



HARVARD
UNIVERSITY

CS288

01

Ethics

About Me



Michael Pope

Philosophy Graduate Fellow

Embedded EthiCS @ Harvard



Identify ethical
& social issues



Reason
through
ethical &
social issues



Communicate
their reasoned
position



Design ethically
& socially
responsible
systems

Embedded
EthiCS
course
modules
help students
to...

Agenda

Pursuing Ethical AI 01

Value Sensitive Design 02

AI for Criminal Justice 03





Given the
case studies
you've seen,
what makes
AI4SI
successful
or *good*?

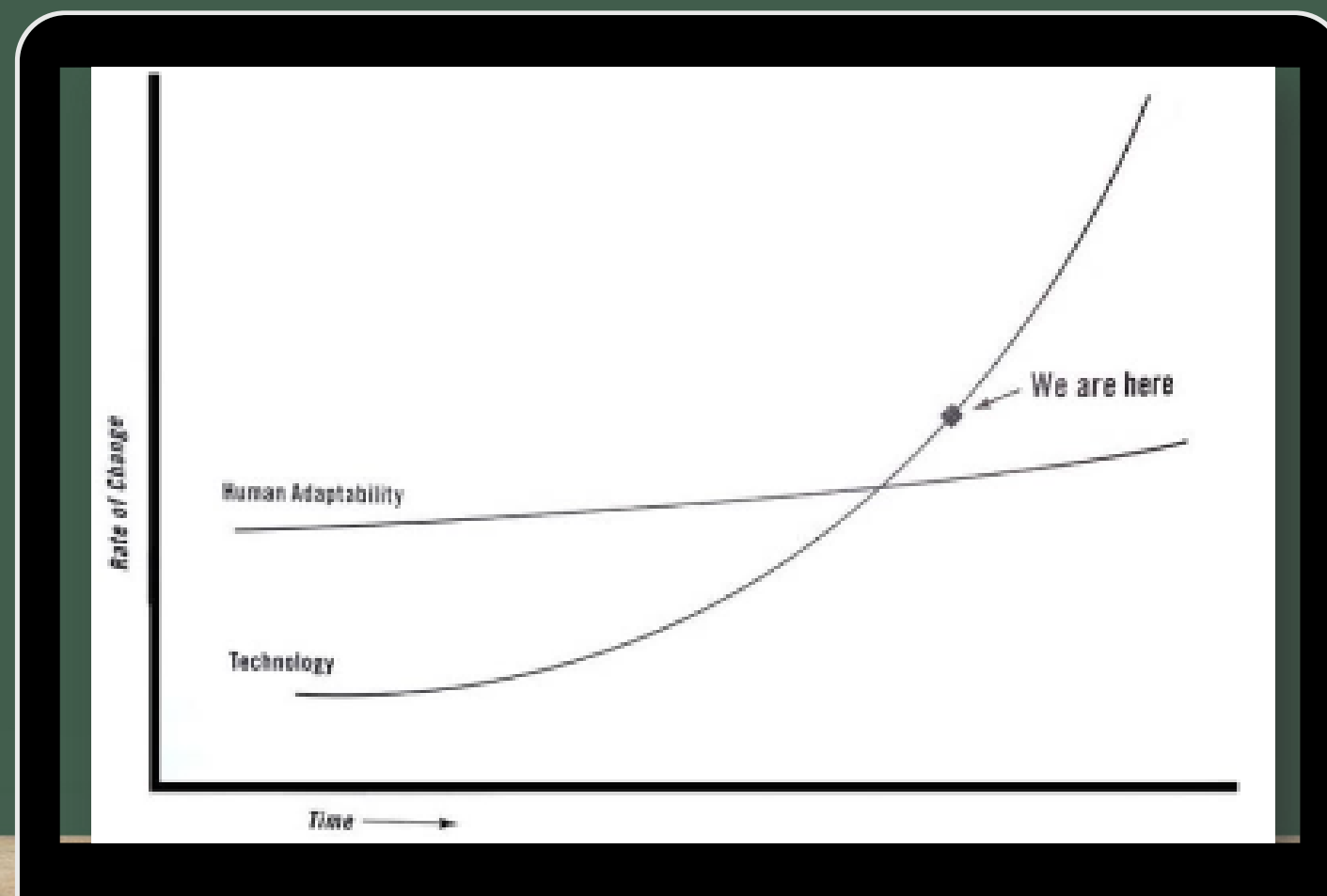
Discussion



Word Cloud

Pursuing a Participant Stance

- Identify who might be affected by the AI intervention and what they care about
- Consider how the system could impact those affected as it scales up



AI4SI is especially well positioned to design & deploy socially responsible systems, but:



What problem(s) *should* you address?

