# Session Title: Algorithms in Society

## Paper 1: On the Use of Sensitive Characteristics in Medical Risk Assessment

On November 10th, 2024, The American Heart Association released a statement describing a new set of equations used to predict heart disease in patients. A notable feature of the new equations is that they no longer include race as a risk factor for heart disease, even though the prevalence of heart disease varies significantly across race. To justify the change, the statement cited "the growing consensus to remove the use of race from clinical algorithms broadly in medicine." The AHA announcement is one of the latest developments in a broader movement that seeks to abolish the practice of using race as a risk factor in healthcare decision-making (Cerdeña et al 2020; Vyas et al 2020; Arenas-Gallo 2024; Coots et al 2025).

This movement is supported by a view in medical ethics that we call *the emerging orthodoxy*. The emerging orthodoxy makes three claims. First, that treating race as a risk factor harms some patient populations by reinforcing a pernicious form of biological essentialism among medical professionals. Second, that treating race as a risk factor is clinically unnecessary, as the predictive value of race can be captured by race-neutral determinants of health. Therefore, race should either *never* be treated as a risk factor, or it should only be treated as a risk factor in *extraordinary cases* where a very high burden of proof is met. The AHA statement takes a particularly strong position on the removal of race, suggesting that race should be removed as a risk factor *a priori*, which is to say it should be removed even before we compare the performance of race-free and race-sensitive risk models.

In this paper, we argue that the emerging orthodoxy should be rejected in favor of what we call *the moderate position.* According to the moderate position, race should be considered as a risk factor in medical decision-making whenever (1) race provides diagnostically relevant evidence that is not screened off by other variables and (2) taking race into account is expected to produce better health outcomes for medically disadvantaged groups than available alternatives. Our paper makes two significant contributions to the debate about the use race as a risk factor in clinical decision-making. First, we show that the evidence for the core commitments of the emerging orthodoxy is weaker than it first appears, and that the emerging orthodoxy generalizes in problematic ways (such as by conflicting with equipoise requirements for clinical trials). Second, we advance a positive argument from cases for the moderate position. We argue that, when physicians fail to treat race as a risk factor for disease under the conditions specified by the moderate position, they thereby violate their professional duties of beneficence and respect for autonomy.

*References*

1. Arenas-Gallo C, Michie M, Jones N, Pronovost PJ, Vince RA Jr. Race-Based Screening under the Public Health Ethics Microscope - The Case of Prostate Cancer. N Engl J Med. 2024 Aug 1;391(5):468-474. doi: 10.1056/NEJMms2402322. PMID: 39083779.
2. Cerdeña JP, Plaisime MV, Tsai J. From race-based to race-conscious medicine: how anti-racist uprisings call us to act. Lancet. 2020;396:1125-1128. [PMID: 33038972] doi:10.1016/S0140-6736(20)32076-6
3. Coots, Soroush Saghafian, David M. Kent, Sharad Goel, A Framework for Considering the Value of Race and Ethnicity in Estimating Disease Risk, Annals of Internal Medicine, 178, 1, (98-107), (2025). https://doi.org/10.7326/M23-3166
4. Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight —reconsidering the use of race correction in clinical algorithms. N Engl J Med. 2020;383:874-882. [PMID: 32853499] doi:10.1056/NEJMms2004740.

## Paper 2: Aspirational Harm: a Concern for Justice

AI-generated products increasingly inform and shape our shared representations and narratives. Whether we are considering arts and entertainment, education, health and well-being, human relationships, or beyond, our resources for interpreting ourselves and others now include AI-generated images, text, audio, and so on.

The notion of aspirational harm has recently been introduced as a distinctive type of harm that can occur in AI systems. Aspirational harm adds a third category to the two already well-established categories of harm that can occur in AI systems: allocative harm (inequitable distribution of resources or opportunity) and representational harms (denigrating or diminishing certain groups along the lines of identity). This third type of harm occurs when shared interpretative resources become diminished or distorted for particular groups in such a way that it hinders the ability for individual members to imagine practical possibilities and alternatives. Informed by the literature on hermeneutical injustice, aspirational harm rests on the insight that shared interpretative resources shape our understanding of the world. Aspirational harm builds on this insight by pointing out how these resources not only shape our understanding of what IS but also what COULD BE. In other words, they shape the contours of our practical imagination. When resources are distorted or diminished, then, not only can an epistemic harm take place (distorting a knower's understanding of the world) but also an agential harm (compromising an agent's ability to imagine and, thus, forge ahead on new paths of action or self-determination).  Conceptual resources can restrict the practical imagination in various ways, including making certain possibilities less intelligible, less desirable, or riskier.

