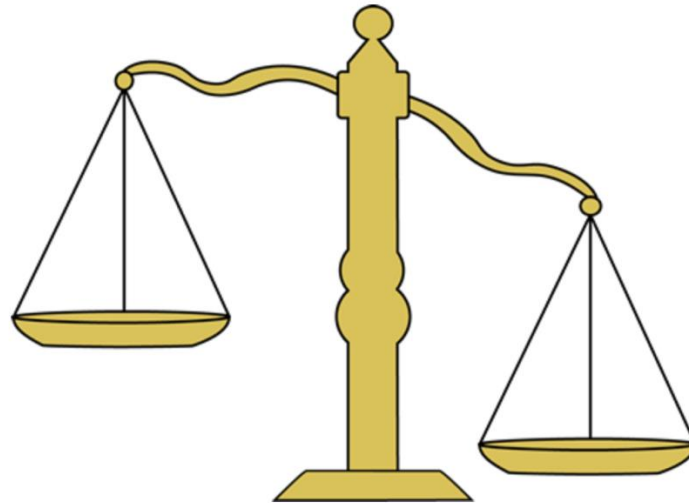


Lecture 14.5

Making Value Judgments

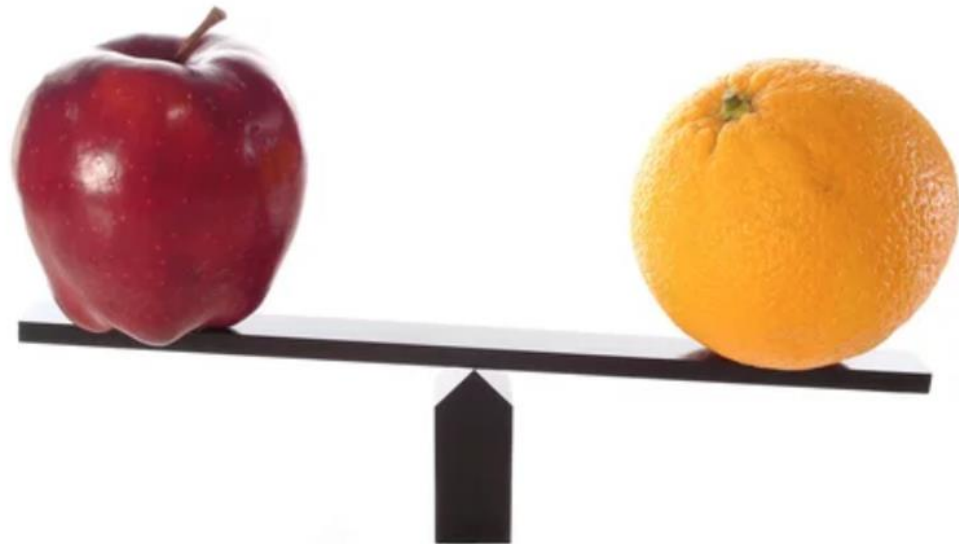


Dan Webber, PhD

Recap: What was our first EthICS lecture about?

Incommensurability!

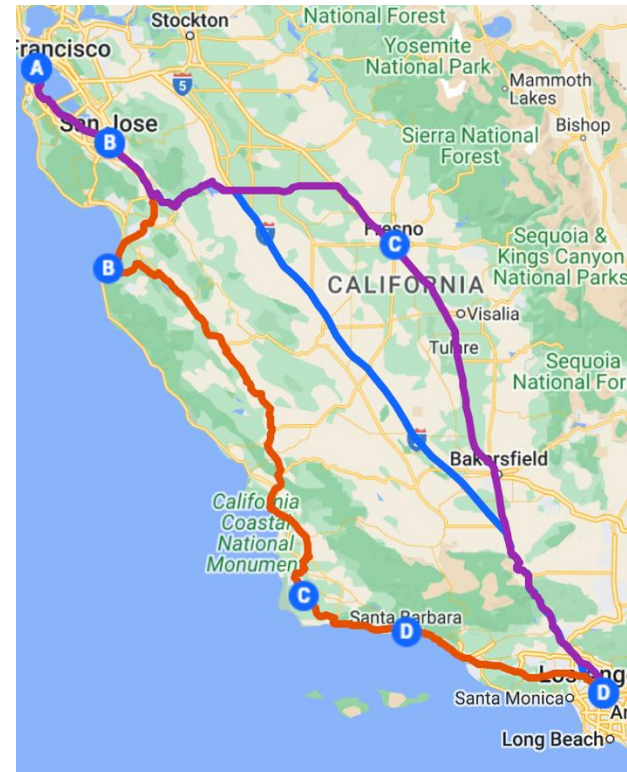
... in the context of sorting!



