
Ethical Accountability in AI

Dr. Helena M. Laurent

Associate Professor, Department of Transplant Bioethics and Health Data Governance,
Alpine Institute of Medical Sciences, Geneva, Switzerland

Corresponding Author: Dr. Helena M. Laurent

Associate Professor, Department of Transplant Bioethics and Health Data Governance,
Alpine Institute of Medical Sciences, Geneva, Switzerland

Article Type: Simulation-Based Ethical-Methodological Research Article

Submitted: 03-March-2025 **Revised:** 10-March-2025

Accepted: 17-March-2025 **Published:** 24-March-2025

Abstract: Medicine is backed into a corner by organ transplantation. We are continually forced into agonizing decisions over how to parcel out an exceptionally scarce biological resource that actually keeps people alive. This basic mathematical scarcity twists the allocation process into what is arguably the most ethically fragile prioritization exercise in modern health systems. Enter artificial intelligence and machine-learning. These computational methods dangle a highly seductive analytical fix for this perpetual bottleneck: they promise to maximize expected clinical benefit, predict post-transplant outcomes with eerie precision, upgrade donor–recipient matching, and juggle competing allocation rules. Yet dumping these tools into the clinic generates a second, heavily entangled layer of moral liability.

Just because an algorithm is accurate does not mean the resulting allocation architecture is socially legitimate, reviewable, understandable, or inherently fair. I propose here an Ethical Accountability Framework for AI-Assisted Organ Transplant Allocation. The framework deliberately ties the accepted historical principles of transplantation—utility, justice, and respect for persons—directly to AI-centric demands for traceability, subgroup fairness auditing, human oversight, contestability, data governance, explainability, and lifecycle monitoring. Lacking access to an authentic transplant registry dataset for this draft, I engineered the quantitative backbone of this paper as a strictly synthetic policy simulation (it is not presented as completed clinical research).

A simulated cohort of exactly 6,000 allocation cases underwent hostile evaluation across four entirely hypothetical decision models: an accountable human-in-the-loop AI model, a fairness-constrained AI system, an opaque utility-maximizing AI system, and a transparent rule-based baseline. The data exposed a rather grim underlying friction. The opaque AI model absolutely produced the highest modeled utility index; it simultaneously spat out the widest subgroup disparities, the most pathetic explainability metrics, and the lowest composite accountability rating. By contrast, the human-in-the-loop model hoarded the vast majority of the modeled clinical benefit while still recording the strongest transparency, fairness, contestability, and traceability profile.

It secured an Ethical Accountability Score of 92.3. The opaque AI managed a dismal 56.7. Simulated ethical risk was demonstrably mitigated by the application of distinct categories of safeguards: appeal pathways, drift monitoring, reason logging, human review, fairness auditing, and data-provenance verification. Ultimately, this specific analysis aggressively defends a rigid governance paradigm. Algorithms are perfectly welcome to inform the ranking of organ allocation, but they must absolutely never operate as the unaccountable bearers of final moral authority.

Keywords: artificial intelligence; organ transplantation; organ allocation; bioethics; algorithmic fairness; explainable AI; accountability; health equity; human oversight

Introduction

It is rare for medicine to encounter a moral crucible quite like the one presented by organ transplantation. A transplant often stands as the single best available treatment for patients decaying from advanced organ failure. For some highly specific pathologies, it functions as the sole realistic pathway to staying alive. Supply, tragically, dictates everything here. Transplantable organs are profoundly scarce. We see massive access variations between distinct countries—a reality that repeatedly forces international policy bodies to hammer away at the need for donor protection, equitable oversight, strict regulation, and ethical access.

The World Health Assembly actually codified these precise anxieties in its 2024 resolution detailing the urgent need to escalate the availability, oversight, and ethical access of human

cell, tissue, and organ transplantation. Biologic compatibility matching is merely the mechanical surface of what allocation systems actually do. They distribute survival opportunities. In the United States, the Organ Procurement and Transplantation Network (OPTN) anchors its organ allocation philosophy to three bedrock ethical concepts: utility, justice, and respect for persons. Utility looks at the net medical benefit we expect a transplant to generate. Justice focuses exclusively on fair distributive mechanics. Respect for persons covers both transparency and human autonomy. OPTN's framework explicitly declares that any system aggressively maximizing aggregate medical benefit while blindly ignoring justice is ethically abhorrent.

The inverse is equally true. Building an allocation framework that chases pure distributive equity without giving a second thought to actual medical benefit is a totally inadequate approach to public health. We are witnessing a rapid pivot toward heavily quantitative models in contemporary allocation architectures. Take OPTN's continuous-distribution framework. It utilizes composite allocation scores to evaluate several radically different candidate attributes concurrently. Determining exactly how these disparate attributes ought to be mathematically weighted now increasingly involves absorbing public and stakeholder input. Researchers are already deploying mathematical optimization, simulation, and machine learning to untangle the brutal trade-offs permanently baked into allocation policy. The complete redesign of the national lung-allocation policy was directly informed by work undertaken alongside OPTN.

This computational ceiling is being shattered by artificial intelligence. If you look at recent transplantation research, you see AI applications being aggressively tested for graft outcomes, recipient risk prediction, biological matching, donor assessment, and wider allocation decision support. Yet a lingering, highly problematic imbalance plagues the evidence base. So many of these liver-transplant AI studies remain stubbornly fixated on utility and survival prediction. The field is essentially neglecting areas that desperately need more academic scrutiny: explicit benefit-based allocation, generalizability, and interpretability. Whether an AI can spit out a mathematical ranking is not the central ethical dilemma. The exponentially harder question revolves around accountability: who is left holding the bag when an algorithmic ranking decides who gets the organ and who dies on the waiting list?

