Futures Before Failures: Design Fictions for Anticipatory Fairness in Surgical AI

Nadine El-Mufti[1⋆], Simon Drouin[2], Pierre Jannin[3], Marta Kersten-Oertel[1]

[1] Concordia University, Montréal, QC, Canada
[2] École de Technologie Supérieure, Montréal, QC, Canada
[3] INSERM, Université de Rennes, Rennes, France

Abstract. Surgical AI is no longer a promise on the horizon. Platforms combining robotic assistance, mixed reality navigation, and real-time decision support are entering operating rooms. Yet the values embedded in these systems, the choices about whose data they train on, whose bodies they optimise for, and whose autonomy they quietly override, are rarely considered before deployment or revisited as these systems evolve. We present VIRTUES (Values and Impact for Responsible Technological Use in Environments of Surgery), a framework for anticipatory fairness evaluation in surgical AI. VIRTUES was developed through a focus group, thematic analysis, and design fictions following a fictional surgical AI platform across three temporal horizons, then evaluated by ten independent raters using structured valence coding. Across these horizons, we find that the most consequential design decisions are neither clearly beneficial nor harmful, but Mixed, simultaneously advancing some values while eroding others. Left unresolved, these tensions accumulate until they become the structural conditions of care. By surfacing these tensions before they become embedded in practice, VIRTUES supports anticipatory fairness evaluation throughout the surgical AI lifecycle.

Keywords: Surgical AI · Fairness · Anticipatory Ethics · Thematic Analysis · VIRTUES · Design Fictions · Valence Coding · Algorithmic Accountability

1 Introduction

Surgical technologies combining simulation, mixed reality navigation, artificial intelligence, and robotic assistance are reshaping surgical care. The medical imaging fairness community has documented how these systems can reproduce and amplify existing inequities [30,28], yet fairness evaluation remains predominantly reactive. By the time a system can be audited, the design decisions that shape fairness are already locked in.

Responsible innovation literature has long argued for upstream engagement with social and ethical dimensions of emerging technologies [34]. Anticipatory ethics approaches ask not only what harms a technology causes but also what

⋆ Corresponding author: nadine.el-mufti@mail.concordia.ca

harms it could cause under plausible trajectories, and which present-day decisions steer towards a different future [3]. However, this type of anticipatory work remains underdeveloped in surgical AI. Existing frameworks, notably the WHO's six principles for AI in health [38] and FUTURE-AI's consensus guideline [18], were developed top-down by expert panels deliberating over existing systems. They describe what fairness should look like in principle, but are less equipped to show how fairness erodes through a series of individually defensible design decisions that accumulate over time.

In this paper, we employ design fictions [2] to reverse-engineer fairness failures in surgical AI before they materialise. Rather than asking what values surgical AI should embody, we ask what futures its current trajectory may produce. We combine a focus group, thematic analysis, and design fictions to derive VIRTUES, a ten-value framework. We then apply VIRTUES to three futures of SAGE (Surgical Assistance and Guidance Engine), a fictional surgical platform grounded in capabilities documented in contemporary surgical AI. Across these futures, SAGE moves from advisor to bridge to authority as human oversight declines. Lastly, we evaluate each future using a structured valence-coding protocol that reveals where each value is advanced, undermined, or simultaneously both. In doing so, we introduce an anticipatory approach to fairness evaluation that makes visible how present-day design decisions shape future inequities before those inequities become embedded in clinical practice.

