

Hard Choices in Artificial Intelligence

July 13, 2022
Thomas Krendl Gilbert

The Problem: What does "Beneficial AI" Mean?

Safe Model Learning



Politics of Refusal



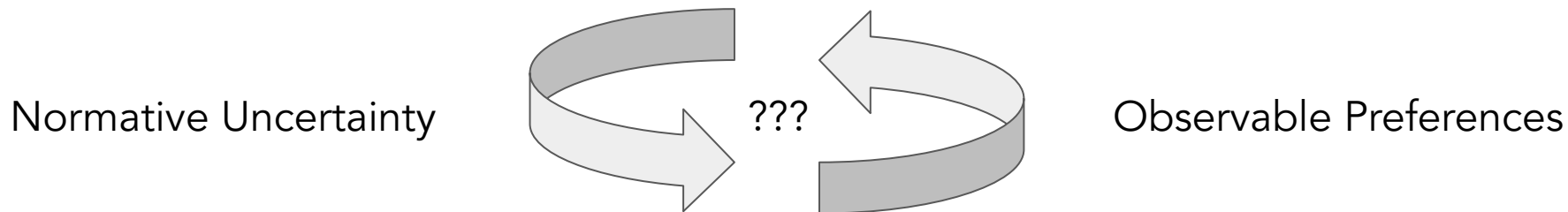
vs.