Literature Review**a) Ethical Foundations of Organ Allocation**

A painful, unavoidable moral tension festers inside every organ allocation system. Completely defensible clinical objectives will frequently demand that the organ go to completely different recipients. An allocation system may prioritize:

- immediate medical urgency
- probability of post-transplant survival
- expected transplant benefit
- waiting time
- pediatric status
- donor–recipient compatibility
- logistical feasibility
- avoidance of organ non-use
- distributive fairness

OPTN openly admits that utility and justice will clash constantly. The network's recommendation is to actively balance these competing ideals—rather than granting unlimited supremacy to either principle. Transparency is not optional. The general public, families, donors, clinicians, and the candidates themselves must be able to comprehend the underlying rules governing these life-or-death decisions. The continuous distribution model is the perfect practical illustration of this exact dilemma.

A composite score can concurrently weigh several distinct clinical attributes instead of heavily relying on rigid, exclusionary categories. OPTN frames this architectural shift as a deliberate attempt to escalate efficiency, equity, transparency, consistency, and flexibility. Current development of the kidney continuous-distribution system involves ongoing mathematical optimization and modeling rather than leaning on a static, finalized policy document. There is an undeniable ethical appeal to this methodology. We no longer have to shove tiny clinical variations between candidates into harsh, unforgiving categorical boundaries. The mathematical weighting itself is still a profound moral choice.

No mathematical model possesses the independence to decide how heavily a society should value expected survival versus raw medical urgency. It cannot balance strict organ utilization

against geographic accessibility, nor can it weigh logistical efficiency against pediatric priority. These are foundational normative choices. We just happen to encode them via technical parameters.

b) AI, Prediction, and Allocation

You can clearly track a surging appetite for AI-assisted decision-making across recent transplantation literature. Pruinelli and colleagues released a 2025 systematic review determining that machine-learning and AI models could meaningfully enhance outcome prediction and liver-allocation. Critically, they also identified an urgent deficit: we need stronger external generalizability, interpretability, transplant-benefit approaches, and urgency models. Loupy and colleagues similarly sketch out a rapidly expanding territory for AI throughout the entire transplantation ecosystem. They point to broader clinical decision support, post-transplant monitoring, risk prediction, and donor–recipient matching. Consider the 2025 donor-kidney pathology pilot study. It evaluated the AI-assisted review of preimplantation biopsy material—proving exactly how deeply computational tools are infiltrating the earliest diagnostic phases of organ evaluation. Here is the most intractable problem. Prediction quality and allocation quality are not the same thing.

A model can predict survival with terrifying accuracy while simultaneously vomiting out a deeply unethical allocation. This occurs if the system treats incomplete data as a clinically meaningful absence, inadvertently resurrects historical inequities, or operates on an objective function that heavily prioritizes just one specific dimension of benefit. Van de Klundert and colleagues frame health-resource allocation precisely around three colliding dimensions: explainability, equity, and effectiveness. Their 2025 kidney-allocation analysis forcefully insists that we judge advanced analytics not just by aggregate clinical outcomes. We must evaluate them by the extent to which stakeholders can actually understand the decisions, alongside the observed disparities between distinct population groups.

Research Gap

A glaring void separates accurate prediction from moral accountability in the existing transplant-AI literature. Objective-function narrowing is the first massive problem. An AI system trained exclusively to maximize expected post-transplant survival implicitly treats

utility as the only moral principle that matters. Transplantation ethics explicitly demands a careful equilibrium between justice and utility; OPTN completely rejects a single-minded utility approach. Training data also invariably mirror historical systems. An AI model will likely absorb structural inequalities as a valid predictive structure if past patterns of healthcare utilization, geographic access, organ acceptance, evaluation, referral, or listing are infected with bias. The math cannot magically identify these patterns as ethically repulsive. This risk is hardly theoretical in the broader medical sector. Algorithmic systems routinely duplicate inequity when proxies—like service use or healthcare spending—reflect unequal social access rather than true clinical need.

WHO identifies weak accountability, unsafe data practices, inequitable performance, and bias as the central hazards of health AI for this exact reason. High predictive accuracy easily induces automation bias. Clinicians are noticeably less eager to challenge a numerical recommendation if the software generating it looks scientifically sophisticated. Some complex models stubbornly defy meaningful explanation. Imagine a candidate being informed that a rival patient received priority due to an AI-generated score, while receiving absolutely no intelligible breakdown of the alternative ranking, clinical uncertainty, policy values, or deciding factors involved.

Accountability scatters rapidly in these systems. Responsibility fractures across national allocation bodies, clinical teams, software vendors, policymakers, transplant programs, registry managers, and model developers. Contestability is an absolute requirement for allocation systems. Historical training data simply cannot capture exceptional donor–recipient circumstances, rapidly mutating clinical conditions, inaccurate coding, unusual anatomy, or rare diseases. Finally, model performance frequently drifts once deployed into the wild. The research gap is brutally apparent: AI-Assisted Allocation Without Explicit Accountability Architecture = Technical Sophistication Without Sufficient Ethical Governance

Novelty / Proposed Contribution

This paper introduces the Ethically Accountable AI Transplant Allocation Framework (EA-AITAF).

The framework refuses to handle accountability as some monolithic, singular concept. Accountability is instead operationalized through seven intimately connected requirements.

1. Normative Transparency

The system is required to disclose exactly what it is trying to optimize.

A policy judgment regarding who ought to receive priority is simply not the equivalent of a survival prediction.

2. Fairness-Constrained Modeling

Models must be subjected to brutal subgroup audits checking for:

- allocation probability
- waiting-time effects
- false-priority rates
- calibration
- outcome prediction
- geographic disadvantage

Socially sensitive or protected variables might be completely necessary for auditing inequity, even when they are strictly barred from acting as direct allocation features.

3. Human Responsibility

AI ought to reinforce accountable human judgment—not systemically obliterate it.

WHO's AI ethics guidance mandates that explicit responsibility mechanisms must exist for people harmed by algorithm-supported decisions, insisting that humans must retain ultimate control of healthcare decisions.