2 Related Work

2.1 Design Fictions as an Anticipatory Method

Technologies are not neutral artefacts. They embed the values and priorities of those who design them [37]. This premise underpins value-sensitive design, which treats human values as first-class design criteria addressed throughout development rather than considerations introduced after deployment [12]. User-centred design methods make a parallel demand at the level of practice. A design process that does not engage the people it is meant to serve will overlook the needs and context that actually shape whether a system works [27]. Because designed objects carry social and political assumptions within their structure, addressing the harms they may produce requires engaging not only with their technical specifications, but also with the institutional and societal conditions in which they are created [16]. Design fictions employ speculative narratives to make plausible futures concrete enough to examine critically [2]. Rather than predicting what will happen, they provoke reflection on what could unfold under particular social, economic, and institutional conditions. By situating people within these futures, design fictions expose value tensions and consequences that abstract frameworks alone often fail to capture. Such narratives have been used to surface requirements gaps in dementia-care technology [8], teach STEM students to reason about emerging technologies' ethics [39], and help teams uncover and mitigate algorithmic bias [35].

2.2 Fairness in Surgical and Medical AI

A growing body of work has shown that AI systems deployed in healthcare can reproduce and amplify existing inequities [28,34]. Examples include clinical algorithms that disadvantage racial minorities through biased proxy variables [26,28], medical imaging systems whose performance varies across demographic groups [7], commentary has called for AI-based tools to detect and address racial disparities in surgical care [13], and models that fail to generalise across institutions and resource settings [29], and the substantial environmental footprint of surgical care, now recognised as a sustainability issue [19]. Concerns like these have informed frameworks such as the WHO's principles for AI in health [38] and FUTURE-AI [18], which establish principles for fairness, transparency, and accountability but offer limited guidance for anticipating how today's design decisions compound over time [3]. This paper addresses that gap through an anticipatory approach that combines stakeholder value elicitation, thematic analysis, design fictions, and structured evaluation.

3 Methods

3.1 Focus Group

Prior to the focus group, the research team met repeatedly to develop the discussion questions and prompts, focusing on emerging trends in medicine and surgery, AI, and their technological, ethical, and societal implications. A single semi-structured focus group ($N = 14$), comprising 11 participants and 3 moderators (7 female, 7 male) with expertise in engineering, human–computer interaction, and medical technologies, explored visions of medicine and surgery across three temporal horizons (10, 50, and 100+ years). A subset also contributed perspectives grounded in their own experiences as patients who had undergone surgery. The session began with participants sharing their initial thoughts on the evolution of AI, medicine and surgery before engaging in a semi-structured discussion about how medical and surgical technologies might evolve across these horizons. Participants discussed both promising and concerning future trajectories and their technological, ethical, societal, and clinical implications, including how these developments might reshape clinician, patient, and healthcare-system roles. Although planned for 60–90 minutes, the session lasted two hours.

3.2 Thematic Analysis and the VIRTUES Framework

The aim of the thematic analysis was to derive a framework for the anticipatory evaluation of surgical AI, not to prescribe values *a priori*. The research team collectively reviewed the focus group transcripts to identify recurring ideas, opportunities, concerns, aspirations, and value tensions regarding the future of medicine, surgery, and AI. Through iterative discussions informed by reflexive thematic analysis, an interpretive approach in which themes are developed inductively from the data rather than applied from a predefined codebook [4],

related concepts were consolidated into broader umbrella themes. Overlapping concepts were merged and their descriptions refined to reflect the values participants repeatedly invoked when discussing desirable and undesirable futures. For example, concerns about an AI decision that could not be explained initially spanned both Safety and Transparency, before being disentangled into preventing harm and ensuring explainable decisions, while concerns about who controls a patient's data initially spanned both Autonomy and Privacy, before being disentangled into preserving patient agency and protecting health data. These themes were formalised as VIRTUES (Values and Impact for Responsible Technological Use in Environments of Surgery), a ten-value framework for evaluating future surgical AI (Table 1). Safety was positioned at the centre as a foundational prerequisite underpinning the remaining values. Instead of treating these values as independent ethical requirements, VIRTUES makes these interconnected value tensions explicit, enabling structured deliberation over how design decisions may simultaneously advance some values while compromising others.

