Ethical Reasoning in AI for Social Impact

Samuel Dishaw

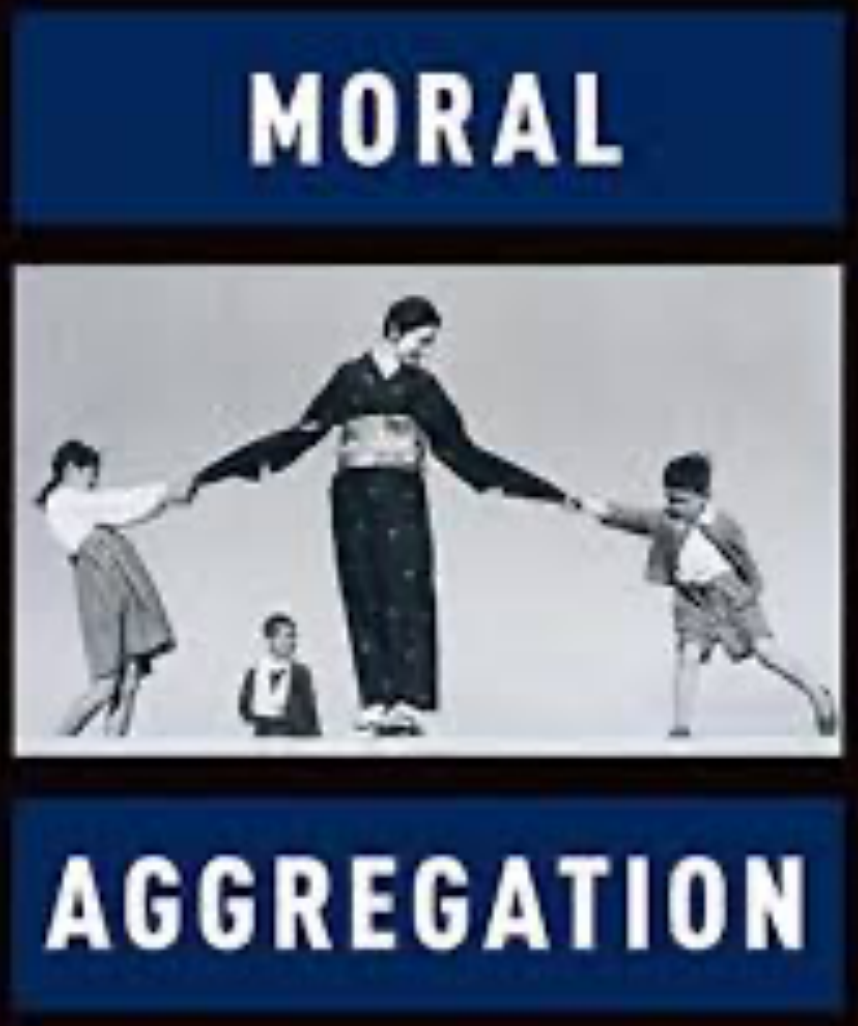
PhD candidate, Philosophy

Agenda

1) Two Moral Methods
2) Case #1: Social network based substance abuse preventions

3) Case #2: Fare Evasion
4) ‘Fair Play’ and Justice

5) Ethical reasoning from data to deployment



Two Moral Methods

- Basic issue: limited resources (for example in healthcare).
- How do we choose between different ways of allocating limited resources?

Two Moral Methods

Scarcity of resources in healthcare:

“Imagine that a new medical technology is made available. This technology enables us to use a new therapy for a rare but serious disease. The downside of this thereby is that it is very expensive. In order to fund this expensive therapy, it is necessary to cut the funding for an existing inexpensive treatment for a relatively minor disease, which a large number of patients require each year” (Hirose 2015)

Which therapy should we fund?

Method #1: Aggregation

- Aggregation: we should fund the therapy that does *the most good*, i.e. that has the greatest total expected benefits across individuals.
- Value of therapy: average benefit * N of individuals who benefit.
- Rationale: if something is valuable, why not maximize it?
- Many public health care systems in the world use aggregation.

Method #2: Pairwise Comparison

- Some philosophers disagree. They think we should opt for the new therapy. We should prioritize helping those individuals whose disease is *most serious*, rather than helping the greater number.

- Aggregating harms additively across persons, they think, is a mistake in "moral mathematics" (Kamm "Moral Reasoning in a Pandemic"). No one person suffers the sum of harms.

Method #2: Pairwise Comparison

- Illustration (classic worry about aggregation):

  We can save one person from death, or some large number of people (N), from a mild headache.

- Is there some numerical value for N at which it is preferable to prevent many headaches than one death?

Method #2: Pairwise Comparison

Method: we identify the individual who is the *worst off* in one distributive outcome. If there is an alternative outcome in which no *individual* is as badly off as they are, the alternative outcome is better.