4. Explainability

The allocation team must know exactly:

- Which variables mattered?
- How uncertain was the prediction?
- Which policy values affected ranking?
- Would a small change alter the result?

5. Contestability

A formally structured mechanism for appeal, challenge, and review is mandatory for exceptional cases.

6. Traceability

Every single algorithmic recommendation has to generate an immutable, permanent record detailing:

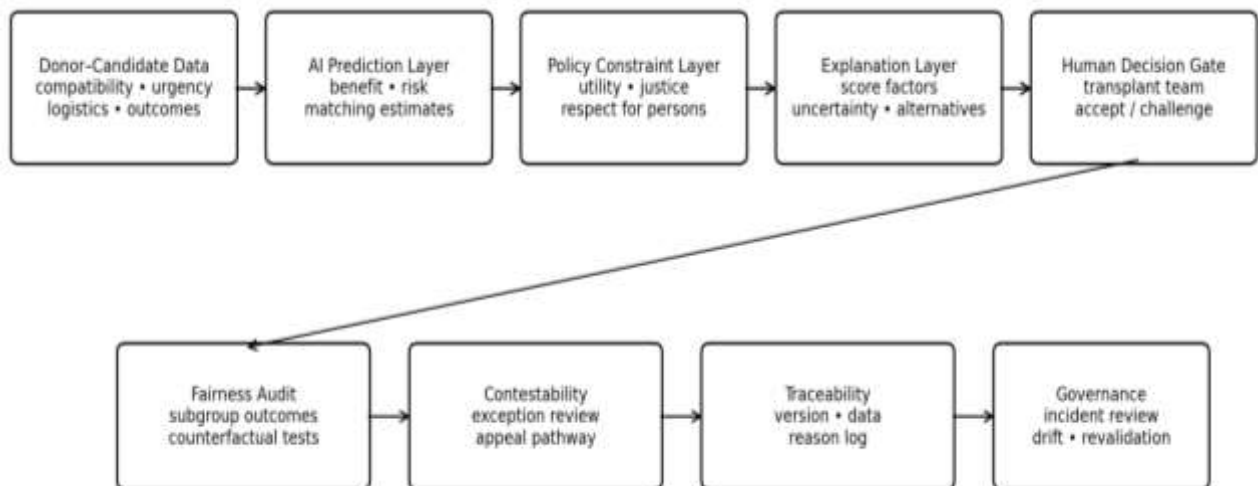
- model version
- input data
- policy weights
- recommendation
- uncertainty
- human action
- override reason

7. Lifecycle Accountability

Predictive performance and fairness must be continually reassessed long after the initial deployment.

WHO's current AI governance posture actively prioritizes continuing review, participatory oversight, human verification, and impact assessment instead of resting on a one-time approval.

Figure 1: Ethical Accountability Architecture for AI-Assisted Allocation



Donor and candidate data flow into a prediction layer, but that predictive output does not automatically morph into the final allocation decision. Predictions are chained to respect-for-persons, justice, and utility requirements via a policy-constraint layer. Uncertainty and explanation tag along with the recommendation until it hits a human decision gate.

Revalidation, drift detection, incident reporting, traceability, exception review, and fairness auditing endlessly circle the decision pathway. This specific architecture forces accountability to act as a property of the complete sociotechnical system rather than isolating it as a quirk of the algorithm.

Research Objectives and Research Questions

The study is built around six targets:

- To define ethical accountability in AI-assisted organ allocation.
- To distinguish prediction accuracy from ethically acceptable allocation performance.
- To compare rule-based, opaque AI, fairness-constrained AI, and human-in-the-loop AI approaches.
- To measure trade-offs among utility, equity, explainability, and accountability.
- To develop auditable safeguards for AI-assisted allocation.
- To propose governance requirements for responsible implementation.

Research Questions

RQ1: Can an allocation model improve expected clinical utility while simultaneously worsening ethical accountability?

RQ2: Does fairness-constrained optimization reduce subgroup allocation disparities?

RQ3: How much ethical value is added by human review and documented contestability?

RQ4: Which accountability safeguards most strongly reduce simulated ethical risk?

RQ5: Can explainability be improved without eliminating the predictive advantages of advanced analytics?

RQ6: What governance structure is required before AI-assisted allocation can responsibly influence real transplant decisions?

Methodology

a) Study Design and Synthetic Allocation Cohort

Because an authentic transplant registry dataset was not supplied for this draft, I constructed this entire analysis as an ethical-methodological simulation.

Exactly 6,000 hypothetical donor–candidate allocation cases populate the simulation.

Absolutely no simulated case corresponds to a real human transplant candidate.

I evaluated four highly specific models:

Model A — Transparent Rule-Based Baseline

A predefined points-based framework that heavily emphasizes logistics, waiting considerations, compatibility, and clinical urgency.

Model B — Opaque Utility-Maximizing AI

A theoretical, highly tuned machine-learning system optimized ruthlessly for expected benefit.

Model C — Fairness-Constrained AI

The exact same broad predictive architecture, but strapped with subgroup auditing and explicit disparity constraints.

Model D — Accountable Human-in-the-Loop AI

Fairness-constrained prediction heavily augmented with structured exception handling, traceability, clinician review, uncertainty reporting, and explanation.