We looked at problems like this:

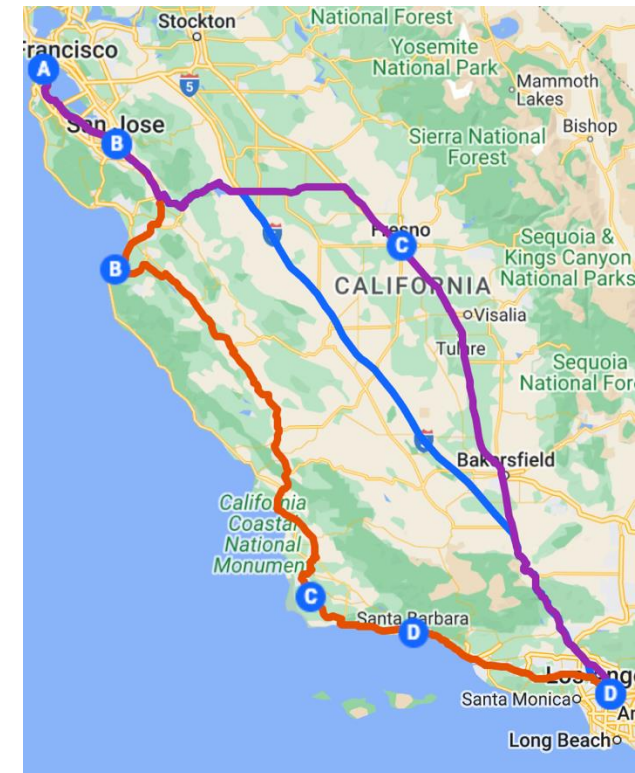
How do we sort when each item to be sorted has multiple values attached to it?

Route Name	Time SF -> LA	Emissions reduction
Coastal	180	2
I-5 Express	135	1.5
Eastern Valley	160	2
...
...



We had an idea: sorting with a weight function

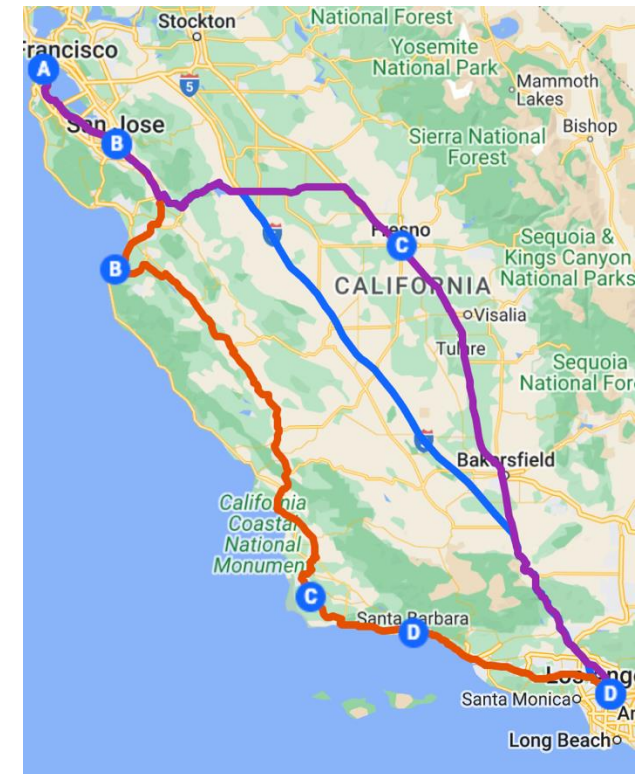
Route Name	Time SF -> LA	Emissions reduction	$f(t, e)$
Coastal	180	2	
I-5 Express	135	1.5	
Eastern Valley	160	2	
...	
...	