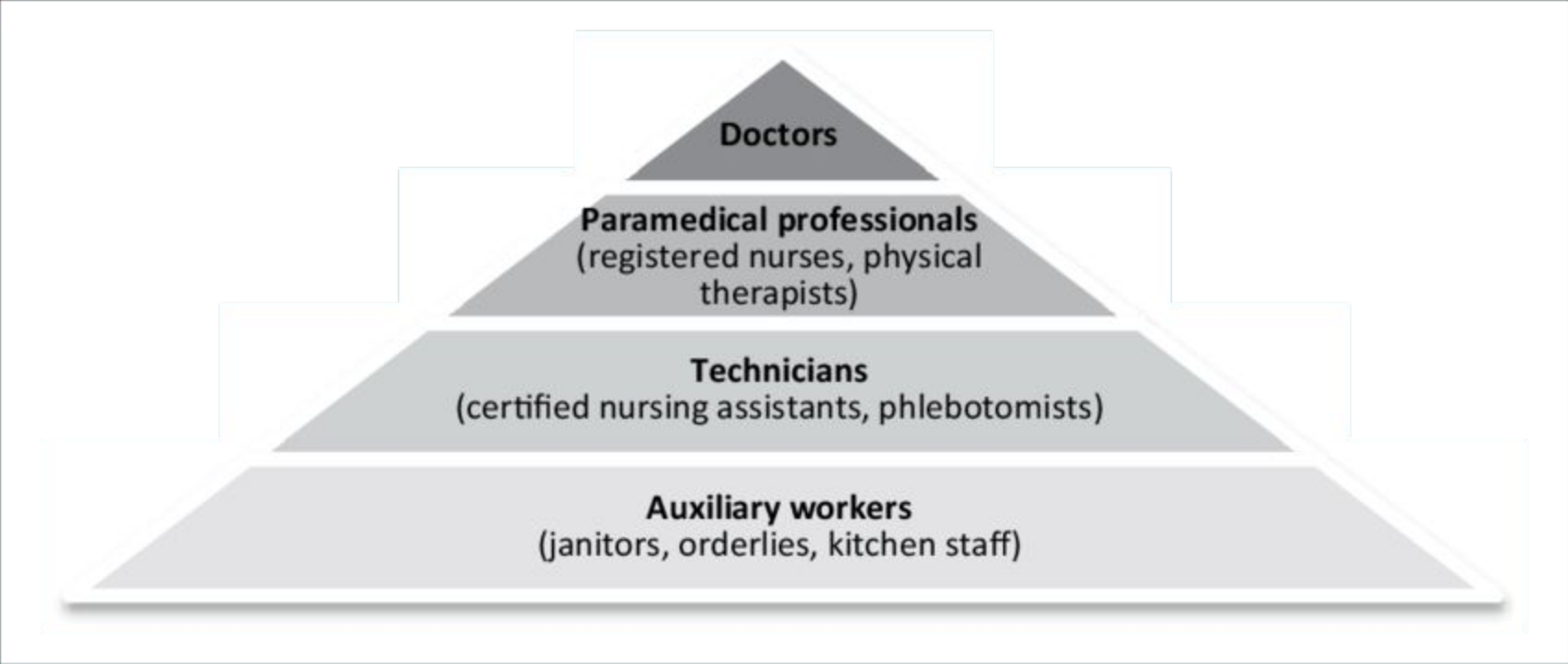
Machine Ethics as Normative Uncertainty

Metanormativism: identify what additional information is needed to aggregate in order to resolve uncertainty. (William MacAskill 2016)

Machine Ethics as Normative Uncertainty?

Metanormativism: identify what additional information is needed to aggregate in order to resolve uncertainty. (William MacAskill 2016)





Doctors

Paramedical professionals
(registered nurses, physical therapists)

Technicians
(certified nursing assistants, phlebotomists)

Auxiliary workers
(janitors, orderlies, kitchen staff)

Expanding the Definition: Normative Indeterminacy

1. Social norms are not simply discoverable or encodable, they are also enacted by the conditions under which AI tools are developed.

Expanding the Definition: Normative Indeterminacy

1. Social norms are not simply discoverable or encodable, they are also enacted by the conditions under which AI tools are developed.
2. Developing safe AI through the crafting of rules and protocols amounts to developing *practices* that affirm distinct value commitments.

Normative Indeterminacy and Diagnosing “Harm”

