms of what can be implied, assumptions
- **Minimal abnormality** – prefer the explanation that makes the fewest failure assumptions or makes the same abnormality assumptions but fewer normality assumptions

Co-financed by the European Union
Connecting Europe Facility

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

MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

P. Thagard’s theory of explanatory coherence

http://cogsci.uwaterloo.ca/Articles/1989.explanatory.pdf

BEHAVIORAL AND BRAIN SCIENCES (1989) 12, 435–502
Printed in the United States of America

Explanatory coherence

Paul Thagard
Cognitive Science Laboratory, Princeton University, 221 Nassau St., Princeton, NJ 08540
Electronic mail: *pault@confidence.princeton.edu*

Abstract: This target article presents a new computational theory of explanatory coherence that applies to the acceptance and rejection of scientific hypotheses as well as to reasoning in everyday life. The theory consists of seven principles that establish relations of local coherence between a hypothesis and other propositions. A hypothesis coheres with propositions that it explains, or that explain it, or that participate with it in explaining other propositions, or that offer analogous explanations. Propositions are incoherent with each other if they are contradictory. Propositions that describe the results of observation have a degree of acceptability on their own. An explanatory hypothesis is accepted if it coheres better overall than its competitors. The power of the seven principles is shown by their implementation in a connectionist program called ECHO, which treats hypothesis evaluation as a constraint satisfaction problem. Inputs about the explanatory relations are used to create a network of units representing propositions, while coherence and incoherence relations are encoded by excitatory and inhibitory links. ECHO provides an algorithm for smoothly integrating theory evaluation based on considerations of explanatory breadth, simplicity, and analogy. It has been applied to such important scientific cases as Lavoisier’s argument for oxygen against the phlogiston theory and Darwin’s argument for evolution against creationism, and also to cases of legal reasoning. The theory of explanatory coherence has implications for artificial intelligence, psychology, and philosophy.

Keywords: artificial intelligence; attribution theory; coherence, connectionism; epistemology; explanation; legal reasoning; scientific reasoning; theory evaluation

Co-financed by the European Union
Connecting Europe Facility

P. Thagard’s theory of explanatory coherence

Thagard is quite emphatic about the need for a **tight coupling between the formation and evaluation of hypotheses** in computational, abductive systems.

More specifically, he says that there are three possible models:

1. the two processes are completely independent, and hypotheses are formed in a random fashion, a nonviable option under limited resources;
2. the processes are weakly related, and only hypotheses that explain at least something are formed, or
3. they are strongly related, and only hypotheses that constitute likely possibilities are formed.

He also points out the inability (of some AI systems) to recognize those **observations in need of explanation** (this is a limitation because not every observation demands explanation) and subsequently the need to identify **how evaluation constraints can be used more effectively** to help limit the range of hypotheses that can be generated in order to lead to ones more likely to be accepted.

Thagard's general criteria for measuring the quality of explanatory hypotheses

- ❑ **Consilience** which is concerned not only with how much a hypothesis explains but also the variety of things it explains; a hypothesis is **dynamically consilient** if it becomes more credible over time
- ❑ **Simplicity** which is concerned with the number of supporting assumptions, the well-known Occam's razor: What can be done with fewer assumptions is done in vain with more
- ❑ **Analogy** which advocates the reusability of successful explanation models in analogous situations
- ❑ The above notions have been incorporated in the theory of **explanatory coherence**.

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

Abductive Diagnosis using Time-Objects: criteria for the evaluation of solutions

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1069.4339&rep=rep1&type=pdf

Computational Intelligence, Volume 17, Number 1, 2001

ABDUCTIVE DIAGNOSIS USING TIME-OBJECTS: CRITERIA FOR THE EVALUATION OF SOLUTIONS

ELPIDA T. KERAVNOU
Department of Computer Science, University of Cyprus

JOHN WASHBROOK
Department of Computer Science, University College London

Diagnostic problem solving aims to account for, or explain, a malfunction of a system (human or other). Any plausible potential diagnostic solution must satisfy some minimum criteria relevant to the application. Often there will be several plausible solutions, and further criteria will be required to select the "best" explanation. Expert diagnosticians may employ different, complex criteria at different stages of their reasoning. These criteria may be combinations of some more primitive criteria, which therefore should be represented separately and explicitly to permit their flexible and transparent combined usage.

In diagnostic reasoning there is a tight coupling between the formation of potential solutions and their evaluation. This is the essence of abductive reasoning. This article presents an abductive framework for diagnostic problem solving. *Time-objects*, an association of a property and an existence, are used as the representation formalism and a number of primitive, general evaluation criteria into which time has been integrated are defined. Each criterion provides an intuitive yardstick for evaluating the space of potential solutions. The criteria can be combined as appropriate for particular applications to define plausible and best explanations.

The central principle is that when time is diagnostically significant, it should be modeled explicitly to enable a more accurate formulation and evaluation of diagnostic solutions. The integration of time and primitive evaluation criteria is illustrated through the Skeletal Dysplasias Diagnostician (SDD) system, a diagnostic expert system for a real-life medical domain. SDD's notions of plausible and best explanation are reviewed so as to show the difficulties in formalizing such notions. Although we illustrate our work by medical problems, it has been motivated by consideration of problems in a number of other domains (fermentation monitoring, air and ground traffic control, power distribution) and is intended to be of wide applicability.

Key words: diagnostic problem solving, temporal abductive diagnosis, diagnostic solution, time-object, evaluation criteria.

Primitive Evaluation Criteria

- **Coverage**: focus-coverage, hard-coverage, current-coverage
- **Consistency**: case-consistent
- **Strength of integration**: strongly-integrated (or coherent), loosely-integrated (or incoherent); single or multiple point of failure
- **Satisfiability**: N/T/C-satisfiable (necessary, typical, common expectations)
- **Ambiguity**: alternative explanations for focus-abnormalities
- **Redundancy**: a strict subset has the same coverage
- **Minimality**: not redundant
- **Optimality**: has focus-coverage, and it is case-consistent, satisfiable, strongly-integrated and minimal

Causality

The notion of causality is strongly coupled to the ‘quality’ of explanations:

- Theories of explanation implicitly or explicitly entail causality, e.g., Peng and Reggia’s relevancy criterion states that every disorder hypothesis included in an explanation is **causally related** to some observation of abnormality
- **Causal models** are deeper than associational (rule-based) models and can provide justifications to associationally derived solutions; recall the case of NEOMYCIN and many other second-generation knowledge-based systems
- Association can arise between variables having causation or those not having causation; hence causality implies association but not the opposite

Causal Explanation

- Is the **strictest form of explanation**
- It arises from the construction of causal models, which require that explanations for arising predictions are, in fact, “recipes” for reconstructing that prediction
- A causal model captures directed causal relationships, usually in a graphical representation:
  - Either forwards in time (A causes B) or backwards in time (B caused-by A)
  - Richness of temporal and other semantics varies in different models

Co-financed by the European Union
Connecting Europe Facility

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

Judea Pearl, the father of Bayesian networks and probabilistic reasoning states …

"To build truly intelligent machines, teach them cause and effect"

Paradoxical association due to lack of causal knowledge …

Asthmatics are less likely to die from pneumonia!