... but that still left us with a problem

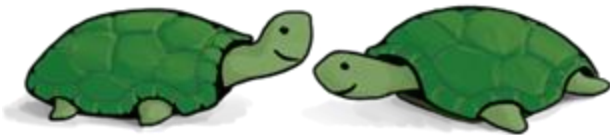
Which function should we use for f ?

Route Name	Time SF -> LA	Emissions reduction	$f(t, e)$
Coastal	180	2	???
I-5 Express	135	1.5	???
Eastern Valley	160	2	???
...
...



Where we ended up

- When incommensurable values are afoot, technical considerations alone won't determine the correct weight function
- Weighing or trading off between different values requires a *value judgment*
- Unanswered question: How do we make good value judgments?



Today

- Value judgments in algorithms: examples
- Philosophical theories for weighing values
 - Pluralism
 - Utilitarianism
 - Prioritarianism
 - Deontology and rights
- Throughout:
 - How would these theories help us make the kind of value judgments our algorithms rely on?
 - What can these theories *not* tell us?



Warning

My hope is that this lecture leaves you better equipped to think about value judgments. But...

- You should *not* expect to leave with a simple, precise algorithm for making value judgments
 - I can't give you this because I don't have it myself! It's an open philosophical question whether there is any such algorithm and, if so, what it is
 - Just because it's an open question doesn't mean it has no answer or that it's not worth thinking about!
Compare: It's an open question whether $P = NP$



Warning, cont.

My hope is that this lecture leaves you better equipped to think about value judgments. But...

- Some of you, I hope, will leave this lecture *less* confident in how to make value judgments than you were before!
- Informed vs. uninformed uncertainty
 - It's better to be uncertain because you're aware of hard questions and nuance than to be uncertain because you're totally in the dark
 - It's also better to be informed and uncertain than uninformed and confident!