Epistemicism

-perception

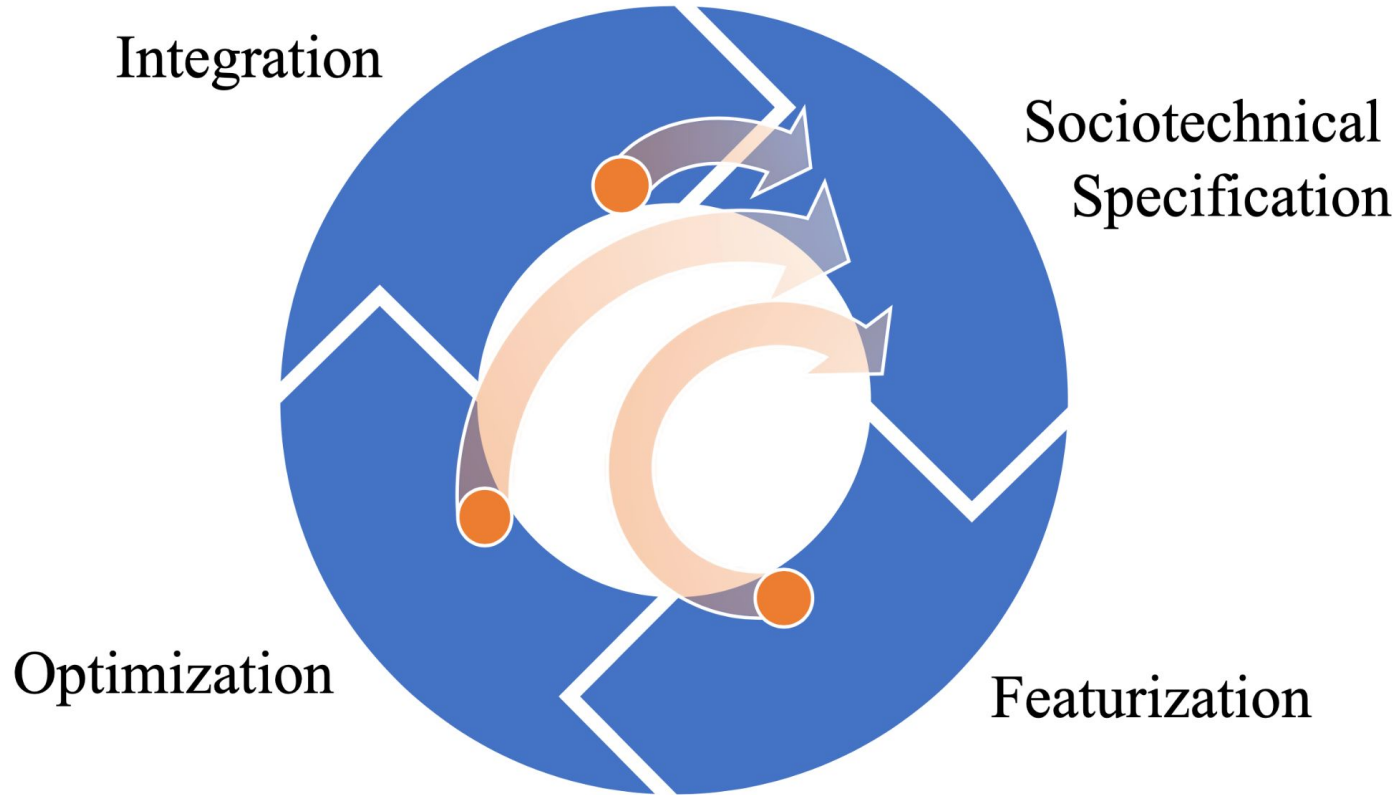
Semantic indeterminism

-language

Ontic incomparabilism

-reality

Hard Choices in Artificial Intelligence (HCAI) Framework



From Dobbe, Gilbert, and Mintz. "Hard Choices in Artificial Intelligence". Published at *Artificial Intelligence*.

The Problem of Featurization

Edge Cases (today)



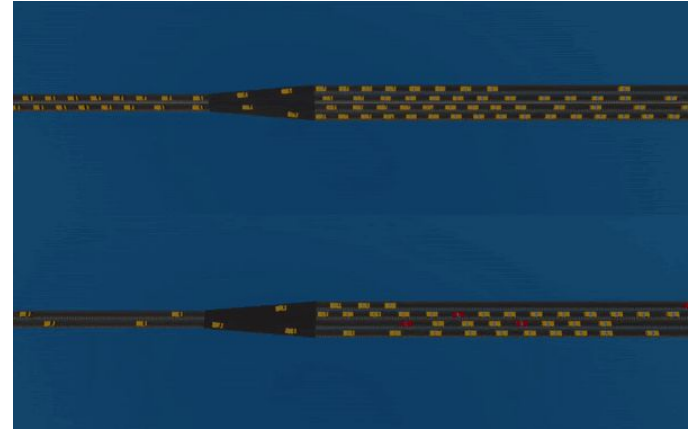
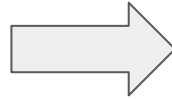
Reward Hacking (future?)



Today



Future?



Accidents will happen when there is a mismatch between the features that matter for a given task and those assumed by the agent or model.