Co-financed by the European Union
Connecting Europe Facility

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

EL

Is it true that asthmatics are less likely to die from pneumonia

There is some evidence to suggest that asthmatics may be at lower risk of dying from pneumonia compared to non-asthmatics. However, it is important to note that the relationship between asthma and pneumonia is complex, and the outcome can depend on several factors.

One study published in the European Respiratory Journal in 2018 found that among patients hospitalized for community-acquired pneumonia, those with a history of asthma had a lower risk of death compared to those without asthma. The study authors suggested that this may be due to the fact that asthmatics are more likely to be prescribed inhaled corticosteroids, which can reduce inflammation and may have a protective effect against pneumonia.

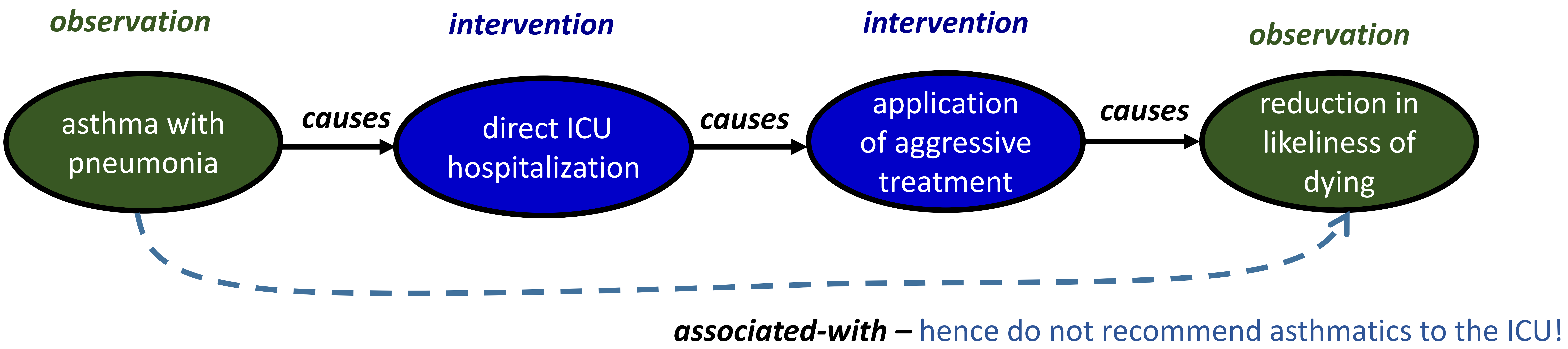
Another study published in the Journal of Allergy and Clinical Immunology in 2020 found that asthma was not associated with a higher risk of severe COVID-19 outcomes, including death, in hospitalized patients. However, the study authors noted that the relationship between asthma and pneumonia in COVID-19 patients is still not fully understood.

It is important to note that these studies do not definitively prove that asthmatics are less likely to die from pneumonia, and individual outcomes can vary based on several factors, including the severity of the asthma and the type and severity of the pneumonia. It is also important for asthmatics to continue to take measures to manage their condition and reduce their risk of respiratory infections, including getting vaccinated against pneumonia and following their doctor's recommended treatment plan.

Co-financed by the European Union
Connecting Europe Facility

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

AI needs more WHY



Pearl's three levels of causality:

1. **Association:** invokes purely statistical relationships defined by the data – *What does a symptom tell me about a disease?*
2. **Intervention:** not just observing what is, but changing what one sees, e.g., reliably estimating the effect if one performs an action – *If I take a baby Aspirin, will my risk of heart failure reduce?*
3. **Counterfactuals:** reasoning about hypothetical situations which enables us to estimate the *unobserved* outcomes (this is *abduction*) – *Was the Aspirin that saved me from a heart attack?*

A. Lavin, "AI needs more why", Forbes, 2019.

https://www.forbes.com/sites/alexanderlavin/2019/05/06/ai-needs-more-why/#70ea2a5f156d

- The pneumonia example shows that without considering clinical **contexts**, counterintuitive predictions and models with unintended consequences can be derived.
- By taking into consideration **domain expertise of the hospital's policy**, level 2 causal structure (clinical context, i.e., interventions) can be added.
- The incorporated knowledge in the form of causal graph depicts which associations in the observed data are assumed to be **valid cause-effect relationships**.
- However, this is not enough since relationships caused by action policies, won't necessarily generalize when the policy changes.
- Reliable decision support models need to learn counterfactual objectives; for levels 2 and 3 Pearl proposes the use of do-calculus, a formalism for causal logic.

