

Where ethics meets the machine: ethics of AI and digitalisation in healthcare

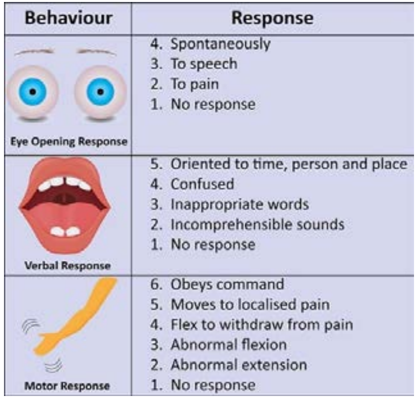
PD Dr. Nadia Primc
Institute for the History and Ethics of Medicine
Primc@uni-heidelberg.de

I. Three scenarios: the transparency and accountability of decisions
II. Medical ethics and its application onto AI use cases
III. AI in organ transplantation

Co-funded by the European Union

The case: ICU day 10

- Male patient, 74 years old
- Severe traumatic brain injury following a fall at home
- Unresponsive on ambulance arrival → admitted to ICU
- **Glasgow Coma Scale (GCS): 4/15** (deep coma, no meaningful response)
- No neurological improvement over 10 days
- Comorbidities: Type 2 diabetes, chronic kidney disease (stage 3), arterial hypertension



Behaviour	Response
Eye Opening Response	4. Spontaneously 3. To speech 2. To pain 1. No response
Verbal Response	5. Oriented to time, person and place 4. Confused 3. Inappropriate words 2. Incomprehensible sounds 1. No response
Motor Response	6. Obeys command 5. Moves to localised pain 4. Flex to withdraw from pain 3. Abnormal flexion 2. Abnormal extension 1. No response

https://wikism.org/w/images/2/27/Glasgow_coma_scale.jpg

The team's clinical assessment tends clearly toward a poor prognosis.

The question on the table: continue or withdraw life-sustaining treatment?

Scenario A — The scores we know

APACHE II *(Acute Physiology and Chronic Health Evaluation)*

- Developed in the US in 1980s; validated in large international ICU populations
- 12 physiological variables + age + chronic health status → score 0–71
- **Our patient's score: ~34 → estimated in-hospital mortality: ~85%**

SOFA Score *(Sequential Organ Failure Assessment)*

- Tracks dysfunction across 6 organ systems (respiratory, coagulation, liver, cardiovascular, neurological, renal)
- Recalculated regularly from routine data
- **Our patient: consistently high, no improvement over 10 days**

Both scores tell the same story. Their formulas are fixed, published, and clinically well understood.

Scenario B — The “black box”

- Hospital has implemented a newly approved commercial DL system for mortality prediction
- Trained on a large dataset of ICU records from multiple European hospitals
- Integrates directly into the EHR; updates continuously as new data arrive
- Validated against the hospital's own historical data, performance better/at least as good as classical scores

Predicted 90-day mortality on Day 10: 90%

- The physician cannot explain how this number was reached as the model architecture is not publicly disclosed; Variables used and their weights are not accessible
- Reasoning behind this specific prediction: unknown to the family, the physician, and the clinical team

Scenario C — The “explainable machine”

- Research-derived ML model designed for time-series data (processes data at 1-hour intervals)
- Trained on large multi-hospital ICU dataset; validated in studies and locally
- Inputs: vital signs, labs, medication records, full electronic health record
- Prediction improves over time: the longer the ICU stay, the more reliable the output

Predicted 90-day mortality on Day 10: 90%

- model uses an explainable AI technique
- For this patient, at this moment, the key drivers are shown: GCS score on Day 10, creatinine trajectory over the past week, age
- Displayed directly in the EHR

Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records

Hans-Christian Thorsen-Meyer, Annelaura B Nielsen, Anna P Nielsen, Benjamin Skov Kaas-Hansen, Palle Toft, Jens Schierbeck, Thomas Strøm, Piotr J Chmura, Marc Heimann, Lars Dybdahl, Lasse Spangsege, Patrick Hulsen, Kirstine Belling, Søren Brunak, Anders Perner

The comparison

	A. Classical Scores	B. Opaque ML	C. Explainable ML
Transparency	High — formula is public	Low — proprietary	Medium — outputs explained
Update frequency	Static / periodic	Continuous	Hourly
Explainability	Yes	No	Yes (XAI)
Family can interrogate?	Partly	No	Yes
Accuracy	good-very good	very good	very good

Which Would You Choose?

- If you were the physician entering that family meeting: Which tool would you be most comfortable using as a basis for the conversation?
- If you were a family member: Which tool would you be most comfortable with as part of the surrogate decision-making process?

