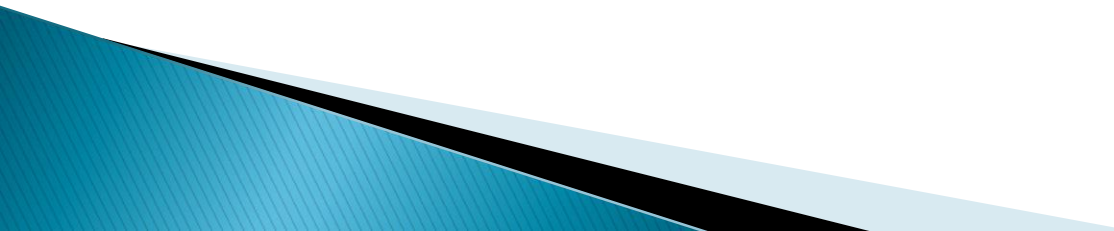


An Ethics Perspective on the Role of Explainability in the Context of AI Applications in Medicine

Elisabeth Hildt



Explainability

- ▶ Characteristic of an AI tool that allows users to understand and reconstruct why and how it came up with its decisions, predictions or recommendations
 - ▶ Term not well defined
 - ▶ Related: explicability, interpretability, transparency
- 

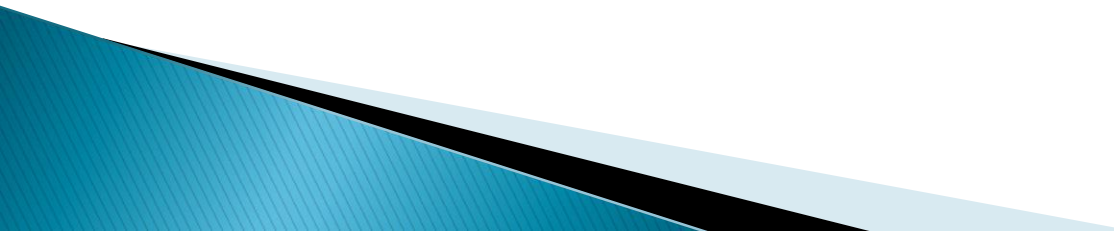
Explainability, human autonomy, agency, and responsibility

- ▶ Autonomous decision-making requires adequate knowledge of the relevant options
- ▶ Human agents are responsible for their autonomous decisions and for actions done voluntarily and intentionally
- ▶ With black box AI, humans do not know what the model output is based on
 - Strictly speaking, it is not humans who decide but the model
 - What humans can do is decide whether they want to rely on the black box model and its output



Lack of explainability risks issues related to individual autonomy, agency, accountability

Explainable AI in medicine

- ▶ Explainability is important for healthcare professionals, patients, informed consent
 - ▶ Explainability versus accuracy
 - ▶ How to balance explainability and accuracy?
 - ▶ Who needs explanation?
 - ▶ How much explanation is needed?
 - ▶ What counts as an explanation in medical contexts?
 - ▶ What is the role of causal explanation?
- 

Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls

- ▶ Health-related emergency calls (112) as part of the Emergency Medical Dispatch Center, City of Copenhagen
- ▶ ML tool developed that seeks to help medical dispatchers to identify cases of out-of-hospital cardiac arrest (OHCA)
- ▶ ML system, trained and tested by using archived audio files of emergency calls; it listens to the calls, produces text output
- ▶ Text output is fed to a classifier, gives alarm when it classifies a case as OHCA, no explanation provided
- ▶ Goal: provide caller with instructions for cardiopulmonary resuscitation (CPR); time factor!

Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls

- ▶ AI tool is faster at recognizing cardiac arrests than the dispatchers
- ▶ Tool has higher sensitivity but lower specificity than the dispatchers alone
 - Detects more OHCA than the dispatchers, higher false-positive rate
 - Problem with low specificity: risk to send out ambulances to false-positive cases, this may detract resources from other patients
- ▶ BUT no positive effect of the use of the AI tool, the patient survival rate could not be increased



Reasons?