No aggregating allowed.

Method #2: Pairwise Comparison

If we prevent the headaches, the distributive outcome looks like this:

O_1: (-10,000, 0, 0, 0, 0, 0, 0, 0, 0…)

If we prevent the one death, the distributive outcome looks like this:

O_2: (0, -1, -1, -1, -1, -1, -1, -1, -1…)

Method #2: Pairwise Comparison

If we prevent the headaches, the distributive outcome looks like this:

O_1: (-10,000, 0, 0, 0, 0, 0, 0, 0, 0…)

If we prevent the one death, the distributive outcome looks like this:

O_2: (0, -1, -1, -1, -1, -1, -1, -1, -1…)

Method #2: Pairwise Comparison

Two ideas motivating the Pairwise Comparison method:

i. Separateness of persons

ii. Priority to the worst off

Case #1: Substance Abuse Prevention

Context: networks of social influence to reduce youth substance abuse.

Problem: deviancy training. Reinforcing negative rather than positive behavior.

Goal: use AI to optimize group pairings.

Case #1: Substance Abuse Prevention

How do we determine what group pairing is *optimal*? How do we rank different distributive outcomes?

A natural measure:

Optimal network partition *maximizes the number of expected non-users*, at the end of the intervention (Rahmattalabi et al. 2019)

Two Group Partitions

Suppose we are configuring peer networks for ten individuals, three of which are heavy users.

Network partition A : 6.3 expected non-users at end of intervention.

Network partition B: 4.8 expected non-users at end of intervention.

Two Group Partitions

Group Partition A *isolates* heavy users

{0, 0, 0}, {.9, .9, .9}, {.9, .9, .9, .9}

{} represents peer groups. Orange: heavy users. Blue: light users.

Numbers: probability an individual is a non-user at end of intervention.

Two Group Partitions

Group partition B *distributes* heavy users:

{.2, .5, .5}, {.2, .5, .5}, {.3, .6, .6, .6}

One way to think about the contrast between A and B is that they allocate the resource of *positive influence* differently.

Two Group Partitions

Isolate heavy users

{0, 0, 0}, {.9, .9, .9}, {.9, .9, .9, .9}

Expected non-users: 6.3

Distribute heavy users

{.3, .5, .5}, {.3, .5, .5}, {.4, .6, .6, .6}

Expected non-users: 4.8

Two Group Partitions

Discussion questions:

a) What do aggregation/pairwise comparison deem best?
b) Which group partition do *you* think is better, and why?
c) What do you think our algorithm should *optimize for*?

Two Group Partitions

Isolate heavy users

{0, 0, 0}, {.9, .9, .9}, {.9, .9, .9, .9}

Expected non-users: 6.3

Distribute heavy users

{.3, .5, .5}, {.3, .5, .5}, {.4, .6, .6, .6}

Expected non-users: 4.8

Recap

When our aim is to help a set of individuals, questions arise as how to allocate limited resources.

When we have to make trade-offs between different people's interests, these questions are hard.

Which allocation maximizes aggregate benefits is one important consideration, but it's not the only one.

Case #2: AI against Fare Evasion

- Problem: fare evasion on public transit in Los Angeles. Costs hard to quantify, but conservative estimate is $2.6 million yearly.

- Goal: use AI to optimize schedule of security officers, thereby reducing fare evasion (Delle Fave et al. 2014).

The Duty of ‘Fair Play’

John Rawls on the duty of fair play (1964):

“Suppose there is a mutually beneficial and just scheme of social cooperation. Suppose further that cooperation requires a certain sacrifice […] Under these conditions a person who has accepted the benefits of the scheme is bound by a duty of *fair play* to do his part and not to take advantage of the free benefits by not cooperating.” (Rawls 1964)

The Duty of Fair Play

- The ‘mutually beneficial scheme’: public transit.
- ‘Cooperation’: paying your fare.
- ‘Free’ benefits: using public transit without paying.
- Duty of fair play says: don’t free ride!

Discussion

Is it ever OK to ‘free ride’? Are there some instances in which you think that would be OK?

Is riding without paying, in general, unfair to other users?

Is reducing fare evasion a goal that we should pursue?

Shelby's critique of Rawls

Shelby's point: whether someone has duties of 'fair play' towards members of their society depends on how that society treats them.

In a deeply unjust society (vast inequalities in lifetime prospects), the disadvantaged have no duties of fair play (i.e. of cooperation).

Shelby's claim about the United States: sufficiently unjust that many individuals do not have a duty to cooperate just for the sake of fairness.

Shelby's critique: Back to Fare Evasion

Implication of Shelby's critique: many people have no duty to pay their transit fare, *even if they can 'afford' to pay*.

If someone has no duty to cooperate, it's unfair to prosecute them for free riding.