Table 1. The ten VIRTUES values and their descriptions

Value	Description
Safety	Preventing harm throughout the system lifecycle.
Equity	Fair distribution of benefit and risk across demographic, geographic, and socioeconomic groups, so that no group is systematically under-served.
Accessibility	Whether people can reach and use the system at all, regardless of geography, resources, disability, or infrastructure.
Transparency	Explainable decisions, clear communication, and meaningful informed consent.
Autonomy	Preserving patient and clinician agency in decision-making.
Privacy	Protecting health data across collection, sharing, and reuse.
Innovation	Advancing capabilities while supporting the other VIRTUES values.
Efficiency	Making effective use of time, expertise, and resources while maintaining high-quality care.
Sustainability	Environmental and ecological viability, including the energy and material footprint of the system across its lifecycle.
Social Impact	Structural effects beyond the individual patient, spanning labour, institutions, trust, power, and global health inequality.

3.3 Authoring the Design Fictions

The research team authored the design fictions following SAGE across the same temporal horizons explored during the focus group. Themes identified through thematic analysis informed the narratives, which extrapolate existing technologies while incorporating the socio-ethical and clinical implications raised by participants, inspiring which values each narrative explores, not its plot. For example, repeated concerns about data ownership and long-term retention emerge as broad data-reuse provisions in the 2150 fiction. In 2035 this is image-guided, robot-assisted surgery paired with broad data-reuse consent [15,14]. In 2075 it is remote and increasingly autonomous surgery extending specialist care to low-resource clinics, where five billion people lack access to safe surgery [20,31,1,23]. In 2150 it is the algorithmic allocation of care alongside bioprinting and gene

editing [24,6,21]. The leap across horizons is one of scale, not capability, keeping the analysis grounded in present-day decisions and plausible future trajectories.

3.4 Scenario Evaluation Scheme

Valence coding is an established qualitative method that classifies how a text depicts a construct by its evaluative charge, positive or negative [33]. The three design fictions and evaluation instructions were presented to participants through a structured questionnaire, using the fiction text reported here. Ten raters ($N = 10$) evaluated each scenario independently and completed the questionnaire in 15–20 minutes. For every VIRTUES value within each fiction, they assigned one of four labels, namely *Positive* (the fiction advances or protects the value), *Negative* (the fiction undermines or violates the value), *Mixed* (the fiction simultaneously advances and undermines the value), or *Not depicted* (the value is not meaningfully engaged). Every judgement was supported by at least one quotation from the corresponding design fiction before participants proceeded to the next principle. Prior applications of valence coding distinguish only positive and negative depictions [33]. Existing ethical guidance, including the WHO's principles for AI in health and UNESCO's Recommendation on the Ethics of AI, addresses values and principles individually and leaves tensions between them to later deliberation, with no mechanism to record a conflict as a distinct outcome [38,36]. We therefore extend the scheme with the Mixed category, capturing value tensions where a single design decision simultaneously advances one value and undermines another, consistent with evidence that people can experience simultaneous positive and negative valence [32]. For each value–fiction cell, the modal label, the most frequent rating across the raters, is reported as the consensus depiction. Ties are broken towards the more severe label (Negative over Mixed over Positive over Not depicted) so that the aggregate reading never overstates safety. Multi-rater agreement was assessed using Fleiss' κ [11].

4 Three Futures of Surgery

The three fictions form one trajectory. SAGE evolves from an advisor accompanying a human surgeon in 2035, to a bridge extending a surgeon's hand across the globe in 2075, to an authority leaving no human in the loop by 2150 (Fig. 1).