ML tool to recognize cardiac arrests in emergency calls: role of (lack of) explainability

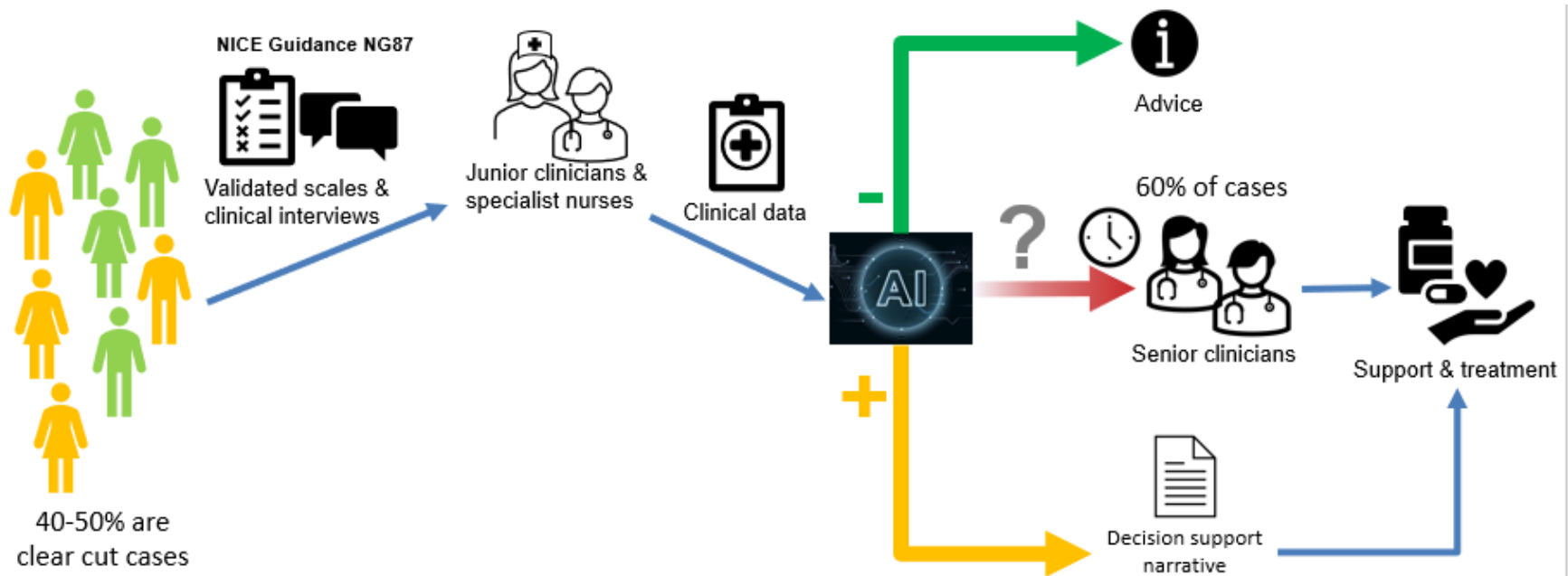
- ▶ Black box, the tool gives alarm but does not provide an explanation
- ▶ No way for dispatchers to find out whether the tool is right
- ▶ The tool does not function as a „second opinion“ to support dispatchers‘, it rather substitutes them
- ▶ Lack of trust?
- ▶ Do the dispatchers ignore the tool?

→ Explainability could increase acceptability and help that the tool is actually used

→ Explanations would allow dispatchers to make informed decisions

AI-supported diagnosis for adult attention deficit hyperactivity disorder (ADHD)

- ▶ ADHD undertreatment and underdiagnosis in the UK
 - Increased ADHD awareness, lack of senior specialists,
 - long waiting lists (1–3 years)
- ▶ AI tool developed to speed up the process (Tachmazidis et al. 2021)



AI-supported diagnosis for adult ADHD

- ▶ Tool: hybrid approach
 - Machine learning model, trained on data of past cases
 - Knowledge model

OUR USP: fusion of a **K**nowledge **M**odel and **M**achine **L**earning




AI-supported diagnosis for adult ADHD: Knowledge model

- ▶ Based on expert knowledge derived from interviews and NICE recommendations
- ▶ If-then rules
- ▶ Rules may conflict with each other
- ▶ Scores to indicate the relevance of assessments for ADHD diagnosis
- ▶ Indicators for overlapping conditions
- ▶ In case of ambiguous results → medical expert
- ▶ The hybrid model is more robust than the individual models as both models have to align for a yes or no answer