Co-financed by the European Union
Connecting Europe Facility

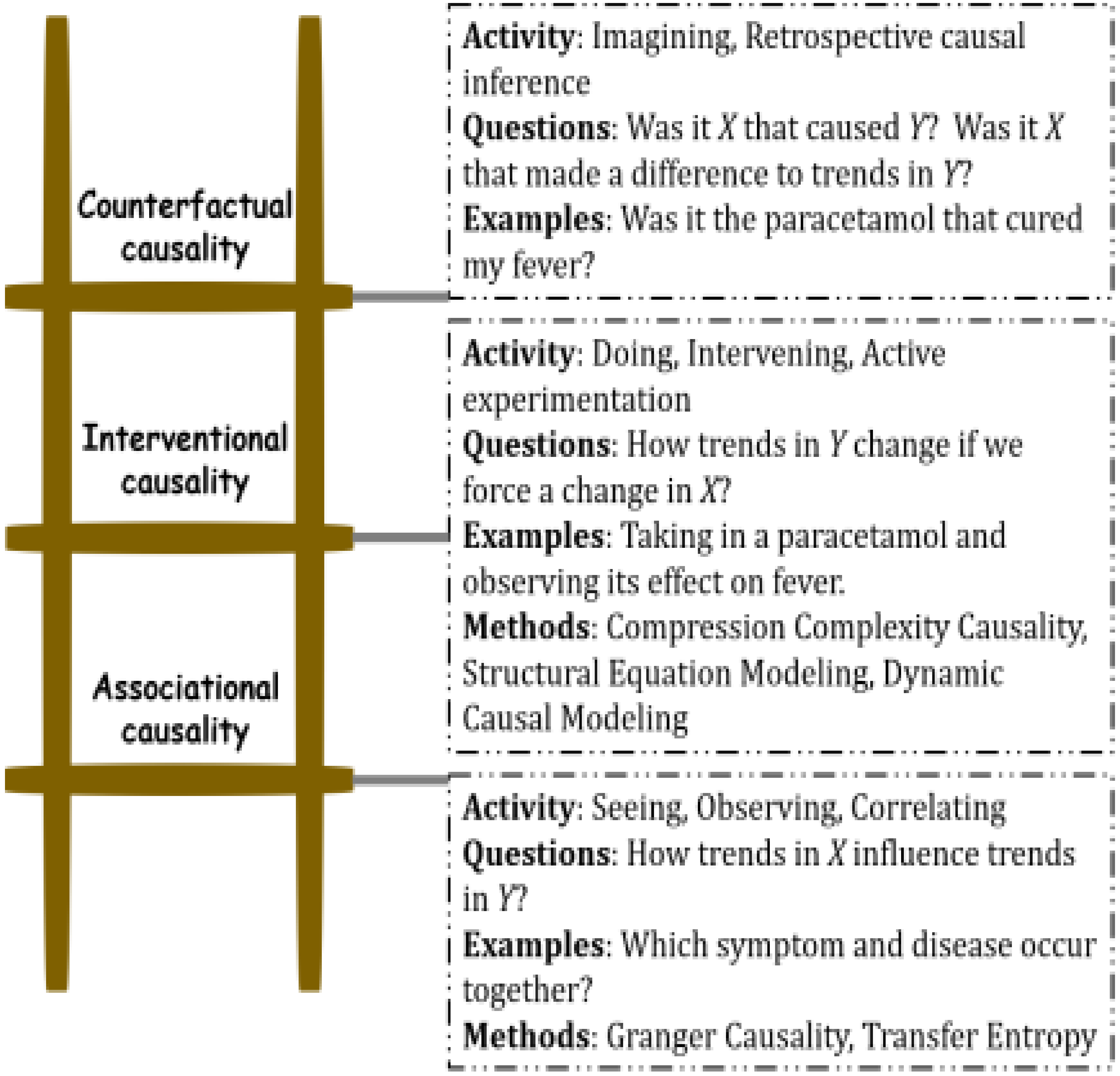
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Challenge: infer causality from purely observational data

J. Pearl's stance: "Causal reasoning is an indispensable component of human thought that should be formalized and algorithimitized toward achieving human-level machine intelligence."

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

The Ladder of Causation

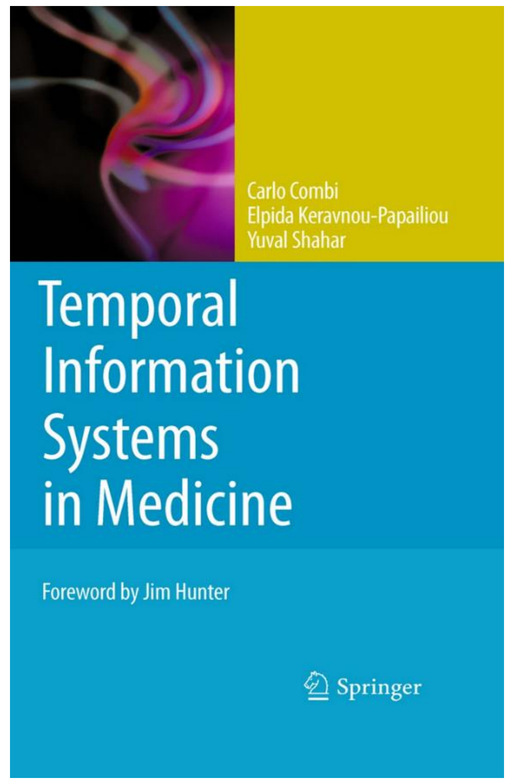


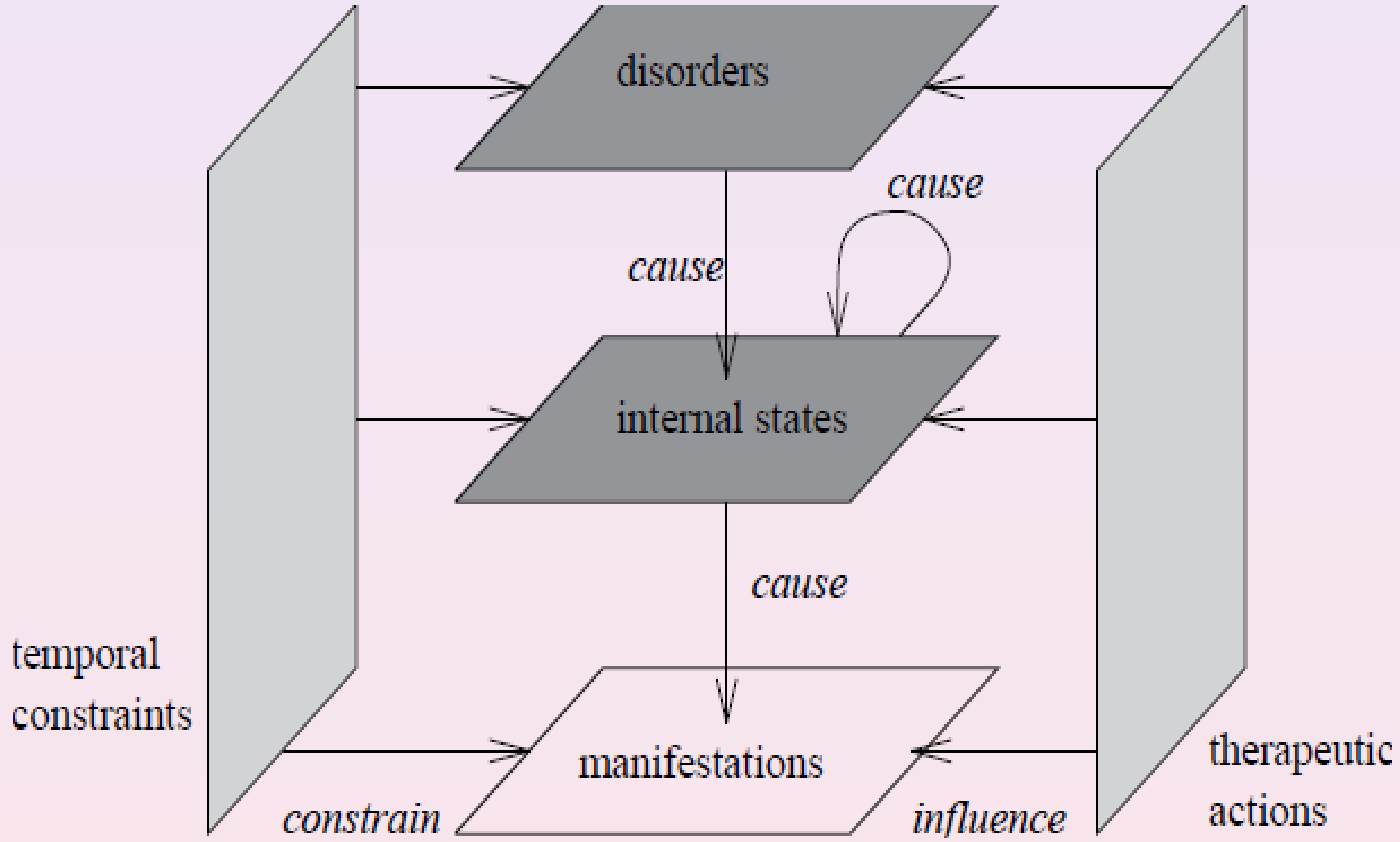
J. Pearl and D. MacKenzie, *The Book of Why: The new science of cause and effect*, Basic Books, 2018

