Hard Choices in Artificial Intelligence: Resolving Normative Uncertainty through Sociotechnical Commitments

Roel Dobbe[1] Thomas Krendl Gilbert[2] Yonatan Mintz[3]

[1]AI Now Institute, New York University [2]Center for Human-Compatible AI, University of California Berkeley [3]School of Industrial and Systems Engineering, Georgia Institute of Technology

Challenges

The ML/AI disciplines are coming to terms with the reality that issues of **safety and fairness cannot be solved through strictly technical means**.

Values like safety or fairness are inherently *vague* in terms of their nature, perception and meaning for different stakeholders and contexts.

Vagueness must be resolved democratically across the AI development and deployment process to ensure that a system is accountably safe for all.

How is safety vague?

Vagueness arises as safety is further specified. We consider three typical instantiations.

Protection

- Active prevention of harm or injury
- Scope of harms
- Unclear how private practices and publicly expected standards are reconcilable

Robustness

- Ability to withstand adverse conditions
- Conditions of harms
- Inconsistent standards across stakeholders (e.g. designers, users, administrators)

Resiliency

- Effective response to stress or difficulty
- Failsafe procedure
- Undetermined what should be done by whom to prevent and minimize harm

Why democratic channels for dissent?

Just as the "stress point" of civil engineering is the agreed-upon strain any bridge can handle before buckling, the critical point for human-compatible AI is the safeguarding of shared moral agency; having power throughout design, training, and deployment.

Sociotechnical Commitments, Dilemmas and Virtues

Stage*	Featurization	Optimization	Integration
Commitments:			
• *formal*	Negotiate what can(not) be modeled	Assess limits of inferred parameters	Assess agency of stakeholders
• *substantive*	Create flexibility for stakeholder input	Negotiate validation with stakeholders	Establish open feedback channels
• *discursive*	Anticipate value-conflicts	Anticipate verification or revisit design	Secure trustworthiness of channels
Dilemma	Model-Based vs Model-Free	Validation vs Verification	Exit vs Voice
Virtue	Context Discernment	Stewardship	Public Accountability

*meant as iterative and not strictly linear

Implications – The Hard Choices Framework can help:

…formulate new interdisciplinary approaches to AI development that center affected stakeholders.

…determine relationship between fairness / safety metrics and procedural justice / accountability

…identify issues where market incentives deprioritize self-determination and ignore safety needs

…address road-blocks to dissent in tech companies for workers (NDAs, lack of IRB, retaliation) and users