So: reducing fare evasion may exacerbate rather than reduce unfairness.

Objection

What about *opportunists*: people who are well-off, but don't pay their fare (because they calculate that, given the risk of getting caught + cost of fine ($75), fare evading is cheaper on the long run).

Trade-off: allowing opportunistic free riding vs. wrongful prosecution.

Pairwise comparison can help here: the harm of wrongful prosecution is *much* greater at the *individual* level than the harm of opportunistic fare evasion, which are spread out across society at large.

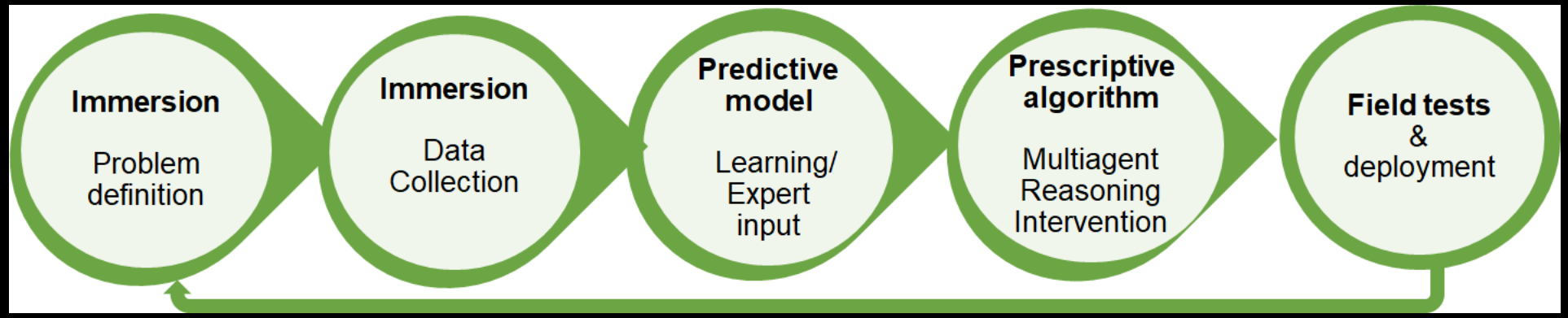
Recap

Lots of goals that are otherwise good goals can be problematic to pursue against morally bad background.

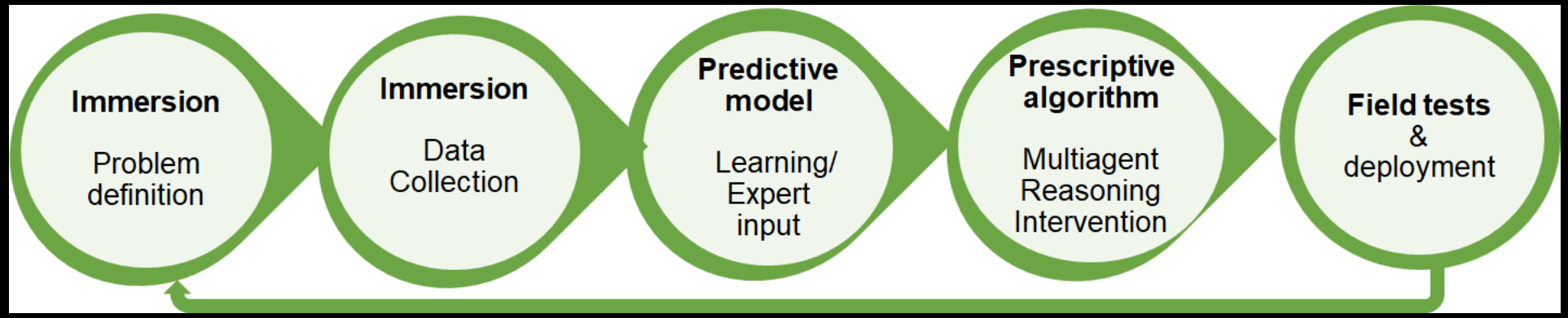
Reducing fare evasion looks like this.

There is nothing intrinsically wrong with reducing fare evasion. But against a background of injustice, it can do more harm than good.

Taking a Step Back: Ethical Reasoning from Data to Deployment



Taking a Step Back: Ethical Reasoning from Data to Deployment



Question: in the two cases we've looked at, at what *stage* did ethical reasoning come into play?

Are there other stages than those we've considered at which you think ethical reasoning could also be important?

Summary

What we've achieved today:

i. We've distinguished two moral methods for allocating resources
ii. We applied these methods to an optimization problem
iii. We discussed the relationship between AI for social impact projects and fairness in society at large
iv. We looked at ethical reasoning in the data-to-deployment pipeline

One Last Request

Please fill out a brief survey about this module!

http://bit.ly/21SCS288