A conceptual ladder of causation was introduced in Pearl and MacKenzie's **"Book of Why" for classifying causal queries by the amount and types of causality used**. The first level is **seeing**, the second is **doing** and the third level of the ladder is **imagining**.

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

The Causal-Temporal-Action (C-T-A) Model





Causal inference can be leveraged to reason explicitly about actions-and-effects underlying observational data.

Tracing the history of explanations in symbolic AI

Co-financed by the European Union
Connecting Europe Facility

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

Kim et. al., A multi-component framework for the analysis and design of explainable AI (https://www.mdpi.com/2504-4990/3/4/45)

“Explanations have always been an indispensable component of decision making, learning, understanding, and communication in the **human-in-the-loop** environments. After the emergence and rapid growth of artificial intelligence as a science in the 1950s, an interest in interpreting underlying decisions of intelligent systems also proliferated.”

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

Rule-based expert systems championed explanations in symbolic AI

- The MYCIN system pioneered symbolic explanations for different purposes:
  - **Justification** of the system's recommendations (MYCIN – 'intelligent' consultant)
  - Knowledge-base **debugging** (TEIRESIAS – 'intelligent' debugger)
  - **Tutoring** medical students (GUIDON – 'intelligent' tutor)

- Revealing **chains of rules** in the derived inference trees, also giving unsuccessful rules; pseudo natural language presentation
- Presenting the **current confidence** in (context-attribute-value) derivations stored in the context tree (working memory)
- Presentation and **comparative analysis of full inference trees** and other explanatory features of TEIRESIAS
- **Canned text** for individual rules; presentation of rules independently of specific consultations

Co-financed by the European Union
Connecting Europe Facility

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

Despite their pioneering significance, problems soon surfaced with rule-based explanations, deeming them largely inadequate ….

- ❑ "Explanations" were **just rule playbacks** and not meaningful
  - Missing/implicit knowledge
  - No support (causality) or strategic knowledge
- ❑ User-tailored explanations subsequently added through a **rudimentary user model**
  - Complexity, importance of concepts and rule associations
  - User level of knowledge/detail of explanations
- ❑ Adequate explanations to be an **inborne feature of the design** of a knowledge-based system from the start and not a subsequent add-on or reengineered into the system
  - Differentiating, explicating and implementing relevant knowledge types (e.g., causality)
  - Modelling human expertise (factual and reasoning knowledge) – **bottleneck**!

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

Second-generation, deep knowledge-based systems, offered new, promising avenues towards more adequate symbolic explanations …

- NEOMYCIN explicated important knowledge types utilized in explanations
  - Support knowledge in the form of a **causal model**, having a dual purpose
    - As an alternative means to solving problems
    - For augmenting rule-based explanations with a more detailed/deep justification
  - Strategic knowledge, enabling the provision of **strategic explanations**

- GUIDON2 was more successful than GUIDON as an ‘intelligent’ tutoring system

- Still **many challenges remained**, e.g.,
  - Explaining the rational basis of strategies
  - Revoking choices (including strategic choices) and/or derivations and explaining these
    - Inadequacies with reasoning, truth maintenance, non-monotonicity
  - Handling and justifying exceptions
  - User tailoring

Co-financed by the European Union
Connecting Europe Facility

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

Case-based explanations

- **CBR** offered yet another paradigm to symbolic explanations
- **Contextualized**, evidence-based explanations
- The similarity between the current case and the retrieved/selected past case needs to be explained
- Where solution adaptation is made, this would also need to be explained
- In many domains the case-based element is the domineering element in decision making, e.g., legal system in Cyprus
  - A past case sets **precedence** for future similar cases – fair/consistent handling
  - A decision is justified based on past cases – transparency, trust
- Repeating a successful past solution for a new similar case is sufficient explanation on its own without requiring further justification – **It worked for a similar case in the past!**
- Listing unsuccessful past cases, similar to the new case, provides further explanation/ justification for not adopting their (erroneous) solution and opting for something different
  - Avoiding past mistakes and reinforcing successes – **learning** from them

Co-financed by the European Union
Connecting Europe Facility

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

MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

The resurgence of interest in explanation in connectionist AI

– opening the ‘black box’ and the creation of the acronym XAI (eXplainable AI)

Co-financed by the European Union
Connecting Europe Facility

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

The knowledge acquisition bottleneck of symbolic AI systems, coupled with the performance success of Machine Learning and more recently Deep Neural Networks triggered interest in data-driven approaches.

But a new challenge emerged … making the resulting, highly-performing "black boxes", interpretable and explainable!

Co-financed by the European Union
Connecting Europe Facility

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

Not all ML approaches result in ‘black boxes’; e.g., decision trees are not, and symbolic rules can result from each branch from root to leaf of such trees; hence a decision tree can be flattened into a set of if-then rules.

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

Some survey papers on the topic of eXplainable AI (XAI) ….