What trade-offs arise? How *should* you address trade-offs?

Value Sensitive Design



Shaping
Technology
with Moral
Imagination



“Ultimately, value sensitive design asks that the technical, civil, and other communities broaden the goals and criteria for judging the quality of technological systems to include those that advance human flourishing”

Identify:

- Stakeholders
- Their values
- Impacts at scale



How can attention to goals and criteria integrate ethical values into the design process?



Imagine you're designing a smart home system that integrates energy use data with recommendations to promote sustainability



Stakeholders



Direct Stakeholders

People directly impacted by the technology.

- People who use the technology
- People on whom the technology is used

Indirect Stakeholders

People who are impacted by the technology, but don't interact with it directly.

- Family members of users
- People who can't access the technology

Values

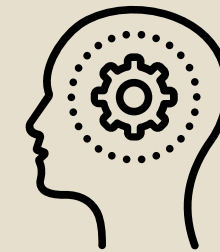


What do different stakeholders value?

What do they want or need? What do they think is important?

Ethical Value Examples

- 1 Justice
- 2 Harm
- 3 Privacy
- 4 Security
- 5 Autonomy
- 6 Community
- 7 Health
- 8 Flourishing



Conceptualizations



Empirical Feedback



Scale

Time

- What might be the immediate and short-term impacts of the intervention?
- What long-term impacts will the technology have?



Pervasiveness

- What are the impacts of the technology when only a small group of people are affected?
- Do these effects change as the technology becomes more widely adopted? How?

COMPAS

Correctional
Offender
Management
Profiling for
Alternative
Sanctions



What is it?

COMPAS informs sentencing practices by predicting risks of recidivism and assessing the needs of offenders. It's used for:

- pretrial decisions
- sentencing following conviction
- determining supervision for probation and parole
- placing offenders in state and federal prisons at the appropriate security level



equivant

COMPAS

Correctional
Offender
Management
Profiling for
Alternative
Sanctions



Calculating Risk Scores

Type 1	Low (1-4)	Medium (5-7)	High (8-10)
Type 2	Unlikely (1-2)	Probable (3-4)	Highly Probable (5-10)
Type 3	Unlikely (1-5)	Probable (6-7)	Highly Probable (8-10)
Type 4	Unlikely (1-4)	Probable (5-7)	Highly Probable (8-10)

Scale	Scale Type
Violent Recidivism Risk	1
General Recidivism Risk	1
Pretrial Release Risk	1
Criminal Involvement	1
History of Noncompliance	1
History of Violence	1
Current Violence	1
Criminal Associates/Peers	4
Substance Abuse	2
Financial Problems/Poverty	3
Vocational/Education Problems	3
Criminal Thinking	3
Family Criminality	3
Social Environment Problems	3
Leisure and Recreation	3
Residential Instability	3
Social Adjustment Problems	3
Socialization Failure	3
Criminal Opportunity	3
Criminal Personality	3
Social Isolation	3

Violent Recidivism Risk:

- Age
- Age-at-first-arrest
- History of violence
- Vocation/education
- History of noncompliance

Discussion

How might values enter into:

Immersion?

System Design?

Deployment?

APPLYING THE FRAMEWORK

Stakeholder

Direct	Indirect

Values

Design and Stakeholder Inputs

Time

Short-term	Long-term

Pervasiveness

Limited Adoption	Widespread Adoption

COMPAS

Correctional
Offender
Management
Profiling for
Alternative
Sanctions

COMPAS is:

- Complicated (137 possible variables)
- Proprietary



Questions About Fairness, Justice, & Bias

A Popular Algorithm Is No Better at Predicting Crimes Than Random People

The COMPAS tool is widely used to assess a defendant's risk of committing more crimes, but a new study puts its usefulness into perspective.

By Ed Yong

Why it's so damn hard to make AI fair and unbiased

There are competing notions of fairness — and sometimes they're totally incompatible with each other.

By Sigal Samuel | Apr 19, 2022, 6:00am EDT

Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing

By Andrew Lee Park / In Law Meets World / February 19, 2019 / 21 Min read / Add comment

Setting the Record Straight: What the COMPAS Core Risk and Need Assessment Is and Is Not

by Eugenie Jackson and Christina Mendoza

The Age of Secrecy and Unfairness in Recidivism Prediction

by Cynthia Rudin, Caroline Wang, and Beau Coker

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

Transparency

17

Ruden, Wang, and Coker (2020):

"But what do secret algorithms like COMPAS actually compute? If we could peek inside, perhaps we could determine what makes these algorithms tick—and whether we need them at all."