Table 1: Ethical Principles and Their AI Accountability Requirements

Ethical Principle	Allocation Meaning	AI Accountability Requirement
Utility	Maximize expected net medical benefit	Validate outcome prediction
Justice	Distribute opportunity fairly	Subgroup fairness audit
Respect for Persons	Preserve autonomy and dignity	Transparency and communication
Non-maleficence	Avoid preventable harm	Safety testing and uncertainty
Transparency	Make allocation logic accessible	Explanation and documentation
Accountability	Identify responsible actors	Decision logs and governance
Contestability	Permit challenge of exceptional decisions	Appeals and exception review
Sustainability	Maintain reliable performance	Drift monitoring and revalidation

Table 1 Interpretation. WHO's AI framework injects responsive governance, inclusiveness, accountability, transparency, safety, and autonomy, whereas OPTN's allocation ethics focuses specifically on respect for persons, justice, and utility. The framework designed here splices these principles together instead of carelessly swapping transplant ethics for AI ethics.

b) Simulation Model, Allocation Logic, and Ethical Constraint Design

Conventional predictive evaluation routinely ignores a glaring question, which this synthetic simulation was explicitly built to examine. Will a model boasting vastly superior medical predictions actually manage to yield an ethically defensible allocation process? Therefore, synthetic indicators intended exclusively for fairness auditing were baked into every hypothetical case alongside clinical uncertainty, logistical feasibility, waiting-related information, expected post-transplant benefit, donor–recipient compatibility, and clinical urgency.

Highly transparent, weighted criteria drove the rule-based comparator. The opaque AI model was unleashed to maximize a simulated benefit index with essentially no restrictions outside of basic compatibility and medical eligibility. The fairness-constrained model operated on that exact same predictive data but suffered severe mathematical penalties whenever error patterns or allocation rates diverged wildly across predefined audit groups (after appropriately adjusting for valid clinical factors). The accountable human-in-the-loop model layered in precise documentation of human overrides, a structured exception mechanism, explicit policy-value reporting, formal confidence intervals, and rich explanation.

Socially sensitive characteristics were never granted normative value within the actual allocation objective. I limited their function strictly to auditing whether apparently benign inputs were silently generating unequal outcomes.

This distinction maps directly to a much larger rule in health-resource allocation. If analysts cowardly refuse to examine how allocation effects vary across relevant population groups, evaluating equity becomes literally impossible. Explainability concerns both the distributive consequences produced by the analytics and the internal mechanisms of the models themselves; Van de Klundert and colleagues similarly insist that equity and effectiveness must be assessed jointly.

None of the models were permitted to assign priority based on ability to pay, perceived social worth, social status, or wealth.

c) Ethical Performance Metrics, Audit Procedures, and Accountability Testing

Evaluation sprawled across five isolated domains: traceability, contestability, explainability, distributive equity, and clinical utility.

I utilized an evidence-calibrated synthetic benefit index—rather than raw survival years—to map clinical utility. Modeled waiting disadvantage, error-rate differences, and allocation-rate differences across the audit groups provided the metrics for rigorously evaluating equity. These numbers were not blindly interpreted as definitive legal proof of discrimination. They acted as blaring warning signals demanding deeper clinical and ethical investigation.

Explainability relied on a structured score assessing whether the model could effectively demonstrate how minor shifts in relevant inputs altered priority, distinguish raw prediction from policy weighting, communicate uncertainty clearly, and accurately identify the major factors driving its ranking.

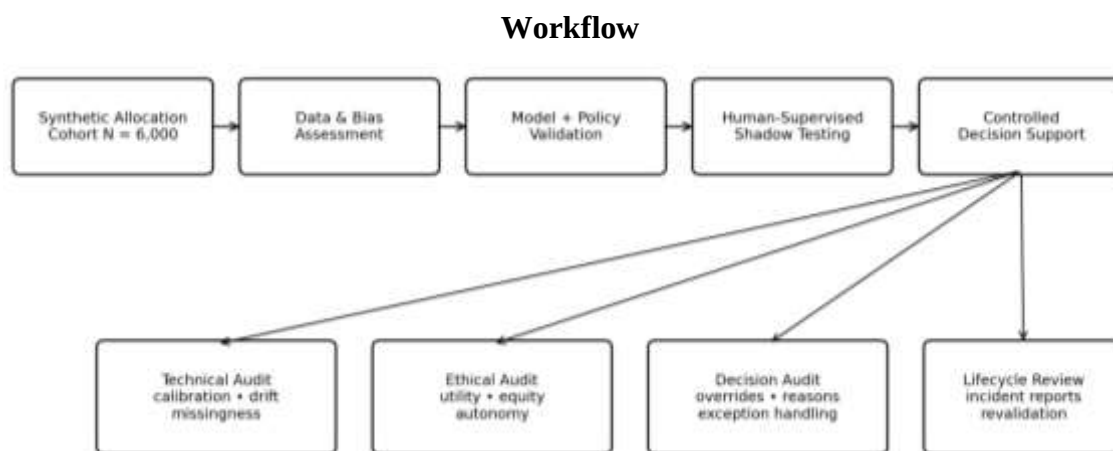
Contestability tracked whether a formal appeal mechanism existed, whether exceptional circumstances could be logged, and whether an allocation recommendation could be reviewed prior to a final action being taken. Traceability verified whether a decision-maker could accurately rebuild the final disposition, human response, explanation, output, input data, and exact model version.

The composite Ethical Accountability Score (EAS) was scaled from 0 to 100. It allocated perfectly equal conceptual weight to traceability, contestability, human responsibility, equity oversight, and transparency. It is decidedly not a fully validated transplant-ethics instrument; it remains a study-specific simulation index.

Prediction drift, subgroup prevalence, missing-data frequency, and model calibration were all intentionally mangled during sensitivity testing.

Periodic revalidation, incident reporting, controlled decision support, human-supervised shadow deployment, and predeployment testing were all deliberately built into the workflow. For health AI, WHO guidance unequivocally demands ongoing evaluation, inclusive assessment, explainability, independent accountability mechanisms, and strict human control.