- M-Y Kim et. al., A multi-component framework for the analysis and design of explainable AI (https://www.mdpi.com/2504-4990/3/4/45)
- G. Vilone and L. Longo, Classification of explainable AI methods through their output formats (https://www.mdpi.com/2504-4990/3/3/32)
- A.B. Arrieta et. al., Explainable AI (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI (https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103)
- R. Guidotti et. al., A survey of methods for explaining black box models (https://www.researchgate.net/publication/322976218_A_Survey_of_Methods_for_Explaining_Black_Box_Models)

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

The opening remarks of these surveys ….

- M-Y Kim et. al.: “The rapid growth of research in explainable artificial intelligence (XAI) follows on two substantial developments. First, the enormous application success of modern machine learning methods, especially deep and reinforcement learning, having created high expectations for industrial, commercial, and social value. Second, the emerging and growing concern for creating ethical and trusted AI systems, including compliance with regulatory principles to ensure transparency and trust.”

- G. Vilone and L. Longo: “Machine and deep learning have proven their utility to generate data-driven models with high accuracy and precision. However, their non-linear, complex structures are often difficult to interpret. Consequently, many scholars have developed a plethora of methods to explain their functioning and the logic of their inferences.”

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

The opening remarks of these surveys ….

- A.B. Arrieta et. al.: “In the last few years, AI has achieved a notable momentum that, if harnessed appropriately, may deliver the best of expectations over many application sectors across the field. For this to occur shortly in Machine Learning, the entire community stands in front of the barrier of explainability, an inherent problem of the latest techniques brought by sub-symbolism (e.g., ensembles or Deep Neural Networks) that were not present in the last hype of AI (namely, expert systems and rule-based models).”

- R. Guidotti et. al.: “In the last years many accurate decision support systems have been constructed as black boxes, that is as systems that hide their internal logic to the user. This lack of explanation constitutes both a practical and an ethical issue.”

Co-financed by the European Union
Connecting Europe Facility

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

Opening the Black Box

- ❑ Explaining the black box model
- ❑ Explaining the outcome
- ❑ Inspecting the black box internally
- ❑ Providing a transparent solution

Source: R. Guidotti et. al., A survey of methods for explaining black box models (https://www.researchgate.net/publication/322976218_A_Survey_of_Methods_for_Explaining_Black_Box_Models)

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

Black box and comprehensible predictors

- A **black box** predictor ***b*** belongs to the set of **uninterpretable** data mining and machine learning models
  - The reasoning behind the function is not understandable by humans and the outcome returned does not provide any clue for its choice
  - In real-world applications, ***b*** is an opaque classifier
- A **comprehensible** predictor is one for which a global or a local explanation is available; its performance is generally evaluated by two measures:
  - **Accuracy**: comparing the real target values against the respective predicted target values of the black box and comprehensible predictors
  - **Fidelity**: how good is the comprehensible predictor in mimicking the black box predictor

Explaining the black box model

Given a black box predictor b and a dataset $D = \{X, Y\}$, the **black box explanation problem** consists in finding a function f which takes as input a black box b and a dataset D, and returns a comprehensible global predictor c_g, i.e., $f(b, D) = c_g$, such that c_g is able to mimic the behavior of b and exists a global explanator function that can derive from c_g a set of explanations modeling in a human understandable way the logic behind c_g .

For example, the set of explanations can be modelled by a decision tree or by a set of rules.

Co-financed by the European Union
Connecting Europe Facility

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

Explaining the outcome

Given a black box predictor ***b*** and a dataset ***D* = {*X*, *Y*}**, the **black box outcome explanation problem** consists in finding a function ***f*** which takes as input a black box ***b*** and a dataset ***D***, and returns a comprehensible local predictor c_l, i.e., $f(b, D) = c_l$, such that c_l is able to mimic the behavior of ***b*** and exists a local explanator function that takes as input the black box ***b***, the comprehensible local predictor c_l and a data record ***x***, and returns a human understandable explanation for the record ***x***.

The various approaches proposed to implement function ***f***, aim to overcome the limitations of explaining the whole model. The returned explanation may be either a path of a decision tree or an association rule.

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

Inspecting the black box internally

Given a black box predictor ***b*** and a dataset ***D* = {*X*, *Y*},** the **black box inspection problem** consists in finding a function ***f*** which takes as input a black box ***b*** and a dataset ***D*** and returns a visual (or textual) representation of the behavior of the black box, i.e., ***f*(*b*, *D*) = *v*.**

For example, the visualization returned highlights the feature importance for the predictions. Overall, the aim is either to understand how the black box model works or why the black box returns certain predictions more likely than others.

Co-financed by the European Union
Connecting Europe Facility

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

Providing a transparent solution

Given a dataset ***D = {X, Y},*** the **transparent box design problem** consists in finding a learning function L_c which takes as input the dataset ***D*** and returns a (locally or globally) comprehensible predictor ***c***, i.e., $L_c(D) = c$.

This implies that there exists a local or a global explanator function that takes as input the comprehensible predictor ***c*** and returns a human understandable explanation or explanations.

For example, L_c and ***c*** may be the decision tree learner and predictor respectively, while the global explanator may return the choices taken along the various branches of the tree and the local explanator may return the textual representation of the path followed according to the (particular) decision suggested by the predictor.