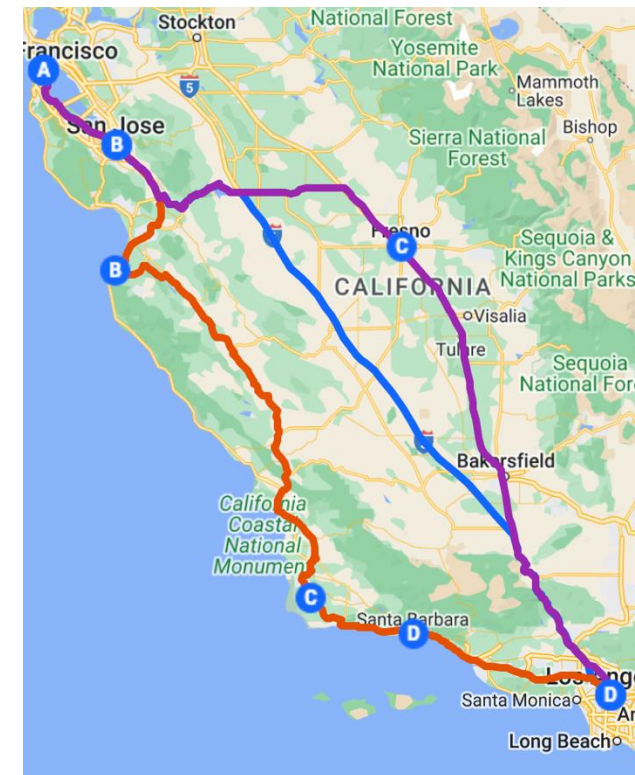
Value judgments in algorithms



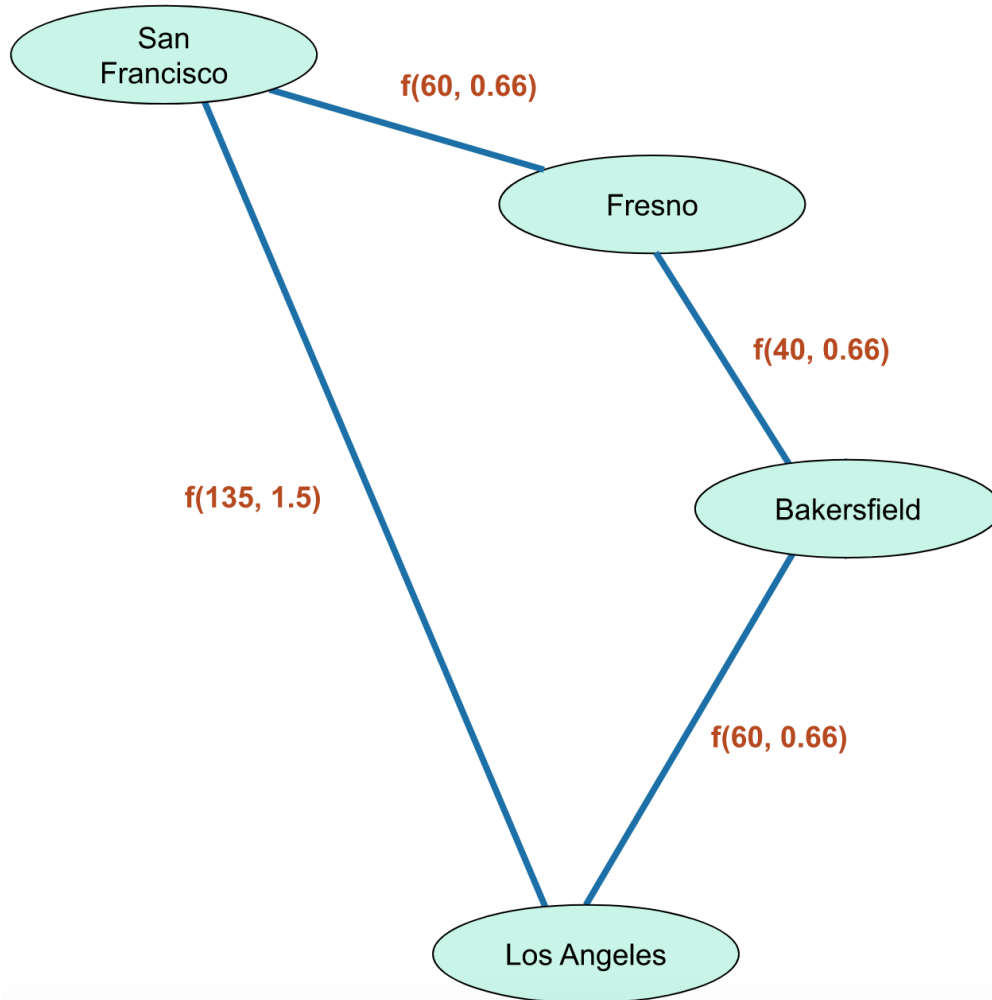
Last time, we found difficult value judgments in a sorting example

Which function should we use for f ?

Route Name	Time SF -> LA	Emissions reduction	$f(t, e)$
Coastal	180	2	???
I-5 Express	135	1.5	???
Eastern Valley	160	2	???
...
...



But many kinds of algorithms presuppose value judgments



For example, we would have faced the same issue if we had framed our high-speed rail problem as a shortest path problem

Another example



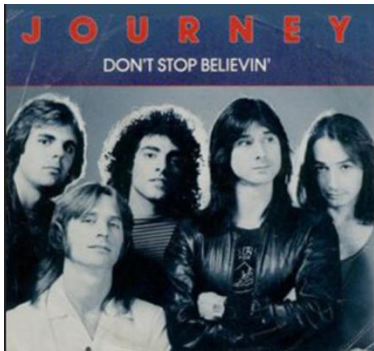
Suppose you're crafting a playlist of classic rock singalong anthems for a drive up to San Francisco.

The drive is 45 minutes long, but there's way more than 45 min of jams you'd like to include.

Does this sound like a problem you've encountered in this class?

It's a knapsack problem!

The playlist (knapsack) can fit 45 minutes worth of tunes



Length (weight):
4:10

Value: ???