In this talk, we aim to: (a) further articulate the agential harm at stake in aspirational harm; and (b) identify when aspirational harm amounts to an injustice. Ultimately, we suggest that this sort of harm poses a distinct threat to autonomy. Put into Rawlsian terms, aspirational harm compromises the second moral power (the capacity to form, revise, and rationally pursue one's own conception of the good). This being the case, having an adequate and healthy supply of interpretative resources looks to function something like a primary good. This raises interesting questions about how the principles of justice might guide AI systems insofar as they produce, distribute, and shape hermeneutical resources.

## Paper 3: Can We Trust Transparent Artificial Intelligence?

Policy proposals about regulating AI often put forward a list of desirable properties that AI systems should be required to have. These lists often include both trustworthiness and transparency. While both properties are plausibly desirable, they turn out to be in tension. In this paper, I argue that if the goal of transparent AI is to promote public knowledge of how AI systems operate, and if the goal of trustworthy AI is to engender public trust in those systems, then the goals  of trustworthiness and transparency inherently conflict.

My argument focuses on doxastic accounts of trust, on which trust is a species of belief. One of the constitutive norms governing belief is that believers ought to update their beliefs by taking available evidence into account (Wanderer and Townsend 2013). If a stakeholder A trusts AI system S to ϕ (e.g., make decisions in a manner that is safe, reliable, nondiscriminatory, etc.), it therefore follows that A believes that S will ϕ, and further that A ought to update their belief that S will ϕ in light of the available evidence. However, this requirement to update one's trust in AI systems on the basis of available evidence conflicts with another constitute norm governing trust: I argue that part of what it is for A to trust that B will ϕ is that A will not take steps to verify whether B will ϕ. (Contra Ronald Regan, "trust but verify" is an oxymoron!)

These two constitutive norms on trust are in tension in a way that creates a previously unrecognized dilemma for the regulation of AI. To promote the goal of transparent AI, institutions must supply the public with evidence that AI systems have normatively desirable properties and encourage the public to update their beliefs about those systems in light of that evidence. However, to promote the goal of public trust in AI, institutions must encourage the public *not* to actively seek out such evidence. The twin goals of trustworthy and transparent AI thus face an inherently tension that has not yet been appreciated. Drawing on the literature in social epistemology on public trust in experts (Nguyen 2022; O'Neill 2002; Christiano 2012; Duijf 2021; Goldman 2001), I argue that policymakers seeking to regulate AI should sacrifice the goal of public-facing transparency for the sake of the goal of public trust in AI.

*References*

1. Baier, Annette. 1986. “Trust and Antitrust.” Ethics 96 (2): 231–60. https://doi.org/10.1086/292745.
2. Christiano, Thomas. 2012. “Rational Deliberation among Experts and Citizens.” In Deliberative Systems: Deliberative Democracy at the Large Scale, edited by Jane Mansbridge and John Parkinson, 27–51. Theories of Institutional Design. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139178914.003.
3. Duijf, Hein. 2021. “Should One Trust Experts?” Synthese 199 (3): 9289–9312. https://doi.org/10.1007/s11229-021-03203-7.
4. Goldman, Alvin I. 2001. “Experts: Which Ones Should You Trust?” Philosophy and Phenomenological Research 63 (1): 85–110. https://doi.org/10.2307/3071090.
5. Nguyen, C. Thi. 2022. “Transparency Is Surveillance.” Philosophy and Phenomenological Research 105 (2): 331–61. https://doi.org/10.1111/phpr.12823.
6. O’Neill, Onora. 2002. A Question of Trust: The BBC Reith Lectures 2002. 2002nd edition. Cambridge: Cambridge University Press.
7. Wanderer, Jeremy, and Leo Townsend. 2013. “Is It Rational to Trust?” Philosophy Compass 8 (1): 1–14. https://doi.org/10.1111/j.1747-9991.2012.00533.x.