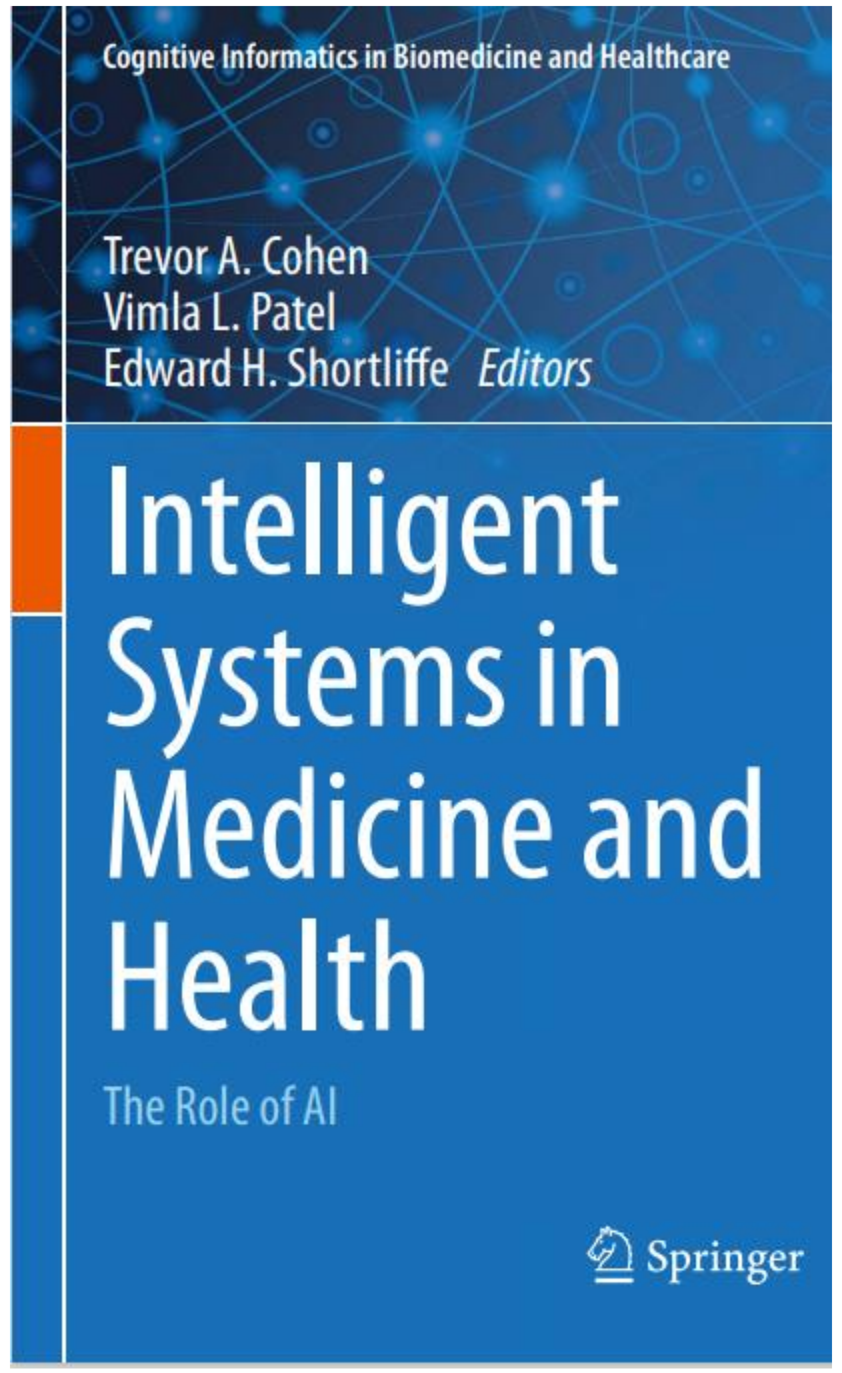
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Agnostic Explanator

- Is a comprehensible predictor, not tied to a particular type of black box, explanation or data type.
- In theory it can explain indifferently a neural network or a tree ensemble using a single tree or a set of rules.

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

Some insights from the following book chapter with respect to ML models



Chapter 8
Explainability in Medical AI

Check for updates

Ron C. Li, Naveen Muthu, Tina Hernandez-Boussard, Dev Dash, and Nigam H. Shah

Machine Learning (ML) Model

- *An ML model is a function learned from data that maps a vector of predictors to a real-valued response.*
- Such a model is considered explainable if the explanation satisfies the following two criteria:
  - It is "interpretable", i.e. the logic the model incorporates to make predictions is understandable by humans, and
  - It has fidelity, i.e. the explanation faithfully reflects the underlying logic of the task model (the model making predictions)

Co-financed by the European Union
Connecting Europe Facility

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

Does explainability truly enhance the usefulness of AI in health care?

Could explainability even lead to harm?

If explanations do not sufficiently satisfy the criteria of interpretability and fidelity, run the risk of giving users a false sense of security.

Co-financed by the European Union
Connecting Europe Facility

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

Explainability in Medical AI cannot be void of the context in which the model is deployed.

- ❑ Each of the following scenarios includes an AI solution.
- ❑ However, the nature of the task performed by the AI enabled tool and how it is incorporated into patient care differ.

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

1. An AI software product is used to analyze chest CTs as part of an automated system for lung cancer screening. Patients with chest CTs that are fagged by the AI software as high risk are automatically referred for biopsy.

2. A physician and nurse for a hospitalized patient each receives an AI generated alert that a patient for whom they both are caring is at risk of developing respiratory failure in the near future and recommends mechanical ventilation. They proceed to meet and discuss next steps for the patient’s clinical management.

3. A consumer smartwatch outfitted with AI capabilities, detects cardiac arrhythmias and notifies a user that an irregular heart rate has been detected recommending that the user consult a physician for further evaluation. After performing a full clinical assessment, the physician orders a continuous cardiac monitoring study for a formal diagnostic evaluation.

Co-financed by the European Union
Connecting Europe Facility

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

1

- *Here the system drives high stakes clinical care without any mediation by human clinicians.*

- As such it may be important for patients, as well as the clinicians, to understand the tool's reasoning behind its conclusions, similar to how a patient would want a physician to explain the reasoning behind a cancer diagnosis.

- The health system employing this AI solution and regulatory bodies may also require in-depth understanding of how the ML model generates its predictions and the level of model performance for quality assurance.

Co-financed by the European Union
Connecting Europe Facility

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

2

- *The AI system interacts with human clinicians who need to synthesize the prediction with the rest of their clinical evaluation in order to make a decision about the patient's management.*

- The clinicians need to trust the tool for its advice to be adopted.

- However, the mechanics of how the ML model generated the prediction may be less important to the clinicians than a conceptual understanding of why the program predicted this patient to be at risk that they can mentally incorporate into the rest of their clinical assessment.