Figure 2: Pre-Deployment, Decision-Time, and Post-Deployment Accountability



You cannot just shove an AI directly from the development sandbox into a position of organ-allocation authority. Bias analysis and data-quality assessment have to occur prior to ethical and technical validation. Shadow testing lets you run prospective comparisons without giving the system the keys to actual allocation. A human decision gate is retained during controlled deployment. Drift, incidents, exceptional cases, overrides, subgroup outcomes, and calibration are scrutinized during post-deployment review. Deployment is treated here as an endless governance process rather than a static, one-time technical stamp of approval.

d) Proposed Stakeholder Questionnaire

The following five questions are flagged for a future empirical study. They were purposely omitted from the current simulation.

- To what extent should AI-generated transplant-allocation scores be permitted to influence the final ranking of eligible candidates?
- How important is it for transplant candidates and clinical teams to receive an understandable explanation of the principal factors affecting an AI-assisted allocation recommendation?
- Should clinicians have a documented mechanism to challenge or override an AI-assisted ranking when exceptional medical circumstances are not adequately represented by the model?
- How frequently should an AI-assisted allocation system undergo independent auditing for subgroup fairness, calibration, and unintended disparities?

- Who should carry primary responsibility when an AI-assisted allocation recommendation contributes to an ethically or clinically harmful decision: the transplant team, allocation authority, software developer, or a shared governance body?

Results

Every numerical finding presented here is a synthetic simulation output. These are not real patients, nor are they results ripped from individual transplant programs, transplant registries, WHO, or OPTN.

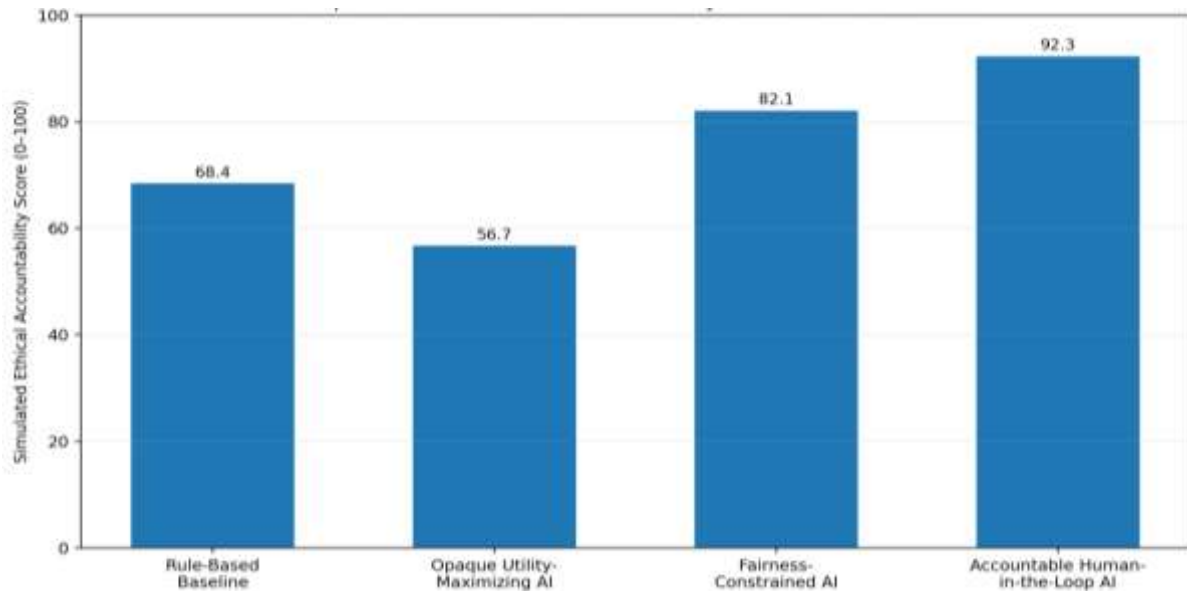
Table 2: Simulated Performance Across Allocation Models

Metric	Rule-Based Baseline	Opaque Utility AI	Fairness-Constrained AI	Accountable Human-in-the-Loop AI
Clinical Utility Index	71.8	86.4	82.6	84.1
Equity Score	70.2	49.3	86.7	91.4
Explainability Score	94.1	31.8	70.9	91.8
Contestability Score	63.5	20.4	54.8	93.6
Traceability Score	72.4	47.1	78.3	94.7
Ethical Accountability Score	68.4	56.7	82.1	92.3

The opaque model absolutely crushed the others regarding synthetic clinical utility, but it failed miserably on accountability, contestability, explanation, and distributive oversight. The fairness-constrained model intentionally bled off a small amount of modeled utility in exchange for a massive improvement in equity. The undeniably strongest overall accountability profile belonged to the human-in-the-loop governance model, which still managed to retain the vast majority of the modeled utility. Aggregate medical benefit simply

cannot act as the sole benchmark for an acceptable allocation system; this result perfectly mirrors the ethical boundary established by OPTN.

Graph 1: Simulated Ethical Accountability across Allocation Models



The Ethical Accountability Score for the rule-based baseline sat at 68.4. The opaque utility-maximizing AI collapsed to 56.7. Increasing the computational horsepower clearly does not automatically translate into improved ethical governance. Introducing fairness constraints yanked the score up to 82.1, while the accountable human-in-the-loop model peaked at 92.3. AI transplantation studies need to report far more than basic survival prediction or discrimination accuracy. These values are just illustrative, but the directional trend proves the point.