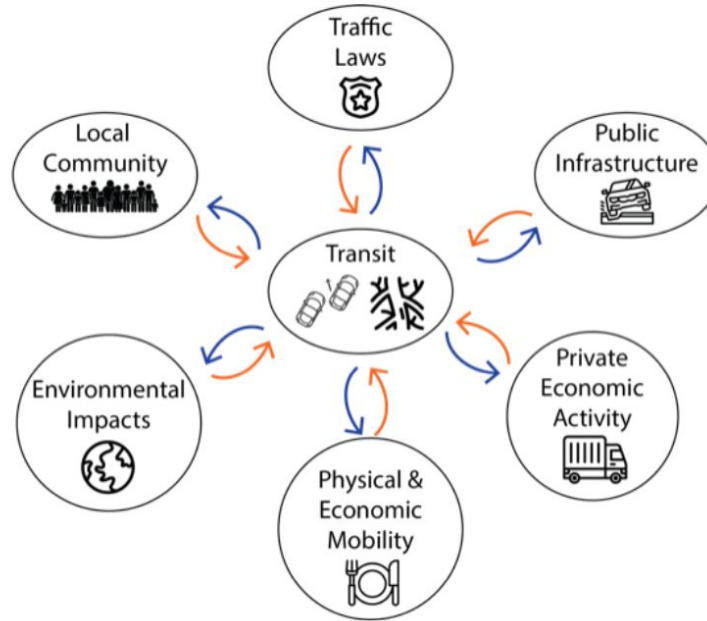
The Problem of Optimization

Online Radicalization



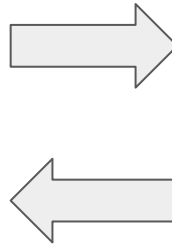
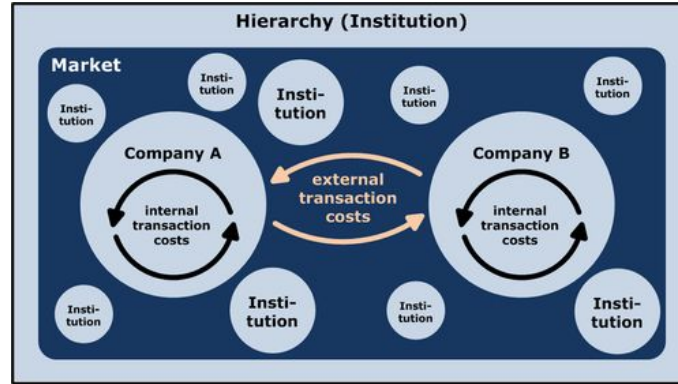
Road Wear





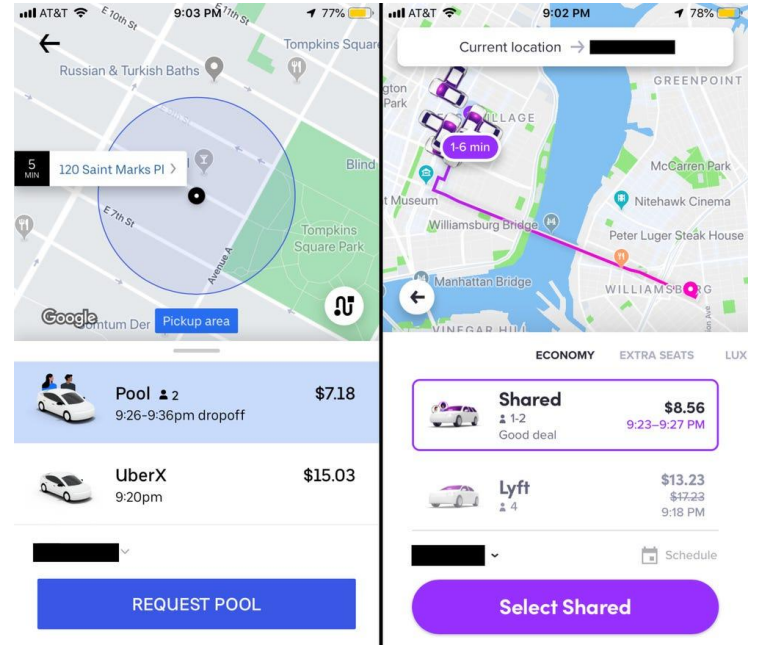
Damage happens when sources of feedback between the system and other activities are ignored or mismatched, generating unstable conditions.

The Problem of Integration: Whither Public Space?





or...



Poor integration happens when a single company figures out how to privatize roads, instead of designing automated systems to support the public interest.

Towards Sociotechnical Specification: Agency and Voice

Designers can only roughly fix the parameters of system performance as they bear on relevant human indeterminacies.

Towards Sociotechnical Specification: Agency and Voice

Designers can only roughly fix the parameters of system performance as they bear on relevant human indeterminacies.



vs.



Questions