(Tachmazidis et al. 2021)

AI-supported diagnosis for adult ADHD: role of explainability (I)

- ▶ From yes – no prediction (ML model) to option „consult senior clinician“  clinical relevance!
- ▶ Intended function of the hybrid model: decision *support* for medical professionals with which they can interact
- ▶ Provides junior clinicians with information about the AI recommendations that they can pass on to patients
- ▶ Provides senior clinical specialists with information of why a patient is transferred to them
- ▶ Increases trust in the system?

AI-supported diagnosis for adult ADHD: role of explainability (II)

- ▶ Knowledge model relies on interviews with one expert
 - Generally accepted rules and scores? Shortcuts? Bias?
- ▶ Explainability may confer the impression that everything relevant has been covered, but:
 - Hybrid model relies primarily on standardized tests results
 - Limited role of expert experience, doctor-patient interaction
 - Model may miss aspects a senior expert would have considered
- ▶ Enough information for clinicians to trust the system?
- ▶ Patient acceptability?
- ▶ New patient category: patients with inconclusive results
- ▶ Deskilling of junior clinicians?
- ▶ Implication: less senior psychiatrists per clinic?

Algorithmic Advisory System for Moral Decision-Making in the Health-Care Sector

- ▶ Methods: fuzzy cognitive maps (FCM), conceptual analysis
- ▶ Principlism-based approach: beneficence, non-maleficence, autonomy (justice omitted)
- ▶ Developed algorithm to automate ethical decision-making
- ▶ → Predictions in favor or against a specified medical intervention
- ▶ Database: 69 clinical ethics committee cases
- ▶ Case parameters: variables that usually play a role in clinical ethics committee discussions
- ▶ The algorithm agreed with the ethicists in 92% of the cases in the training set and 75% of the cases in the test set (treatment yes or no)

(Meier et al., 2022)

<https://journals.library.iit.edu/index.php/CEPE2023/article/view/251>



Overview

- Start
- Patient
- Beneficence
- Non-Maleficence
- Autonomy
- Results

Beneficence

Please describe the expected *positive* effects of the intervention in question.

You have indicated that - untreated - the patient would probably have 5 years left to live.

How many years of life would he or she probably have left to live if the intervention was begun (or continued)?



You have rated the patient's current quality of life as "poor".

If the intervention has a potential positive effect on the patient's quality of life, how will it be affected?

- no positive effect
- marginally improved
- moderately improved
- markedly improved

For how many years can this positive effect on the quality of life be expected to persist?