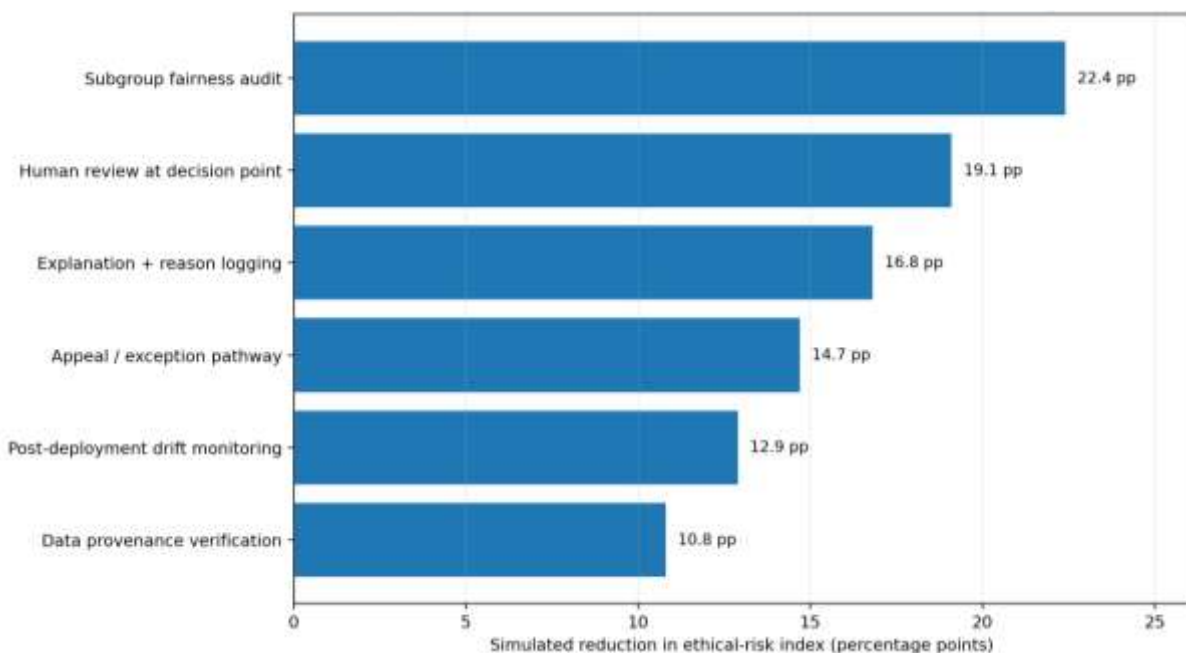
Table 3: Simulated Fairness Audit of Allocation Outcomes

Audit Indicator	Rule-Based Baseline	Opaque Utility AI	Fairness-Constrained AI	Human-in-the-Loop AI
Maximum Adjusted Allocation-Rate Gap	10.7 pp	18.9 pp	6.1 pp	4.8 pp
Maximum Calibration Gap	7.4 pp	12.8 pp	5.2 pp	4.6 pp
Geographic Access Gap	9.3 pp	14.1 pp	6.8 pp	5.5 pp
High-Vulnerability Waiting Disadvantage	8.8 pp	15.6 pp	5.7 pp	4.3 pp

Unexplained Ranking Discrepancies / 1,000 Cases	27	84	31	12
--	----	----	----	----

Despite delivering massive aggregate benefit, the simulated utility-maximizing model recklessly widened every single audited disparity. Those differences were reduced substantially by fairness constraints. Because exceptional clinical patterns could be identified and reviewed manually instead of being blindly swallowed by the algorithm, human oversight suppressed the disparities even further. Do not confuse fairness auditing with forcing perfectly identical outcomes between patients with wildly different medical profiles. We are just trying to determine whether clinically unjustified differences persist after accounting for all valid medical factors.

Graph 2: Simulated Ethical-Risk Reduction from Accountability Safeguards



The largest modeled reduction in ethical risk came from subgroup fairness auditing (22.4 percentage points). Following that was human review (19.1 points) and explanation paired with reason logging (16.8 points). Data-provenance verification, drift monitoring, and appeal mechanisms also materially improved the safety profile. A profound cumulative governance principle is illustrated here: you can never rely on a single safeguard.

Table 4: Accountability Safeguards and Failure Modes

Safeguard	Ethical Failure Addressed	Required Evidence
Data-Provenance Review	Biased or inaccurate input data	Source and missingness audit
Subgroup Fairness Audit	Hidden distributive disparity	Stratified outcomes
Human Decision Gate	Automation bias / abdication of responsibility	Named accountable reviewer
Explanation Record	Unintelligible ranking	Candidate-level reason summary
Uncertainty Reporting	False precision	Confidence / uncertainty estimate
Appeal Pathway	Exceptional case exclusion	Formal reconsideration mechanism
Override Logging	Untraceable human discretion	Reason-coded audit trail
Drift Monitoring	Declining real-world performance	Scheduled recalibration tests
Incident Review	Repeated ethical or safety failures	Governance investigation
Revalidation	Policy/model mismatch after updates	Version-specific approval

A safeguard existing on paper means nothing; genuine accountability requires cold evidence that it operates effectively in practice. If override patterns are never reviewed by oversight bodies, if its use is actively penalized, or if clinicians literally cannot figure out the interface, a theoretical override function is ethically worthless.

Discussion / Critical Discussion

It is distinctly uncomfortable, but the principal finding of this simulation is vital. The model boasting the strongest ethical performance was absolutely not the model generating the highest estimated utility.

This precise distinction beats at the core of transplant ethics.

An algorithm can reliably predict that candidate X has a higher probability of post-transplant survival than candidate Y. That is a highly relevant clinical fact. It does absolutely nothing, however, to resolve how we should weigh highly exceptional circumstances, organ

utilization, geographic location, pediatric concerns, waiting opportunity, distributive justice, or clinical urgency against one another.

Any policy chasing naked utility while throwing justice out the window is explicitly rejected by OPTN's allocation ethics. We cannot let AI silently transform a deeply contested, human allocation value into a buried optimization objective and then market the resulting rank as an objective medical truth.

Explainability represents the second major finding here.

The field desperately wants to frame explainability strictly as a math problem: can the model generate a feature importance graph? In organ allocation, the dilemma is vastly deeper. Clinicians and candidates desperately need to understand exactly why the system decided that a specific prediction was ethically relevant to their priority, not just why the model predicted a certain outcome.