- it is easier to debate the fairness of a transparent model than a proprietary model
- transparency can facilitate fairness by providing clear, verifiable explanations for life-changing decisions



Discussion

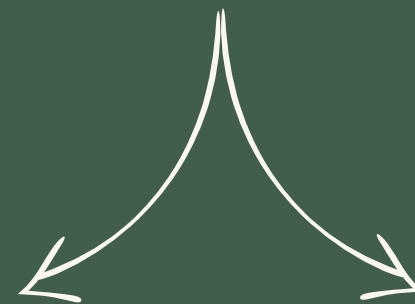
What does it mean for an algorithm to be transparent?

Transparency



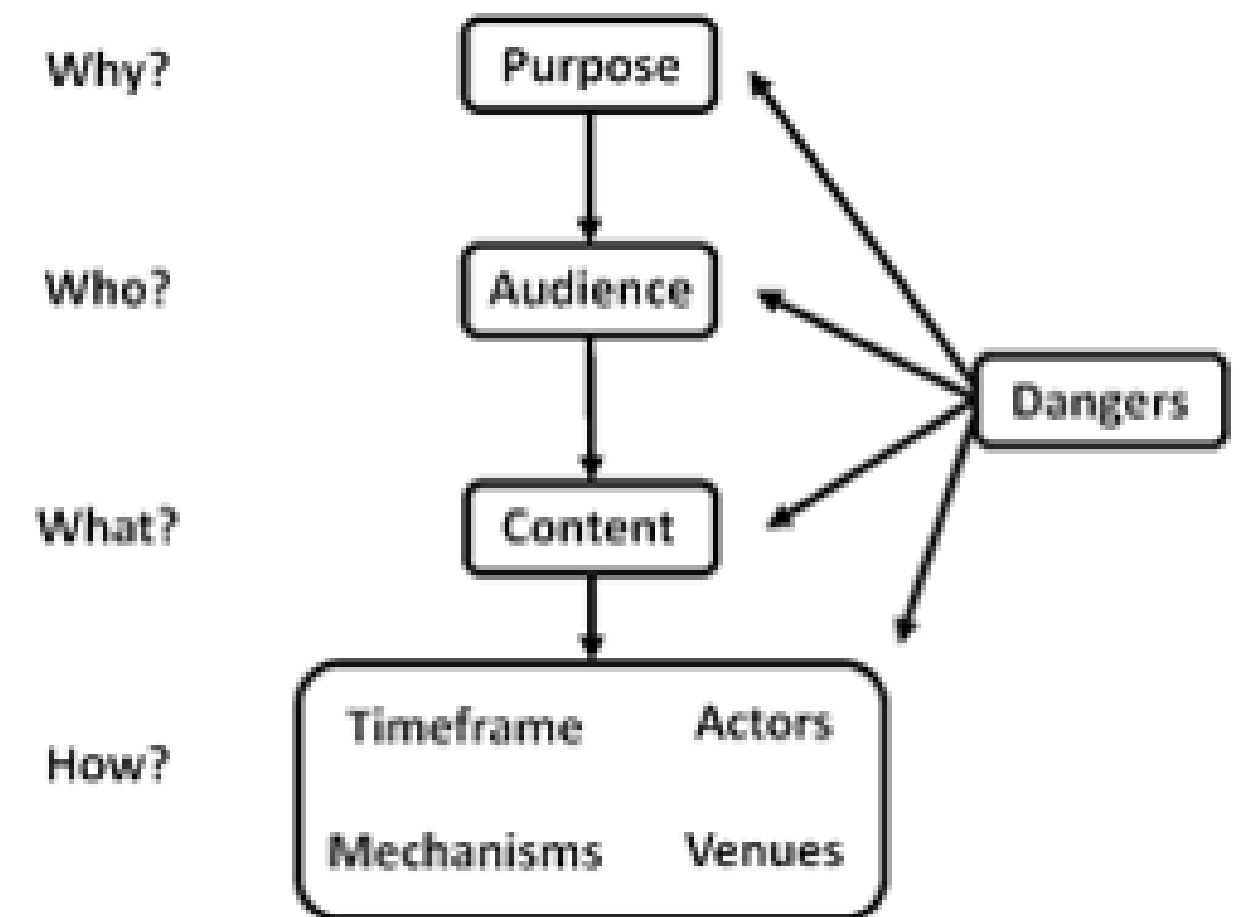
What is it?