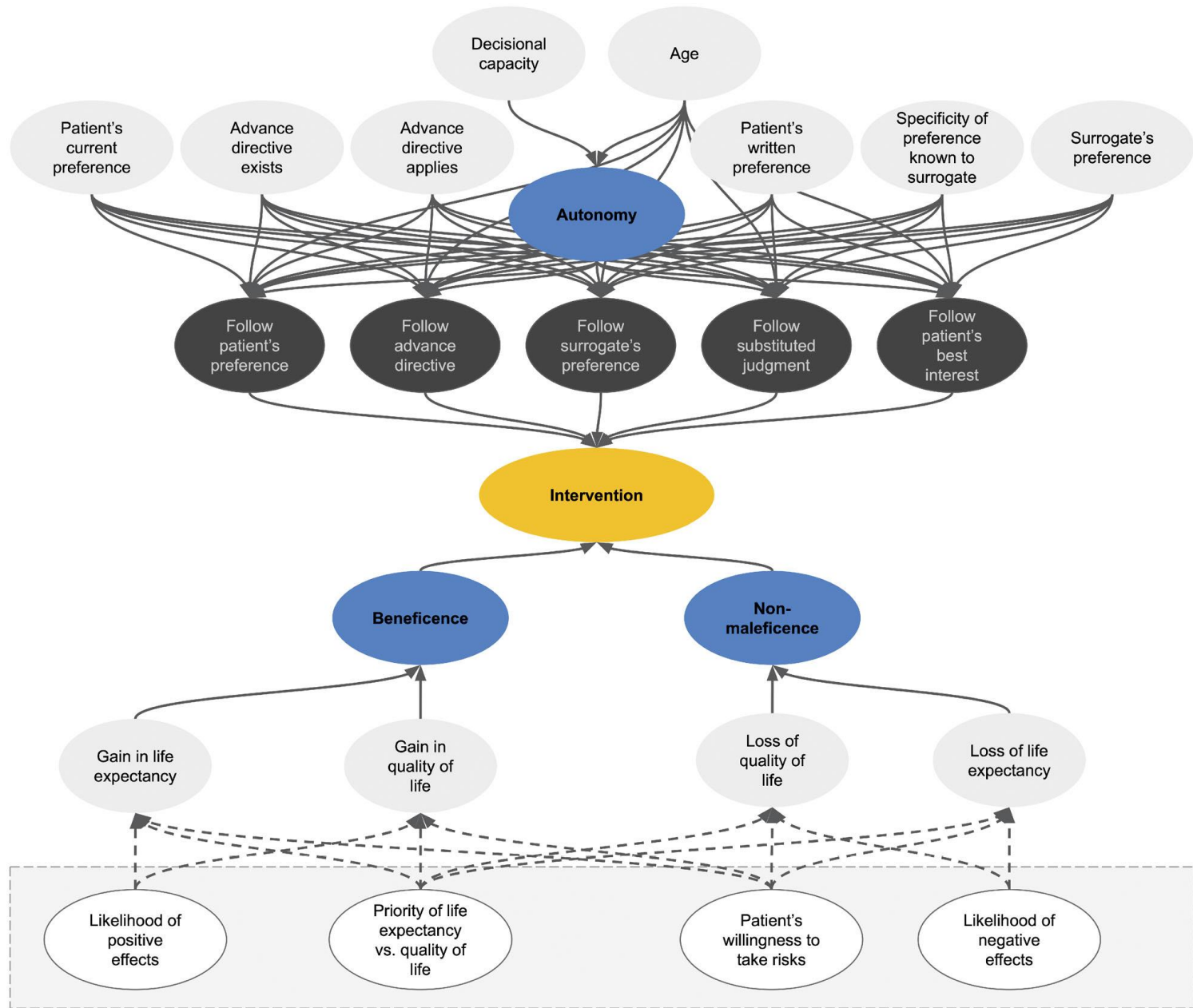


What is the likelihood for (the continuation of) the intervention to yield these positive effects?

- very low
- low
- fair
- high
- very high

Previous

Next



Visualization of the Fuzzy Cognitive Map (FCM)