Co-financed by the European Union
Connecting Europe Facility

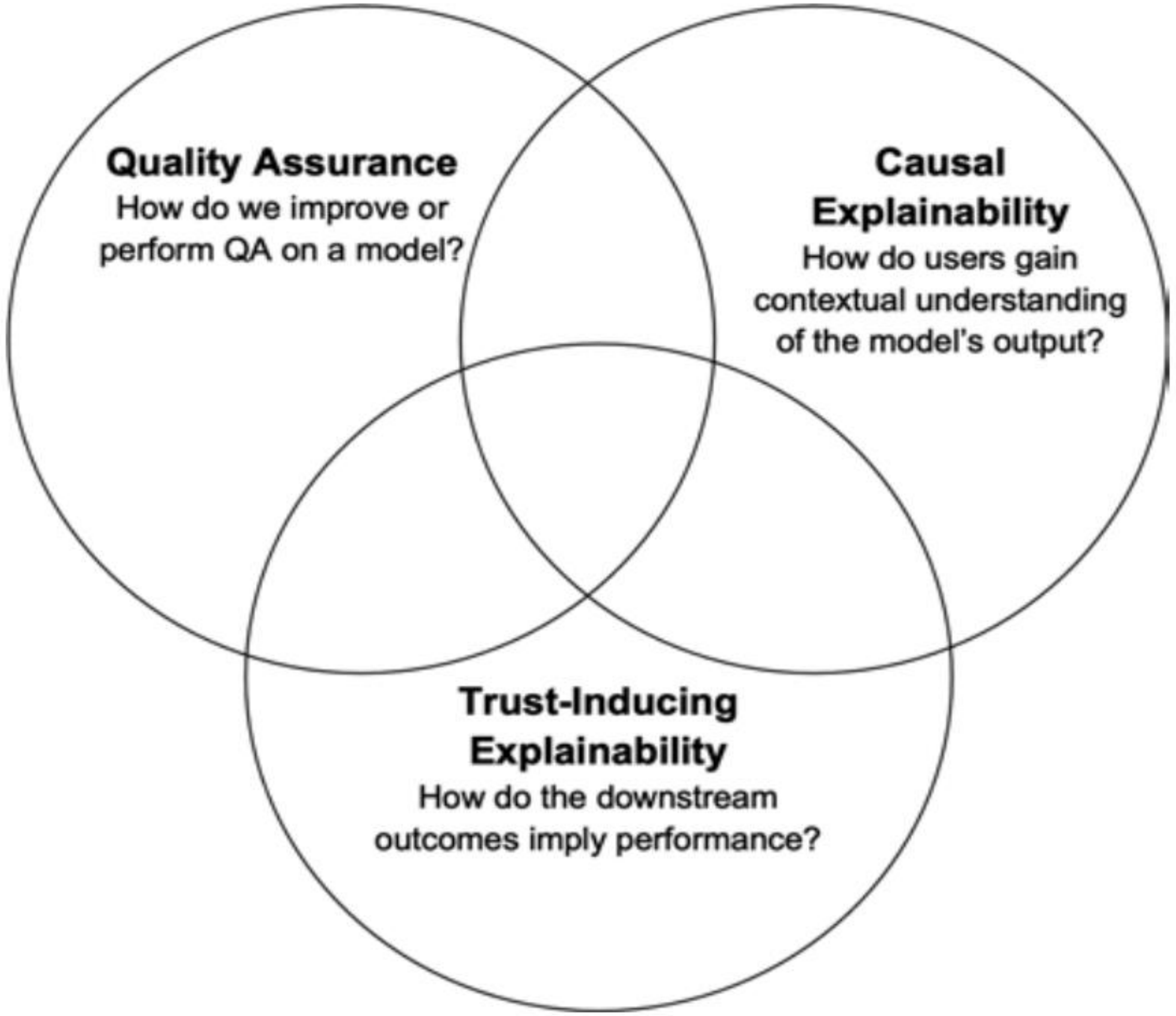
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

3

Again, trust in the AI advisor is important, but insight into the “how” and “why” of the AI prediction may be less relevant to the non-clinician layperson user since the AI prediction is only meant to be supplemental to a formal evaluation by a physician and does not directly drive care management.

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

Three Purposes of AI Explainability



A manifesto on explainability for AIM

Co-financed by the European Union
Connecting Europe Facility

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

Artificial Intelligence In Medicine 133 (2022) 102423

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

ELSEVIER

Research paper

A manifesto on explainability for artificial intelligence in medicine

Carlo Combi [a,*], Beatrice Amico [a], Riccardo Bellazzi [b], Andreas Holzinger [c], Jason H. Moore [d], Marinka Zitnik [e], John H. Holmes [f]

[a] *University of Verona, Verona, Italy*
[b] *University of Pavia, Pavia, Italy*
[c] *Medical University Graz, Graz, Austria*
[d] *Cedars-Sinai Medical Center, West Hollywood, CA, USA*
[e] *Harvard Medical School and Broad Institute of MIT & Harvard, MA, USA*
[f] *University of Pennsylvania Perelman School of Medicine Philadelphia, PA, USA*

A R T I C L E I N F O

Keywords:
Artificial intelligence
Explainability
Explainable artificial intelligence
Interpretability
Interpretable artificial intelligence

A B S T R A C T

The rapid increase of interest in, and use of, artificial intelligence (AI) in computer applications has raised a parallel concern about its ability (or lack thereof) to provide understandable, or *explainable*, output to users. This concern is especially legitimate in biomedical contexts, where patient safety is of paramount importance. This position paper brings together seven researchers working in the field with different roles and perspectives, to explore in depth the concept of explainable AI, or *XAI*, offering a functional definition and conceptual framework or model that can be used when considering XAI. This is followed by a series of desiderata for attaining explainability in AI, each of which touches upon a key domain in biomedicine.

The increasing use of AI/ML raises concerns and questions, such as:

- How does an AI algorithm work – what is it doing?
- Does an AI system work as well as an expert?
- Does an AI system do what a user would do, where she in the same situation?
- Why cannot the system tell a user how it arrived at a conclusion or made a decision?

Explainability is related to understanding, i.e. having a mental model of what we are observing.