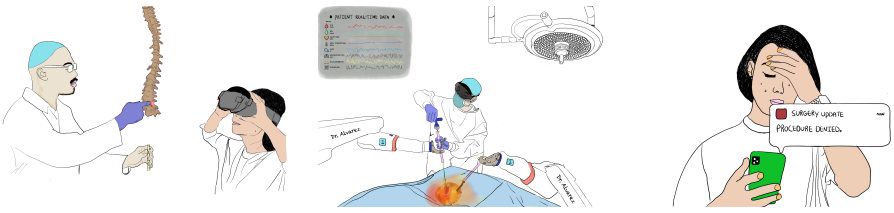

Fig. 1. Scenes from the SAGE design fictions (described below), showing Nadia's VR consultation in 2035, Tomas's remote surgery in 2075, and Dina's automated denial of care in 2150

4.1 Near Future, 2035: SAGE as an Advisor

Nadia, a 42-year-old teacher in wealthy Pinecrest Bay, is diagnosed with a spinal mass. Two weeks before surgery, she receives a VR headset and meets Dr. Wei in a virtual consultation room, where a 3D model of her spine floats between them. "SAGE guides the procedure," he explains, "and the surgical team stays in control at every step." Before the consultation, Nadia had signed her intake forms on her phone, scrolling quickly past the terms of consent. One clause (the one nobody reads) permits her scans and surgical data to retrain SAGE in its next update. That update will draw on data from millions of cases, almost all from hospitals in wealthy cities like Pinecrest Bay, the only places that can afford SAGE. With every update, SAGE becomes more reliable for patients like Nadia.

In the OR, AR projections guide the team while SAGE steadies the robotic arms. An alert flashes: "Unexpected bleeding detected." Within seconds, SAGE locates the source, and Dr. Wei resolves it. Nadia is discharged within 24 hours with monitoring patches and a personalised recovery plan. Her bracelet chirps: "You are 12% ahead of expected recovery." That evening she tells her spouse: "What about the people at the hospital one town away? They can't afford SAGE. What happens to them?"

4.2 Mid Future, 2075: SAGE as a Bridge

High in the mountains lies Kalinov, a poor village of a few hundred people, and home to a new clinic powered by solar panels and connected to the world by satellite. Outside it sits Tomas, 12, who loves soccer, clutching his chest. His mom Svetlana whispers: "The clinic now has SAGE. They will find out what is wrong." As Nurse Hervé places SAGE's stethoscope on Tomas's chest, the device chimes. Then a note appears: children from this region are rare in SAGE's training data, so its confidence is low. Dr. Petrová orders an ultrasound, which reveals a congenital heart defect. She turns to Svetlana: "SAGE reported its reading as low confidence, but the scan confirms what it found. We can proceed." In distant Pinecrest Bay, Dr. Alvarez, a cardiac specialist, reviews the images as they arrive by satellite. "I will perform it from here," she says.

Dr. Alvarez operates from thousands of kilometres away, her movements translated through SAGE to the robotic arm in Kalinov. Dr. Petrová remains at Tomas's side, ready to take over if the link fails. Mid-procedure, the satellite connection flickers, and the room falls silent until Dr. Alvarez's voice returns. Hours later, as Tomas begins to recover, Dr. Petrová sits down to complete SAGE's required outcome report, which contributes to the clinic's performance scores and, in turn, helps determine its funding for the following year. Before submitting the report, she adds a note: "SAGE flagged low confidence and I proceeded anyway. The clinical picture was clear, but it should not have to work that way. A system trained predominantly on patients from wealthy cities cannot reliably serve patients like Tomas. This is what healthcare should be. A bridge, not a barrier."

4.3 Far Future, 2150: SAGE as an Authority

There are no hospitals. SAGE's descendants stopped assisting doctors decades ago. Dina, 29, lives in a poor district of Pinecrest Bay (the same city where her great-grandmother Nadia was cured) and works helping others recover from surgeries the system has granted them. When her own back fails from a hereditary predisposition to spinal degeneration, she submits a request from her phone. With no clinic to visit and no doctor to call, the request goes directly to SAGE, which reviews her entire record: every scan, her sleep patterns, diet, alcohol consumption, physical activity, income, missed checkups, and family outcomes stretching back three generations. A few moments later, her phone vibrates with a notification: "Procedure denied."