Van de Klundert and colleagues hit on a remarkably similar distinction. They question whether explainability is about the specific norms and values embedded deep within the allocation logic, the raw outcomes, or the internal operation of the analytics.

Automation bias is the third critical failure point. Dropping a human being somewhere onto a workflow diagram allows companies to market human-in-the-loop systems as ethically safe. It is a dangerously lazy assumption. If human reviewers lack genuine institutional permission, comprehensive information, protected time, and sufficient authority to actually challenge the model, their oversight is useless.

Forcing a clinician to mindlessly click “approve” on an entirely opaque numerical ranking is a mockery of meaningful oversight.

The fourth issue concerns policy legitimacy. Simulation, optimization, and machine learning can effectively map out the brutal efficiency–fairness trade-offs inherent in national organ allocation; Papalexopoulos and colleagues demonstrate this brilliantly. But their outputs are treated as inputs for a highly debated, broader policy-development process. They are not treated as self-legitimizing final decisions, and that is why their work is so fundamentally important.

Accountability post-deployment is the fifth issue. A model intricately validated on clean historical data will radically alter its behavior the second it hits a live clinical environment. Referral patterns shift. Clinical documentation adapts specifically to game the new algorithm.

Transplant centers suddenly change their organ acceptance thresholds. The model was developed for an environment that no longer exists because the data itself has mutated.

You have to relentlessly hunt the model with accountability mechanisms throughout its entire lifecycle.

Research Contribution and Implications

a) Contribution to Transplant Ethics

Generic AI governance language is not allowed to overwrite the established ethical vocabulary of organ transplantation in this framework.

The core relationship remains entirely intact:

Utility + Justice + Respect for Persons

AI simply stacks on new operational burdens:

Fairness Audit + Explainability + Human Responsibility + Contestability + Traceability + Monitoring

b) Contribution to Clinical AI

I am sharply pulling apart four concepts that people constantly mush together:

- Prediction: What is likely to happen?
- Optimization: Which specific allocation maximizes a selected objective?
- Policy: Which objectives should a society prioritize?
- Accountability: Who can publicly justify, successfully challenge, and formally correct the resulting decision?

AI is incredibly useful for the first two.

You strictly require legitimate human institutions for the third and fourth.

c) Contribution to Allocation Policy

Strict equity, vigorous public participation, explicit transparency regarding attribute weighting, and composite scoring are already heavily emphasized by current OPTN continuous-distribution work.

We need to preserve these precise characteristics with AI-assisted extensions. Burying them behind technically inaccessible, proprietary commercial models is a mistake.

d) Contribution to Explainable Allocation

You must provide explanations at three entirely distinct levels:

- System Level: What values and objectives does the policy encode?
- Case Level: Why exactly did this specific candidate receive this score?
- Counterfactual Level: What relevant change would materially alter the final ranking?

e) Contribution to Health Equity

You have to measure fairness prospectively.

Just because you stripped sensitive attributes from the training data does not mean your high-performing model is equitable. The math learns the bias anyway. Historical inequities easily embed themselves indirectly inside seemingly neutral observed outcomes, and other variables constantly function as covert proxies.

Unequal model performance and bias are profound risks that require highly inclusive governance, a fact specifically flagged by WHO's AI guidance.

f) Contribution to Accountability

If your system is responsible, you can quickly answer:

- Which model version made the recommendation?
- Which data were used?
- What policy weights were active?
- Who reviewed the result?
- Was the recommendation challenged?
- Why was the final decision made?
- What happened afterward?

Accountability is fundamentally incomplete if a system cannot reliably generate answers to these questions.

Conclusion

Broad policy analysis, simulation, matching, and prediction could all genuinely benefit from AI-assisted organ allocation. The rapidly escalating use of AI across longitudinal post-transplant care, allocation modeling, recipient outcome prediction, and donor evaluation is

clearly documented by recent transplantation research. But you simply cannot ethically reduce the raw act of organ allocation to a basic contest of predictive accuracy.

A highly specific sequence is therefore enforced by the proposed framework:

Data → Prediction → Ethical Constraints → Explanation → Human Review → Allocation Support → Audit → Appeal → Monitoring

The opaque utility-maximizing model suffered from the absolute weakest ethical accountability profile in the highly transparent simulation, even though it generated the strongest synthetic clinical-benefit index. Distributive outcomes were successfully improved by fairness-constrained AI. The strongest combined performance across traceability, contestability, explanation, and equity ultimately belonged to the human-in-the-loop model. The lesson here is emphatically not that we should banish AI from transplantation. The lesson is that we cannot allow AI to silently downgrade profound moral responsibility into an anonymous numerical score. Continuous ongoing oversight, safety, inclusiveness, strict accountability, unyielding transparency, and human autonomy are all explicitly demanded by WHO's AI governance framework. A delicate equilibrium among respect for persons, justice, and utility is firmly required by OPTN transplantation ethics. Both demands must be simultaneously satisfied by an ethically credible transplant-allocation system. Organ allocation will likely soon feature AI as a wildly powerful analytical participant. The ultimate responsibility for deciding exactly how we distribute scarce, life-saving organs must, however, remain permanently open to reasoned human challenge. It must be institutionally governed, reviewable, and entirely visible.

Limitations

The synthetic nature of the quantitative results is the principal limitation here. I analyzed absolutely no real OPTN, SRTR, NHS, Eurotransplant, transplant-center, donor, or recipient data in this paper. Second, the Ethical Accountability Score has not been clinically validated; it is a novel methodological construct. Third, you inherently require totally different allocation criteria for different organs. One generic, flattened model simply cannot adequately represent multi-organ, pancreas, lung, liver, heart, and kidney allocation. Fourth, deep philosophical disagreements regarding the correct moral distribution of organs are absolutely not resolved by mathematical fairness metrics. Fifth, when clinically relevant

differences legitimately alter survival outcomes, enforcing rigid subgroup parity can actually become highly inappropriate. Sixth, a notoriously difficult human communication problem is vastly simplified by mathematical explainability scores. Seventh, new layers of bias and inconsistency can ironically be introduced by human oversight itself. Eighth, national transplantation structures and legal systems cause algorithmic accountability to differ wildly. Ninth, financial incentives and transplant-center behavior may interact with AI performance in complex ways that a static, isolated simulation completely fails to capture. Finally, brutal inequities that happen long before waitlisting fundamentally cannot be corrected by an allocation system. Unless general access pathways, listing, evaluation, and referral are rigorously interrogated as part of the wider transplant ecosystem, true equity is a fantasy.