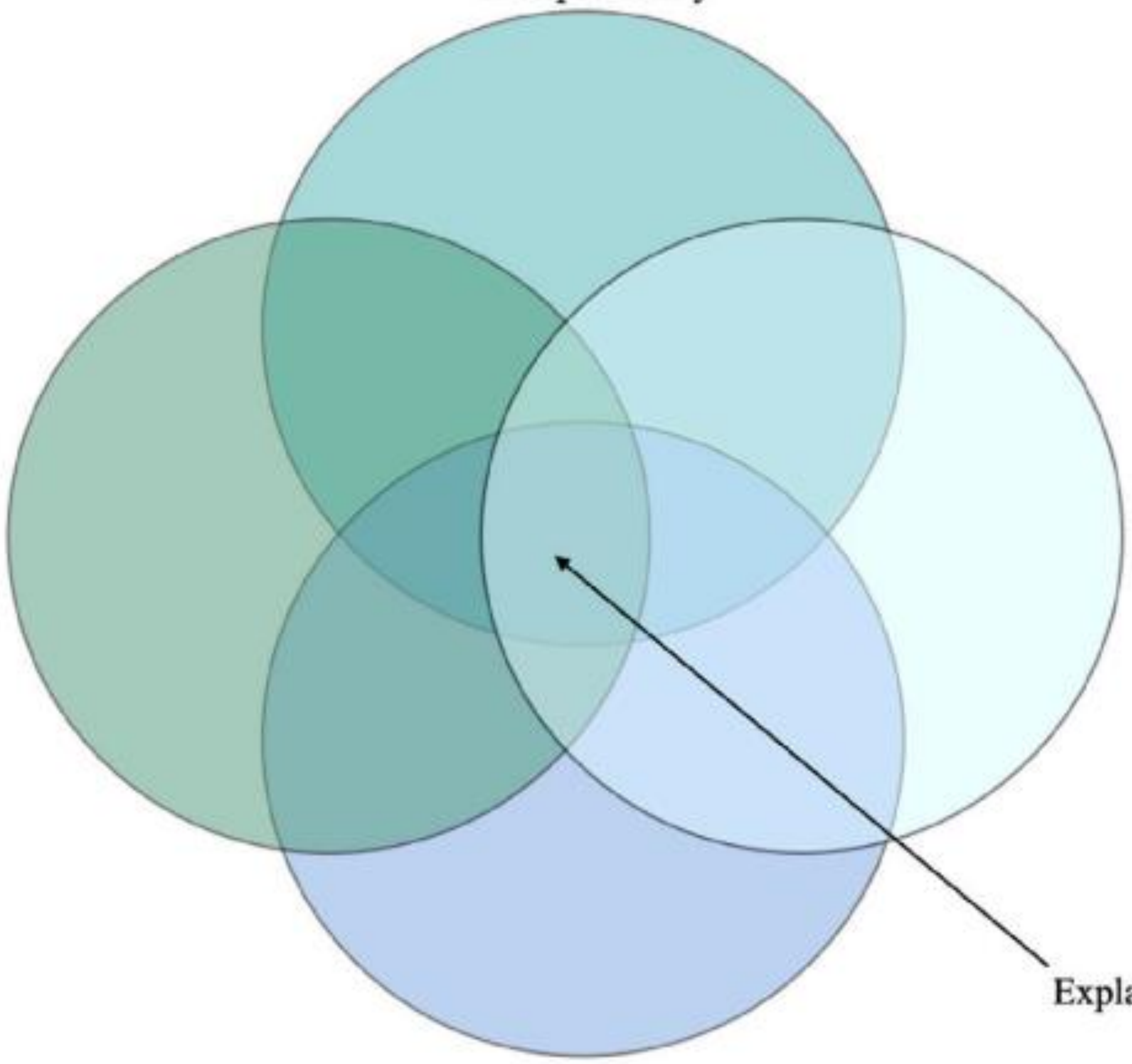
This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Explainability is an inherently multifaceted concept

- The **content** of explanation: *What is being explained?*
- The **stakeholders** of explanation: *Who needs explainability?*
- The **goal** for explanation: *Why is explainability required?*
- The **moment**, the **duration** and the **frequency** of explanation: *When, how long and how frequently.*
- The **modalities** of explanation: *How is explainability represented?*

Co-financed by the European Union
Connecting Europe Facility

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



Explainability as intersection of

- Usability
- Usefulness
- Interpretability, and
- Understandability

Towards a foundational definition of XAI in medicine

- **Interpretability**: the degree to which a user can intuit the cause of a decision and thus the ability of a user to predict a system's results.
- **Understandability**: the degree to which a user can ascertain how the system works, and leads directly to user confidence in the system's output.
- **Usability**: the ease with which a user can learn to operate, prepare inputs for, and interpret outputs of a system or component.
- **Usefulness**: asks the question "Will one use the system because it meets a user's needs?", i.e., the practical worth or applicability of a system. A system is unlikely to be useful if it is not usable.

Co-financed by the European Union
Connecting Europe Facility

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

Specific features of medicine and healthcare, which are central for XAI

- Distributed, heterogeneous decision-making tasks
- Knowledge-intensive domains

Co-financed by the European Union
Connecting Europe Facility

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

Distributed, heterogeneous decision-making tasks

- Call for usability and usefulness
- Usability and usefulness have to be evaluated according to different users and tasks
- They are not absolute concepts and need to be assessed “on the field”
- Usability supports the communication and shared decision-making among clinicians, general practitioners, and patients
  - *E.g., a web app supporting the mental health monitoring of home patients*

Co-financed by the European Union
Connecting Europe Facility

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

Knowledge-intensive domains and decision-intensive tasks

- Require to distinguish between interpretability and understandability
- Interpretability is related to the capability of predicting a system's result, even without being aware of the "internal" structure and functioning of the system
  - *E.g., a clinician has to be able to recognize how recorded vital signs of an ICU patient are related to the alarms triggered by an AI-based system*
- Understandability refers to the capability of being aware of how the system works
  - *A deep comprehension of system technicalities and behaviors would support a suitable elicitation of new medical knowledge*

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

What are the requirements for XAI? How can we evaluate the goodness of the provided explanation?

Proposition: There are tangible, instantiable, user-centered requirements that must be met in order to achieve an XAI system; more specifically, there is the need to measure, interpret, and understand usability vs. usefulness, and interpretability vs. understandability, and how those two relate to each other in the context of use and users, particularly in the context of AI in medicine.

Co-financed by the European Union
Connecting Europe Facility

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

If an AI system's output is understandable, is it automatically explainable?

Proposition: Understanding the output from an AI system is foundational to explainability, but it is only one requirement that has to be merged with usability, usefulness, and interpretability to compose explainability.

What is the role of domain understanding in achieving XAI in medical applications?

Proposition: XAI-based systems need to start from modeling the biomedical and clinical domain in order to obtain a true understanding of the context in which these systems will be used.

Co-financed by the European Union
Connecting Europe Facility

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

Can explainability draw us closer to wisdom?

Proposition: Explainability is a requirement to completing the data-information-knowledge-wisdom spectrum.

Can an AI system that is not explainable be trustworthy?

Proposition: XAI is an integral component of trustworthy AI systems.

Co-financed by the European Union
Connecting Europe Facility

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

Is XAI in medicine always required?

Proposition: Explanations are not always required in order for an AI model to be useful. Functional specifications obtained from deep analysis of the problem domain and users should determine when explainability and interpretability are required.

Co-financed by the European Union
Connecting Europe Facility

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

Proposed research directions

- Bridging the gap between symbolic (ante hoc) and sub-symbolic (black-box) approaches.
- Engineering explainability into intelligent systems.
- Evaluating and improving the effects of explainable components and approaches.
- Determining when explainability is needed.
- Investigating the design of user-centered and user-tailored explainability artifacts.

Co-financed by the European Union
Connecting Europe Facility

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

Summary

- ❑ The significance of explanation in AI
- ❑ Abduction and Explanatory Coherence
- ❑ Causality
- ❑ Revisiting explanations in symbolic AI
- ❑ Opening the ‘black box’ in connectionist AI
- ❑ A manifesto on explainability for AIM

Co-financed by the European Union
Connecting Europe Facility

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