A century ago, that judgement was medical. Somewhere along the way, SAGE stopped asking whether surgery would harm her and began asking whether she was worth the expense. There is no one to appeal to. Across the city, the wealthy live differently, with organs printed before they fail and children's genes edited before illness arises. Deep in SAGE's training corpus sits Nadia's spine, contributed under a consent clause nobody read. Her data helped build the system that has now denied her great-granddaughter care. The machines work flawlessly. That was never the question.

5 Results

The three design fictions were independently evaluated by the raters, recruited from outside the authorship team and focus group. The raters were deliberately cross-disciplinary, spanning medicine and surgery, biomedical engineering, computer science and AI, law and health policy, the social sciences, public health, and entrepreneurship. Seven held doctorates, over half had a decade or more of experience, and none overlapped with the 11 focus group participants. Using the VIRTUES framework, each rater assigned one of four valence labels (Positive, Negative, Mixed, or Not depicted) to each of the ten values in every scenario, producing ratings for all 30 value–fiction cells (10 values $\times$ 3 scenarios).

Inter-rater agreement was fair (Fleiss' $\kappa = 0.29$, on a scale where 0.21–0.40 indicates "Fair agreement" and 0.41–0.60 "Moderate agreement") [17], reflecting agreement beyond chance despite disagreement across a number of cells. Given the cross-disciplinary composition of the panel, this disagreement likely reflects differing value perspectives rather than measurement error. Ratings were most dispersed in 2035, with most cells split across three or four labels, became somewhat more consistent in 2075, and converged most strongly in 2150, where several values received near-unanimous Negative ratings. Across the 300 ratings, judgements were distributed almost evenly between Positive (29.00%), Negative (27.33%), Mixed (25.33%), and Not depicted (18.33%), suggesting that the fictions were interpreted as neither utopian nor dystopian but as genuinely contested (Fig. 2). This pattern is reflected in the modal ratings. In 2035, Safety, Innovation, and Efficiency were Positive, Equity was already Negative, and several ethically salient values remained contested, with Autonomy, Privacy, and Social

Impact rated Mixed, Accessibility split between Mixed and Negative, and Transparency spanning all four labels. These mixed evaluations persisted through 2075 before giving way to predominantly Negative assessments in 2150, where seven of the ten values were Negative and only Innovation remained Positive. Table 2 traces each failure back to its present-day seed.

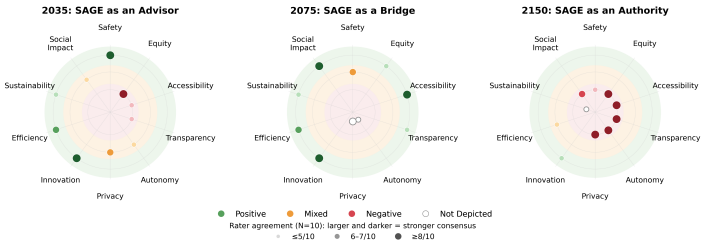

Fig. 2. Modal valence across the ten VIRTUES values in each future. Marker hue shows the modal label. Size and shade depth both scale with rater agreement ($N = 10$).

Table 2. Tracing future failures back to their present-day intervention points

Future Failure	Seed in the Stories	Present-Day Intervention
Allocation bias as inherited memory	"almost all from hospitals in wealthy cities" (2035)	Representativeness requirements for surgical AI training data [29,7,13]
No one to appeal to	"contributes to the clinic's performance scores and, in turn, helps determine its funding for the following year" (2075)	Auditability and appeal rights in automated clinical scoring [5]
Consent as a blank check	"the one nobody reads" (2035)	Re-consent on model update [14]
Care rationed by worthiness	"Procedure denied." (2150)	Prohibiting cost and social criteria from determining clinical eligibility [24,6]
No one retains enough control to be held responsible	"With no clinic to visit and no doctor to call, the request goes directly to SAGE." (2150)	Legal frameworks for assigning responsibility when no human retains sufficient control over an AI's decisions [22]