The „transparency“ of clinical judgment

METIS

Cobler-Lichter et al. (2025): > 50% deaths after traumatic brain injury (TBI) follow the decision to withdraw life-sustaining therapy (WLST)

- ML model to analyze data from the American College of Surgeons’ National Trauma Databank (2017–2021) to predict WLST in TBI patients
- institutional withdrawal culture is a strong independent determinant of WLST, irrespective of clinical condition (age, GCS as further important factors)

Patients with identical clinical profiles could receive different decisions, depending on which hospital was treating them.

- self-fulfilling prophecy: institutional culture, not just clinical data, shapes the decision.

RESEARCH ARTICLE

Machine Learning Models to Predict Withdrawal of Life-Sustaining Therapy in Patients With Severe Traumatic Brain Injury

Michael Cobler-Lichter,[1] Jessica M. Delamater,[1] Fernanda J.P. Teixeira,[2] Ana M. Reyes,[1] Talia R. Arcieri,[1] Brian Manolovitz,[2] John A. McKeown,[2] Tulay Koru-Sengul,[3] Jonathan Jagid,[4] Joacir Graciolli Cordeiro,[4] Nina Massad,[2] Mohan Kottapally,[2] Amedeo Merenda,[2] Kristine O'Phelan,[2] Nicholas Namias,[1] and Ayham Alkhachroum[2]

Correspondence
Dr. Alkhachroum
axa2610@med.miami.edu

Neurology® 2025;105:e214249. doi:10.1212/WNL.0000000000214249

Rethinking “transparency” and the reshaping of clinical judgment

Classical scores are transparent, but they feed into a human decision process that is not fully transparent

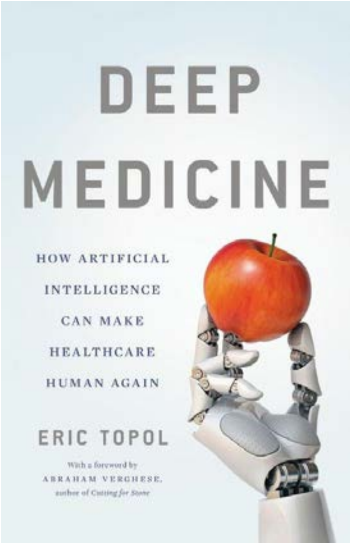
- Institutional culture, local norms, and implicit assumptions shape decisions in ways we cannot easily see

The question is not: Can we see inside the tool?

But: Can we see inside the decision and justify the decision? And what would it mean to make that process genuinely accountable?

AI does not simply replace human judgement, it shapes human judgment!

Beyond the ICU – a broader map



AI applications/use cases across the clinical pathway, different areas of care, professions, tasks, etc.:

- referral, scheduling of appointments
- management of waiting lists/triage
- diagnosis, clinical decision-making
- planning and support of therapy/surgery
- patient education and counseling
- data collection and analysis
- documentation
- biomedical research

.....

Recurrent ethical challenge:

→ tools that process large amount of data that the clinician cannot fully process

→ produces an output the clinician may not fully understand

→ shapes a decision with significant consequences for the patient.

How to make the use of these tools accountable?

I. Three scenarios: the transparency and accountability of decisions

II. Medical ethics and and its application onto AI use cases

III. AI in organ transplantation

Ethical guidelines and frameworks for AI

United Nations Educational, Scientific and Cultural Organisation (UNESCO)	AI Ethics Recommendation (2021)
World Health Organisation (WHO)	Guidance on Ethics and Governance of Artificial Intelligence for Health (2021)
European Union (EU)	Ethics Guidelines for Trustworthy AI (2019)
European Union (EU)	EU AI Act (Regulation (EU) 2024/1689)
Organisation for Economic Co-operation and Development (OECD)	AI Principles (published in 2019, updated in 2024)
Institute of Electrical and Electronics Engineers (IEEE)	IEEE Ethically Aligned Design (2019)
Microsoft	Responsible AI Standard (published internally in 2019, then openly in 2022)
Google	AI Principles (2018; updated 2025)

transparency and explainability, accountability, fairness and non-discrimination, privacy and data protection, safety and reliability, and human oversight, as often cited principles

→ not specific to medicine/healthcare

The „Georgetown mantra“

Beauchamp & Childress (1979). Principles of biomedical ethics.

(1) autonomy = to respect the patient’s right to decide

(2) Beneficence = to act in the patient’s best interest

(3) Non-maleficence = to avoid unnecessary harm

(4) Justice = to distribute benefits and burdens fairly

These principles do not always point in the same direction.

Medical ethics is the practice of reasoning carefully about how to balance them.