Length (weight):
3:23

Value: ???



Length (weight):
8:36

Value: ???



Length (weight):
5:55

Value: ???

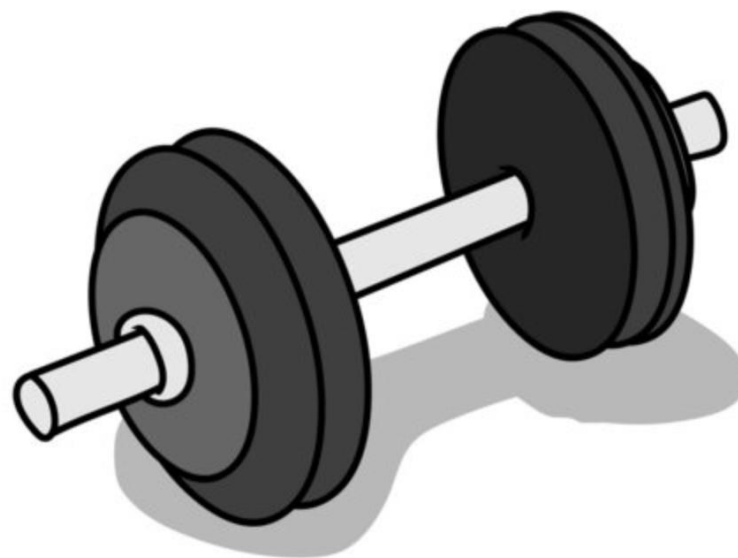
...

How much value does “Sweet Caroline” add to the playlist compared to “Don’t Stop Believin’”?

Lots of algorithms
operate on items that
have a value or weight

In this class, you usually just
take the values or weights for
granted, as arbitrary input into
your algorithm

When solving problems in real
life, sometimes the hardest
part isn't the algorithm—it's
figuring out the correct values
or weights!



Even harder when the problem involves multiple values and people!

Suppose you're responsible for allocating \$N of discretionary government funding. This funding can be used for anything: repairing roads, expanding the hospital system, funding an endowment for the arts, mitigating climate change, etc.

How would you begin to approach this problem?

(Is it an instance of a kind of problem you've seen before?)

Yep, it's another knapsack!



Road repair

Hospital expansion

Arts endowment

Climate mitigation ...



Cost (weight):
\$W (per unit)

Value: ???



Cost (weight):
\$X (per unit)

Value: ???



Cost (weight):
\$Y (per unit)

Value: ???



Cost (weight):
\$Z (per unit)

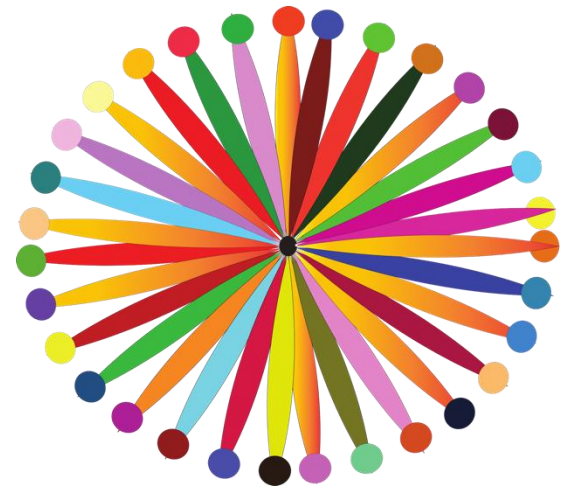
Value: ???

The \$N question: What's the *value* of each of these projects?



Ethical Theory

One Answer: Pluralism



- There are many kinds of value, and no single value in terms of which they're all commensurable.
- There's no common measure of the value of smooth roads, good healthcare, and the arts.
- To assign comparable numerical values to each of these things, we just exercise our best judgment about which are more or less valuable—but not *in terms of* anything else!

But are all these values *really* incommensurable?



Aren't smooth roads valuable because people *prefer* to drive on them?

Isn't good healthcare valuable because people *like* to be healthy?

Isn't art valuable because people *enjoy* it?