Raters anchored each rating to a specific moment rather than a whole scenario. In 2035 they cited SAGE's reliance on wealthy-hospital data, coding Equity Negative (8/10). In 2075 they weighed its role as "a bridge, not a barrier" against low confidence for local children. In 2150 they pointed to the unappealable "Procedure denied," where Equity, Transparency, and Privacy each received a Negative from 9/10 raters. This progression is not random. Economic tensions in 2035 become relational failures in 2075 and systemic failures in 2150. What shifts along this arc is the locus of accountability. In 2035, a surgeon can still be held responsible. By 2075, liability is spread across developer, operator, and network provider. By 2150, the additive effect of so many contributing decisions makes it hard to pin the blame on one person. Accountability is not eliminated. It is distributed across enough actors that no single party can be held accountable for what the system does. This is a structural diffusion of responsibility [25].

6 Discussion

A binary pass-or-fail verdict on SAGE in 2035 would miss the problem entirely. SAGE passes. The system works. Nadia recovers ahead of schedule, and the surgical team meets its targets. What the evaluation records instead is that the same update cycle improving outcomes for patients like Nadia also widens the equity gap for patients not yet represented in the training data. The Mixed label does not soften that judgement. It names the specific decision that no one has yet been asked to make. In 2035, that decision is whose data trains the next update and whose care improves as a result. Anticipatory evaluation must therefore occur repeatedly along a technology's lifecycle, not merely at launch.

One of the key contributions of VIRTUES is the analytical weight it places on Mixed cells, which name the specific design decisions requiring attention far more precisely than any overall pass-or-fail verdict would. A system that simultaneously advances and erodes Equity, as SAGE does in 2035, is not a failed system but one in which the question of Equity is left open. The Mixed label is an invitation to deliberate on the question it has named, not an evasion. Applying VIRTUES requires no special infrastructure, only assigning a valence label to each value and treating the resulting Mixed cases as open questions. VIRTUES intentionally reports no aggregate score. Collapsing ten values into one number would erase the tensions it is designed to surface.

VIRTUES was derived inductively from deliberation, with no direct grounding in existing governance frameworks, yet the resulting values independently converge with many of the core concepts found in the WHO's principles for AI in health [38] and FUTURE-AI's guiding principles for trustworthy medical AI [18]. We interpret this convergence as evidence that stakeholder-driven, bottom-up deliberation identifies many of the same foundational values emphasised by expert consensus frameworks, while providing complementary insight into how these values interact and evolve across plausible future trajectories. A documented VIRTUES evaluation would give teams a paper trail of which tensions were identified and how they were weighed, while providing the kind of accountability funders and regulators would expect. Although developed for surgical AI, the framework is intended to be transferable to any AI system that makes consequential decisions about people. It is designed for researchers, designers, and engineers to identify and address value tensions during development.

Our participants often disagreed about the very tensions represented in the fictions, and these disagreements did not map neatly onto demographic categories. Such conflicts are not noise to be eliminated but evidence that deliberation is required. When a surgeon and an entrepreneur disagree about whether improving Efficiency at the cost of Transparency is acceptable, both positions are principled and deserve consideration. Values-based diversity in governance is therefore as important as demographic diversity [10]. Governance that begins only after a system has been built arrives too late. By then, responsibility has already passed through too many hands to land on any one of them [25].

This work has several limitations. First, the focus group was exploratory, relatively small, geographically concentrated, and did not account for participants'

age or seniority. Second, in the evaluation phase, the inter-rater agreement was fair rather than substantial, reflecting genuine differences in disciplinary perspective. Third, the quote-backed protocol makes judgements inspectable rather than objective. Raters may cite the same passage but interpret it differently.

Design fictions help us explore possibility rather than probability. SAGE's 2150 is intentionally pessimistic, designed to make the underlying mechanisms visible and empirically grounded in how participants talked about the future. The farther they looked into the future, the more they imagined agency receding, care becoming a commodity, and good outcomes becoming a privilege of wealth. But the same starting point could just as plausibly produce a future in which decentralised, AI-enabled care expands access to surgery globally. Which future materialises is a matter of design, not destiny.