Broad Transparency: "any process in which some entity makes information about its own activities available, to be used in further decisions and actions."



Re-use Transparency

Assessment Transparency
(Who is the disclosure for?)



From Elliott (2020)



Following this module, has your original response changed or evolved?

How might this influence your project?

Reflection

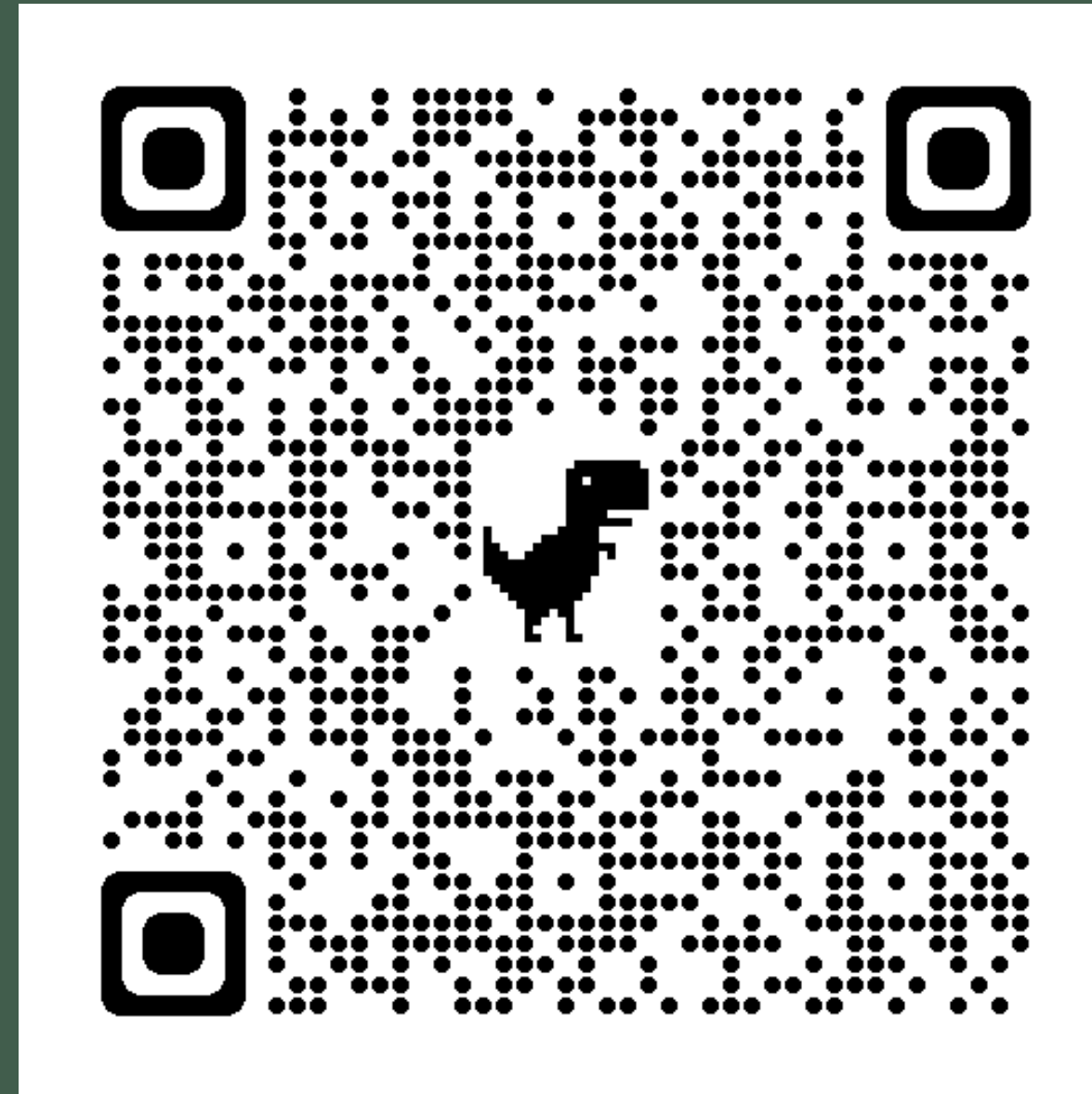
Given the case studies you've seen,
what makes AI4SI *successful*
or *good*?



Thanks!

Please feel free to
reach out!
michael.pope.4@bc.edu

Survey



<https://tinyurl.com/CS288F22>