Autonomy in the ICU: surrogate decision-making

Scenario	**Can the family interrogate the reasoning?**
A Classical scores	Yes — variable by variable, with clinical support
B Opaque ML	No — the model is proprietary; no one in the room can explain it
C Explainable ML	Partly — key drivers are visible, but not full reasoning

The patient cannot speak for himself. The family must decide ***what he would have wanted****, not what they want for him. Surrogate decision making demands comprehensible disclosure, not merely disclosure.*

Transparency is not binary. It comes in degrees. For a surrogate end-of-life decision, the standard required is very high.

Beneficence and non-maleficence: competence as part of the duty of care

Beneficence requires more than good intentions: it requires the ability to assess whether tools are valid and appropriate for this patient.

- **Scenario A**: The physician knows the formula, its variables, its limitations, and its derivation. Her competence is intact.
- **Scenario B**: She cannot scrutinise the model. She cannot know if it performs as claimed.
- **Scenario C**: Variable contributions are visible, but drift and population fit remain unknown.

Using a tool she cannot assess is not negligence, but structural opacity and a constraint on her ability to fulfil her duty of beneficence. Non-maleficence applies symmetrically: continuing futile treatment may itself be an unjustified harm.

Justice: Whose history is in the training data?


Science
Current Issue First release papers Archive About Submit manuscript
HOME > SCIENCE > VOL. 366, NO. 6464 > DISSECTING RACIAL BIAS IN AN ALGORITHM USED TO MANAGE THE HEALTH OF POPULATIONS
RESEARCH ARTICLE
Dissecting racial bias in an algorithm used to manage the health of populations
ZIAD OBERMEYER, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN Authors Info & Affiliations
SCIENCE • 25 Oct 2019 • Vol 366, Issue 6464 • pp. 447-453 • DOI: 10.1126/science.aax2342

The principle: Fair distribution of benefits and burdens in healthcare. Care must not be influenced by clinically irrelevant factors (e.g. race, gender, or socioeconomic status)

In our clinical case: Justice is less central than the other three principles, as the decision is not a question of distributing scarce resources.

But: AI and the risk of reproducing injustice

e.g. Obermeyer et al. (Science, 2019)

- algorithm systematically underestimated the health needs of Black patient´s: healthcare costs were used as a proxy for health need
- Result: structural inequity in the data amplified inequity in the output

AI systems learn from historical data and may reproduce the inequalities that history contains.→ focus of Lecture 3.

What the 4 principles reveal: the limits of generic AI ethics

Generic AI ethics: *Is the tool transparent/fair? Is there a human in the loop?*
→ Questions are formulated at a level of abstraction that is **context-independent**.

„Transparency“ means something different, when it serves different ethical obligations.

Principle	What transparency actually requires
Autonomy	Comprehensible disclosure enabling genuine surrogate decision-making
Beneficence	Sufficient understanding of a tool's basis/limitations for professional judgment
Justice	Ability to detect whether training data encodes/amplifies structural inequity

Medical AI ethics requires medical ethics — not just AI ethics.

I. Three scenarios: the transparency and accountability of decisions

II. Medical ethics and and its application onto AI use cases

III. AI in organ transplantation

The ethical landscape shifts across the transplantation process

Patient-centred ethics (Stages 1–3)
Individual rights · Autonomy · Informed consent · Donor & family obligations

Allocation-centred ethics (Stage 4)
Distributive justice · Urgency · Utility · Fairness · Public trust

STAGE 1 — Referral to Transplant Centre
Potential recipient
- Prediction of referral benefit vs. continued medical management
- Risk stratification for urgent referral
- Patient navigation & information support (LLM)

STAGE 2 — Recipient Evaluation & Listing Decisions
Recipient / waitlisted patient
- Shared decision support (dialysis vs. transplant)
- Waitlist mortality prediction beyond MELD / MELD 3.0
- AI-based clinical decision support for listing (AI-CDS)

STAGE 3 — Donor Identification & Evaluation
Post-mortem donor | Living donor | Donor family
- ICU identification of potential post-mortem donors
- Living donor suitability & complication risk evaluation
- Graft quality & steatosis assessment from imaging

STAGE 4 — Organ Allocation & Donor–Recipient Matching
All waitlisted patients | Donor / donor family | Society / public
- Post-transplant survival prediction (utility)
- Waitlist mortality & urgency scoring
- Donor–recipient matching
- Behavioural prediction (adherence, substance use)

Continued →

Patient-centred ethics (Stages 5–7)
Professional responsibility · Adherence & graft survival · Long-term follow-up · Re-listing equity

STAGE 5 — Surgery & Perioperative Care
Recipient | Living donor
- Biliary anatomy mapping & virtual hepatectomy planning
- Graft weight prediction from imaging
- Organ preservation & transport optimisation