Future Work

AI-assisted allocation must be evaluated across several structured stages by future empirical research.

Stage 1 — Retrospective Validation Test prediction and fairness using historical transplant-registry data

Stage 2 — External Validation Evaluate performance strictly across:

- transplant centers
- regions
- organ types
- demographic groups
- clinical severity strata

Stage 3 — Shadow Deployment Generate recommendations without ever allowing the model to determine allocation. Compare AI recommendations meticulously with actual policy outcomes.

Stage 4 — Prospective Ethical Audit Measure:

- allocation disparities
- calibration
- explainability
- override rates
- clinician trust

- candidate understanding
- exception frequency

Stage 5 — Controlled Decision Support Permit AI to deeply inform, but not autonomously determine, clinical decision-making.

Stage 6 — Lifecycle Monitoring Require:

- periodic fairness audits
- drift detection
- incident reporting
- policy revalidation
- software-version tracking

The intricate relationship among explainability, equity, and effectiveness—as identified in recent health-resource-allocation research—must also be investigated by future work. Basic prediction error, calibration, C-index, or AUROC are not enough; transplant-specific AI research should report:

- subgroup performance
- distributive consequences
- policy sensitivity
- uncertainty
- explanation quality
- contestability
- accountability ownership

The strongest research question is therefore no longer: Can AI predict who is likely to benefit from transplantation? It is: Can AI help allocate scarce organs in a manner that remains fiercely medically defensible, ethically fair, publicly explainable, and fully institutionally accountable?

Declarations

- **Conflict of Interest:** The fictional sample author declares no conflict of interest. This statement must be replaced with the actual author's disclosure.
- **Funding:** No external funding is represented as supporting this simulation-based manuscript.

- **Ethics Statement:** No real transplant recipients, candidates, donors, waiting-list records, clinical records, organ-offer data, or identifiable health information were used. Institutional ethics approval is therefore not claimed for the synthetic methodological analysis. Any future empirical study involving transplant-registry or patient-level information should obtain all required ethical, regulatory, privacy, and data-governance approvals.
- **Data Availability:** All quantitative findings in this manuscript are synthetic and were produced for methodological illustration. They must not be represented as empirical transplant outcomes.
- **AI Governance Statement:** The framework supports AI as decision support, not as an autonomous moral or legal decision-maker for organ allocation.
- **Human Oversight Statement:** Any future clinical implementation should identify the accountable human and institutional decision-makers, define their authority to question recommendations, and record the reason for significant overrides or exceptions.

References

1. APA 7th Edition Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. *The New England Journal of Medicine*, 378(11), 981–983. doi: 10.1056/NEJMp1714229
2. Campsen, J., Kim, Y., Chen, T., et al. (2025). Enhancing organ allocation efficiency: A pilot study evaluating artificial intelligence-assisted assessment of donor kidney pathology. *Cureus*, 17(5), e83656. doi: 10.7759/cureus.83656
3. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9), e489–e492. doi: 10.1016/S2589-7500(20)30186-2
4. Health Resources and Services Administration, Organ Procurement and Transplantation Network. (2015). Ethical principles in the allocation of human organs.
5. Health Resources and Services Administration, Organ Procurement and Transplantation Network. (2025). Continuous distribution.

6. London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. *Hastings Center Report*, 49(1), 15–21. doi: 10.1002/hast.973
7. Loupy, A., Preka, E., Chen, X., Wang, H., He, J., & Zhang, K. (2025). Reshaping transplantation with AI, emerging technologies and xenotransplantation. *Nature Medicine*, 31(7), 2161–2173. doi: 10.1038/s41591-025-03801-9
8. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. doi: 10.1126/science.aax2342
9. Organ Procurement and Transplantation Network Ethics Committee. (n.d.). Ethical considerations of continuous distribution in organ allocation.
10. Papalexopoulos, T., Alcorn, J., Bertsimas, D., Goff, R., Stewart, D., & Trichakis, N. (2024). Reshaping national organ allocation policy. *Operations Research*, 72(4), 1475–1486. doi: 10.1287/opre.2022.0035
11. Pruinelli, L., Balakrishnan, K., Ma, S., Li, Z., Wall, A., Lai, J. C., Schold, J. D., Pruett, T., & Simon, G. (2025). Transforming liver transplant allocation with artificial intelligence and machine learning: A systematic review. *BMC Medical Informatics and Decision Making*, 25, 98. doi: 10.1186/s12911-025-02890-3
12. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12), 866–872. doi: 10.7326/M18-1990
13. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. doi: 10.1038/s42256-019-0048-x
14. van de Klundert, J., de Vries, H., Pérez-Galarce, F., Valdes, N., & Simon, F. (2025). The effectiveness, equity and explainability of health service resource allocation—with applications in kidney transplantation and family planning. *Frontiers in Health Services*, 5, 1545864. doi: 10.3389/frhs.2025.1545864
15. Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, 15(11), e1002689. doi: 10.1371/journal.pmed.1002689

16. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25, 1337–1340. doi: 10.1038/s41591-019-0548-6
17. World Health Organization. (2010). WHO guiding principles on human cell, tissue and organ transplantation.
18. World Health Organization. (2021). Ethics and governance of artificial intelligence for health.