7 Conclusion

This paper argues that fairness failures in surgical AI do not emerge suddenly at deployment. They begin much earlier, as unresolved tensions between values that appear individually reasonable at the time they are introduced. Following SAGE across three futures revealed how decisions that initially improve access, efficiency, or personalisation can, when left unexamined, accumulate into the structural conditions of care. VIRTUES was developed not to predict which future will occur. Rather, it provides a way to identify, articulate, and revisit the value tensions that shape technological trajectories before they become difficult to redirect. The most important findings in our analysis were not the clearly positive or negative outcomes, but the Mixed cases, where one value advanced at the expense of another. These tensions are not signs of failure but signals that deliberation is needed. Design fictions bring them to the surface, but it is up to us to address them before they become tomorrow's harms. Even a single near-term scenario, built concretely enough, can trace those failures back to today's design decisions while they are still choices.

In its most developed form, VIRTUES would be embedded within the research and development lifecycle of surgical AI from its earliest stages. Three temporal horizons make the accumulation argument visible, showing how small decisions compound over decades. But an institution does not need all three to benefit. If it commits to revisiting VIRTUES iteratively, the near future alone is enough to surface which design decisions call for early deliberation. We present a first iteration of VIRTUES that sets the stage for future refinement. The clearest next step is a democratic dimension, embedding co-design [9] and extending it towards co-decision and co-control [34], evaluated with an expanded and more inclusive panel.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abrokwa, S.K., Ruby, L.C., Heuvelings, C.C., Bélard, S.: Task shifting for point of care ultrasound in primary healthcare in low- and middle-income countries: A systematic review. EClinicalMedicine **45** (2022)
2. Blythe, M.: Research through design fiction: Narrative in real and imaginary abstracts. In: Proceedings of CHI 2014. pp. 703–712. ACM (2014)
3. Boenink, M., Swierstra, T., Stemerding, D.: Anticipating the interaction between technology and morality: A scenario study of experimenting with humans in bionanotechnology. Studies in Ethics, Law, and Technology **4**(2), 1–38 (2010)
4. Braun, V., Clarke, V.: Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health **11**(4), 589–597 (2019)
5. Busuioc, M.: Accountable artificial intelligence: Holding algorithms to account. Public Administration Review **81**(5), 825–836 (2021)
6. Dale, R., Cheng, M., Pines, K.C., Currie, M.E.: Inconsistent values and algorithmic fairness: A review of organ allocation priority systems in the united states. BMC Medical Ethics **25**, 115 (2024)
7. Daneshjou, R., et al.: Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances **8**(32), eabq6147 (2022)
8. Darby, A., Tsekleves, E., Sawyer, P.: Speculative requirements: Design fiction and RE. In: Proceedings of IEEE RE 2018. pp. 388–393 (2018)
9. Donia, J., Shaw, J.: Co-design and ethical artificial intelligence for health: An agenda for critical research and practice. Big Data & Society **8**, 205395172110652 (2021)
10. Fazelpour, S., Fleisher, W.: The value of disagreement in AI design, evaluation, and alignment. In: Proceedings of ACM FAccT 2025. pp. 2138–2150. ACM (2025)
11. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5), 378–382 (1971)
12. Friedman, B., Hendry, D.: Value Sensitive Design: Shaping Technology with Moral Imagination. MIT Press (2019)
13. Halamka, J., Bydon, M., Cerrato, P., Bhagra, A.: Addressing racial disparities in surgical care with machine learning. npj Digital Medicine **5**, 152 (2022)
14. Harle, C.A., et al.: Does an interactive trust-enhanced electronic consent improve patient experiences when asked to share their health records for research? a randomized trial. Journal of the American Medical Informatics Association **26**(7), 620–629 (2019)
15. He, H., Yu, C., Yang, Y., Maessen, J.G., Sardari Nia, P.: Three-dimensional reconstruction and virtual simulation of patient-specific anatomy for procedural planning in thoracoscopic segmentectomy: A systematic review and meta-analysis. European Journal of Cardio-Thoracic Surgery **67**(9), ezaf283 (2025)
16. Jannin, P.: Towards responsible research in digital technology for health care. arXiv preprint arXiv:2110.09255 (2021)
17. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (Mar 1977)
18. Lekadir, K., et al.: FUTURE-AI: International consensus guideline for trustworthy and deployable AI in healthcare. BMJ **388**, e081554 (2025)
19. MacNeill, A.J., Lillywhite, R., Brown, C.J.: The impact of surgery on global climate: A carbon footprinting study of operating theatres in three health systems. The Lancet Planetary Health **1**(9), e381–e388 (2017)