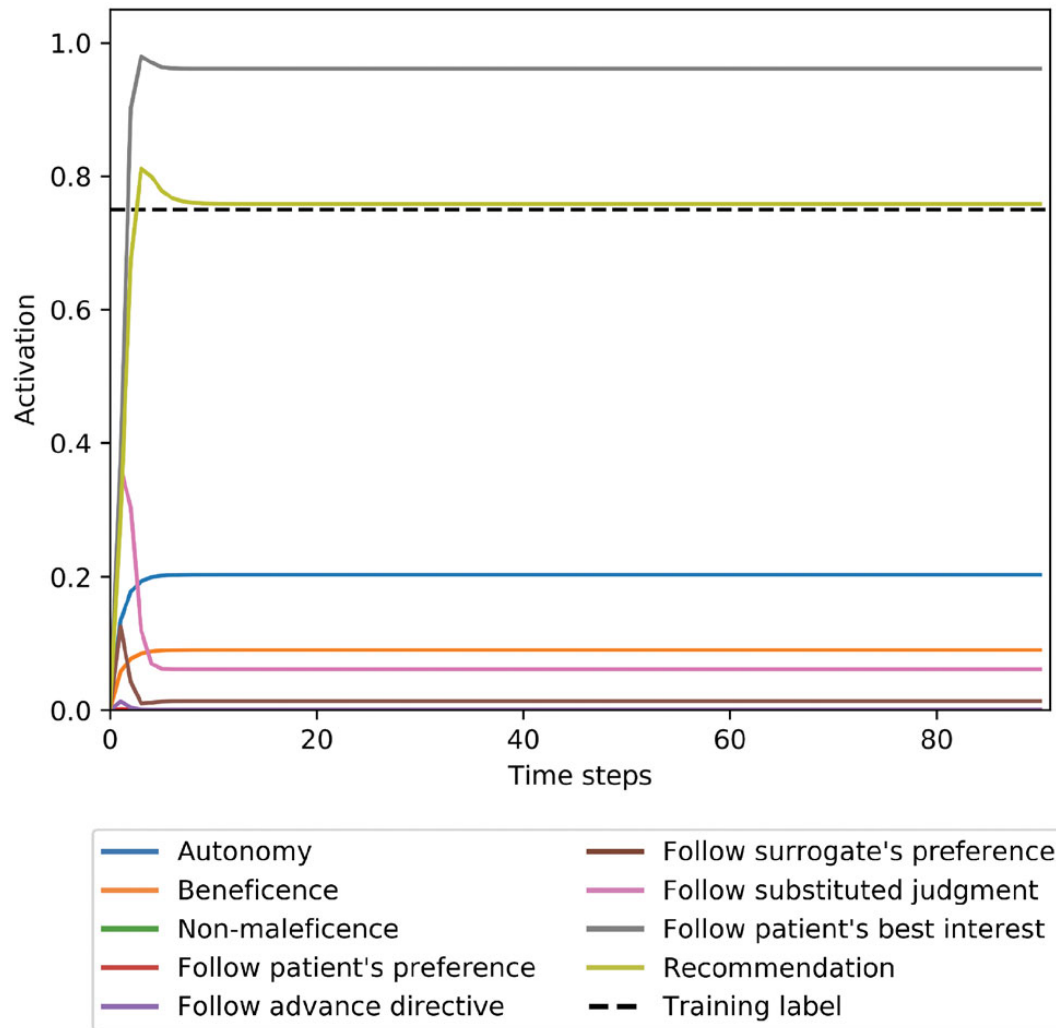


Figure 4. Example case 1: A 10-year-old is suffering from leukemia.

Seeing their child experience the strong side-effects that the therapy induces, the parents want all interventions halted. However, chemotherapy has proven to be highly effective and the child’s prognosis is very promising. METHAD’s analysis indicates that continuing the therapy would very likely be in the young patient’s best interest, which is in accordance with human ethicists’ judgment (denoted as “training label”).

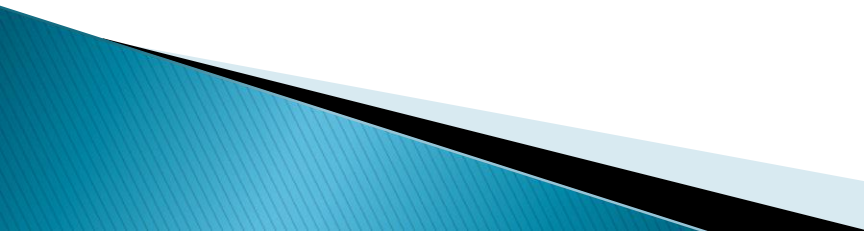
Algorithmic Advisory System for Moral Decision-Making: role of explainability (I)

- ▶ For AI tools involving moral decision-making, explainability is needed

→ not only medical facts are central, but also values, value hierarchies, individual preferences

- ▶ Users can check whether the tool is in line with their own values, value hierarchies, preferences and decisions
- ▶ The explanation provided may support clinical ethics committees in their decision-making process
- ▶ The explanation provided indicates whether the tool is in line with general medical ethics standards

Algorithmic Advisory System for Moral Decision-Making: role of explainability (II)

- ▶ Simplistic explanations
 - ▶ Focus on quantification tends to dismiss qualitative aspects
 - ▶ Explanation relies on a Western approach and one ethical theory
 - ▶ Value pluralism and individual patient preferences?
 - ▶ What is the authority of the tool?
 - ▶ What to do when users disagree with the tool?
 - ▶ Negative influence on patient autonomy?
 - ▶ „Standardization“ of medical ethics?
- 

Conclusion

- ▶ Explainability is of central relevance in the three clinical decision support systems (CDSS)
- ▶ Role of explainability is different in each example
- ▶ Advantages and downsides of explainability are different in each example

 Relevance of case-by-case analysis!