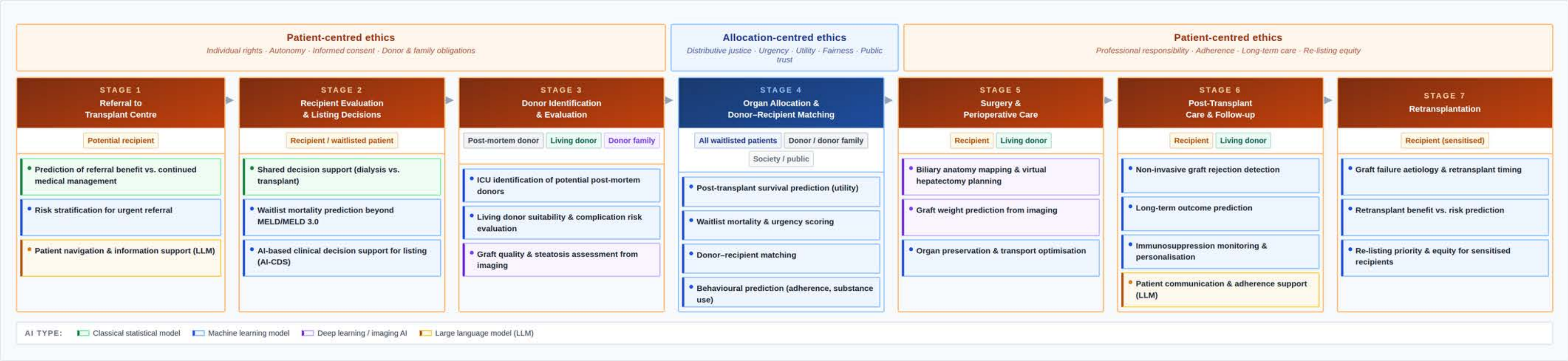
STAGE 6 — Post-Transplant Care & Follow-up
Recipient | Living donor
- Non-invasive graft rejection detection
- Long-term outcome prediction
- Immunosuppression monitoring & personalisation
- Patient communication & adherence support (LLM)

STAGE 7 — Retransplantation
Recipient (sensitised)
- Graft failure aetiology & retransplant timing
- Retransplant benefit vs. risk prediction
- Re-listing priority & equity for sensitised recipients

AI TYPE: Classical statistical model · Machine learning model · Deep learning / imaging AI · Large language model (LLM)

Abbreviations: *AI = artificial intelligence; LLM = large language model; MELD = Model for End-Stage Liver Disease; ICU = intensive care unit; AI-CDS = AI-based clinical decision support. All machine learning and deep learning applications are at research or pilot stage unless otherwise stated.*

AI in organ transplantation: same technology, different ethical contexts



In Brief

iChoose Kidney for Treatment Options: Updated Models for Shared Decision Aid

Jennifer C. Gander, PhD,[1] Mohua Basu, MPH,[2] Laura McPherson, MPH,[2] Michael D. Garber, MPH,[3] Stephen O. Pastan, MD,[4] Amita Manatunga, PhD,[5] Kimberly Jacob Arriola, PhD,[6] and Rachel E. Patzer, PhD[2,3]

Pruinelli *et al. BMC Medical Informatics and Decision Making* (2025) 25:98
https://doi.org/10.1186/s12911-025-02890-3

BMC Medical Informatics and Decision Making

SYSTEMATIC REVIEW Open Access

Transforming liver transplant allocation with artificial intelligence and machine learning: a systematic review

Lisiane Pruinelli[1,2*], Kiruthika Balakrishnan[1], Sisi Ma[3,4], Zhigang Li[5], Anji Wall[6], Jennifer C. Lai[7], Jesse D. Schold[8], Timothy Pruett[9] and Gyorgy Simon[3]

Summary: working from the inside out

The ethical significance of the ICU case depends on whether the tools used enable:

- Genuine informed surrogate decision-making (autonomy)
- A competent professional assessment of validity and limits (beneficence)
- Visible and attributable harms (non-maleficence)
- Freedom from the contamination of historical inequity (justice)

The core claim of this lecture:

- Generic AI ethics frameworks, formulated at a level of abstraction that is context-independent, cannot capture contextually grounded requirements that shift as the clinical context changes.
- A framework adequate to this complexity must be built from the inside out: from the specific ethical obligations that apply at each stage of the clinical process — not from generic principles applied uniformly from above.

Brief (!) Survey

Next lecture: Virtual dissections in human anatomy teachings

09/04/2026 at 16:00

Prof. Dr. Paola ALBERTI