Isn't climate change mitigation valuable because it's necessary for us to live *happy lives*?

Utilitarianism

- All valuable goods are commensurable in terms of a single fundamental value: *happiness*
- The value of a thing x is the sum, over all people, of the happiness it produces:

$$v_x = \sum_p h_{p,x}$$

- You should choose the highest value; that is, you should maximize overall happiness
- (What about *unhappiness*?)

Utilitarianism makes it easy to formulate our problem—right?

Road repair



Cost (weight):
\$W (per unit)

Value:

$$\sum_p h_{p, roads}$$

Hospital expansion



Cost (weight):
\$X (per unit)

Value:

$$\sum_p h_{p, hospital}$$

Arts endowment



Cost (weight):
\$Y (per unit)

Value:

$$\sum_p h_{p, arts}$$

Climate mitigation ...



Cost (weight):
\$Z (per unit)

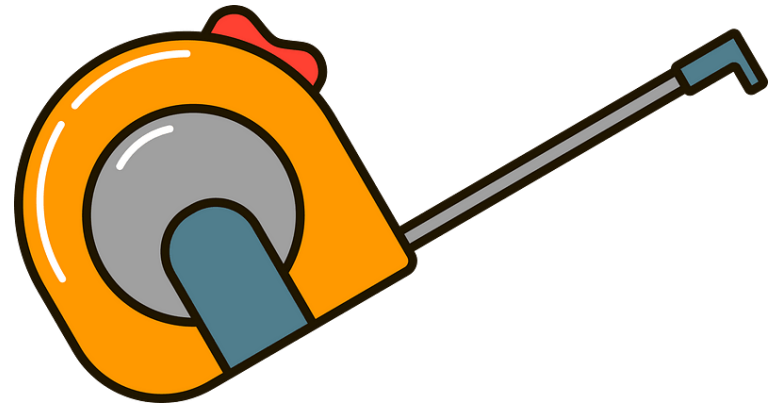
Value:

$$\sum_p h_{p, climate}$$

Okay... but how do we evaluate these sums?

Utilitarianism: measurement problems

- How do we measure happiness?
- Even harder: how do we measure happiness *precisely enough* to allow for commensurability?
- How do we measure the happiness of *future people* when we don't know what the future holds, or when it depends on what we do now?



Case Study: Effective Altruism

- *Effective Altruism* (EA) is a social movement, often inspired by utilitarian thinking, that advocates doing the most good possible with your charitable giving
- In the 2010s, EA consensus was that the way to do this was to donate to efforts to prevent malaria
- Malaria kills lots of people and can be prevented with cheap mosquito nets—lots of bang for the buck, happiness-wise!

Case Study: Effective Altruism

- Recently, many EAs have endorsed *longtermism*: the idea that we can do more good by trying to improve the lives of people in the distant future
- The way to do this, many of them claim, is by investing in safeguards to prevent AI from taking over the world and killing us all

Mosquito Nets



Cost (weight):

\$5n

Value:

$$\sum_p h_{p, nets} = h_{life} * n$$

Stop Killer Robots

Cost (weight):

\$X

Value:

$$\sum_p h_{p, AI} = h_{life} * 10^{58} * c$$



Utilitarianism is just the tip of an iceberg



The top-line view is seductively simple: produce as much happiness as you can

But this simplicity is deceptive: below the surface are hard questions about how happiness can be measured and compared (especially at global scale!)

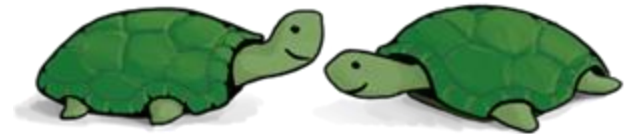
That doesn't mean utilitarianism is false. Morality is complex! But it's not a silver bullet that makes value judgments easy

(Is it really more helpful than pluralism?)

... and is utilitarianism even true?

Whoo boy, *I* don't even know, and trying to answer questions like that one is *literally my job*

Let's try an easier one: Why *might* someone think it's false? That is, what might be wrong with just trying to maximize overall happiness?