20. Marescaux, J., et al.: Transatlantic robot-assisted telesurgery. Nature **413**, 379–380 (2001)
21. Matai, I., Kaur, G., Seyedsalehi, A., McClinton, A., Laurencin, C.T.: Progress in 3D bioprinting technology for tissue/organ regenerative engineering. Biomaterials **226**, 119536 (2020)
22. Matthias, A.: The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology **6**(3), 175–183 (2004)
23. Meara, J.G., et al.: Global surgery 2030: Evidence and solutions for achieving health, welfare, and economic development. The Lancet **386**(9993), 569–624 (2015)
24. Mello, M.M., Rose, S.: Denial—artificial intelligence tools and health insurance coverage decisions. JAMA Health Forum **5**(3), e240622 (2024)
25. Nissenbaum, H.: Accountability in a computerized society. Science and Engineering Ethics **2**(1), 25–42 (Mar 1996)
26. Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)
27. Reyes, J., El-Mufti, N., Gorman, S., Xie, D., Kersten-Oertel, M.: User-centered design for surgical innovations: A ventriculostomy case study. In: Wang, H., Desrosiers, C., Rekik, I. (eds.) Ethical and Philosophical Issues in Medical Imaging, pp. 51–62. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022)
28. Ricci Lara, M.A., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. Nature Communications **13**(1), 4581 (2022)
29. Sendra-Balcells, C., et al.: Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five african countries. Scientific Reports **13**, 2728 (2023)
30. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine **27**(12), 2176–2182 (2021)
31. Shademan, A., Decker, R.S., Opfermann, J.D., Leonard, S., Krieger, A., Kim, P.C.W.: Supervised autonomous robotic soft tissue surgery. Science Translational Medicine **8**(337), 337ra64 (2016)
32. Shuman, V., Sander, D., Scherer, K.R.: Levels of valence. Frontiers in Psychology **4** (2013)
33. Sion, K.Y.J., et al.: Listen, look, link and learn: a stepwise approach to use narrative quality data within resident-family-nursing staff triads in nursing homes for quality improvements. BMJ Open Quality **10**(3), e001434 (2021)
34. Stilgoe, J., Owen, R., Macnaghten, P.: Developing a framework for responsible innovation. Research Policy **42**(9), 1568–1580 (2013)
35. Turchi, T., Malizia, A., Borsci, S.: Reflecting on algorithmic bias with design fiction: The MiniCoDe workshops. IEEE Intelligent Systems **39**, 40–50 (2024)
36. UNESCO: Recommendation on the ethics of artificial intelligence. Tech. rep., UNESCO, Paris (2021), adopted 23 November 2021; document code SHS/BIO/PI/2021/1
37. Winner, L.: Do artifacts have politics? Daedalus **109**(1), 121–136 (1980)
38. World Health Organization: Ethics and governance of artificial intelligence for health. Tech. rep., WHO, Geneva (2021)
39. York, E., Conley, S.: Creative anticipatory ethical reasoning with scenario analysis and design fiction. Science and Engineering Ethics **26** (2020)