Prioritarianism

- Maybe there's more that matters about happiness than *how much* there is—maybe it also matters how it's *distributed*
- The value of a thing x is the *weighted* sum, over all people, of the happiness it produces:

$$v_x = \sum_p (h_{p,x})(w_p)$$

- The worse off you are, the greater your weight—that is, the more value there is in increasing *your* happiness



Prioritarianism

- Prioritarianism doesn't solve the measurement problems of utilitarianism
- And it adds another problem: What's the function from a person's level of utility to their weight in the sum?
- But if we're given some happiness numbers and a weight function, we can contrast the two views

Prioritarianism vs Utilitarianism

Suppose we use a simple step function: people who are more well off than average have a weight of 1, while people less well off have a weight of 2

Road repair



Value: 10 util for each
of all 1mm citizens

Hospital expansion



Value: 1000 util for
each of 9,000 citizens
... who are all less well-
off than average

Value beyond utility

Maybe some things, like the environment or artistic achievement, have value beyond their effect on human happiness



If we think this, we are saying that utilitarianism or prioritarianism is incomplete. We are back to the pluralist view we started with: there are multiple values that we need to weigh against each other

Constraints on pursuing the good

But we might go even further: we might think that value judgments are shaped not just by the values to be weighed, but also by values that can't be weighed against others

Stop Killer Robots

Cost (weight):
\$Y (per unit)

Value:

$$\sum_p h_{p, AI}$$



Stop Killer Robots by Defrauding Customers

Cost (weight):
\$0 (per unit)

Value:

$$\sum_p h_{p, AI} - u_{p, fraud}$$



Constraints on pursuing the good

Road repair



Cost (weight):
\$W (per unit)

Value: ???

Hospital expansion



Cost (weight):
\$X (per unit)

Value: ???

Arts endowment



Cost (weight):
\$Y (per unit)

Value: ???

Promised: all \$N

Climate mitigation ...



Cost (weight):
\$Z (per unit)

Value: ???

Considerations like these don't just weigh against others (like happiness). They seem to *override* the balance of other considerations

Deontology and Rights

- Views that have these sorts of constraints are sometimes called *deontological*
- Deontologists think that there are *moral principles* that shouldn't be violated no matter how much value we could produce by doing so
- Often, these principles are taken to express people's *rights*. Rights limit what we can do to people even in pursuit of valuable goals

Deontology and Rights

Road repair



Cost (weight):
\$W (per unit)

Value: ???

Hospital expansion



Cost (weight):
\$X (per unit)

Value: ???

Arts endowment



Cost (weight):
\$Y (per unit)

Value: ???

Climate mitigation ...



Cost (weight):
\$Z (per unit)

Value: ???

What are some other ways that moral principles or people's rights might play a role in our decision about how to allocate funding?



Today

- Value judgments in algorithms: examples ✓
- Philosophical theories for weighing values
 - Pluralism ✓
 - Utilitarianism ✓
 - Prioritarianism ✓
 - Deontology and rights ✓
- Throughout:
 - How would these theories help us make the kind of value judgments our algorithms rely on? ✓
 - What can these theories *not* tell us? ✓


Satisfied?
I hope not!

I hope you feel a burning need
for answers:

- What's the *right* way to weigh different people's interests in deciding what to do?
- Why *should* we care about moral principles or rights if they stop us from doing more good?
- How can we *know* which moral views to accept?



... but I hope you also feel you learned *something*:

- Lots of algorithms—not just sorting—work on input data that have *values* attached
 - To solve real-world problems algorithmically, you first need to determine these values
 - That can be hard to do, even with the help of a theory like utilitarianism!
 - Measurement problems abound
 - Utilitarianism might be too simple
 - It also seems to matter how goods are *distributed*, and that we respect people's *rights*
- 

Embedded Ethics survey!

- Coming soon, be on the lookout
- Your thoughts on Embedded Ethics in your current courses
- First 800 participants get \$10 gift card
- Whether you choose to participate or not will **not** affect your grade in any way



Want to talk more
about ethics?

Dan Webber
webberdf@stanford.edu

Email to set up a meeting!