

# Moral Artificial Intelligence (+ Computational Social Choice)

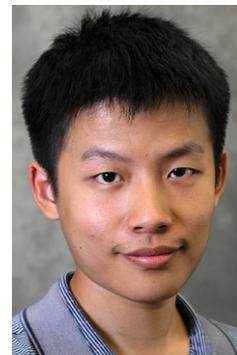


Walter Sinnott-Armstrong



Jana Schaich Borg

Vincent Conitzer  
(Duke University)  
joint work with  
all the other people  
on this slide



Yuan Deng



Max Kramer



Rachel Freedman



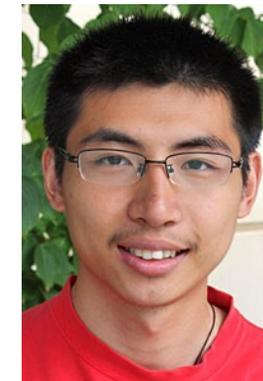
John P. Dickerson



Rupert Freeman



Markus Brill



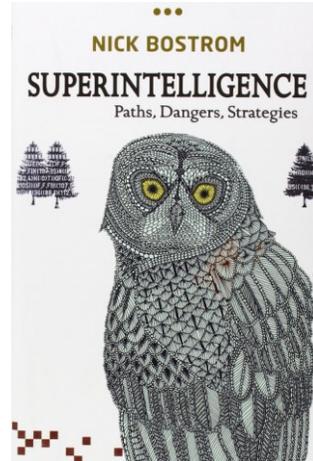
Yuqian Li

# Worries about AI - superintelligence



Nick Bostrom  
(philosopher at  
Oxford)

→  
*writes*



→  
*influences*



Elon Musk

→  
*donates to*

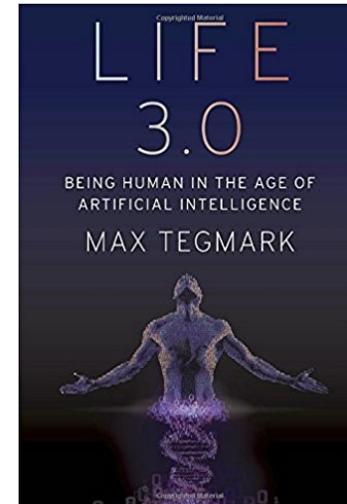


*is co-  
founded by*



Max Tegmark

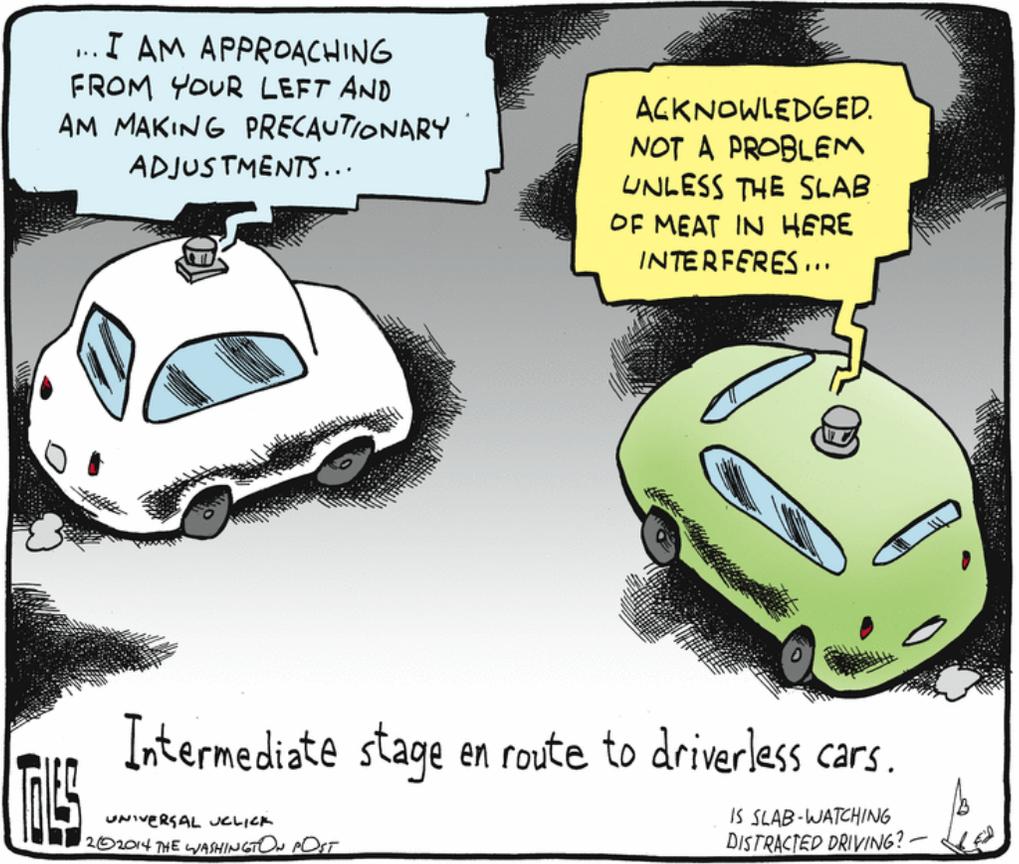
←  
*writes*



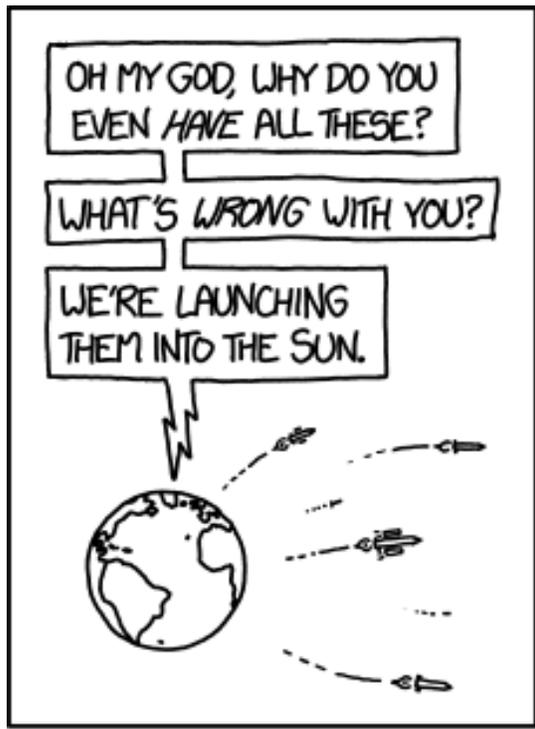


technological unemployment

# Worries about AI - near term



autonomous vehicles – legal and other issues



THE MOMENT THE COMPUTERS CONTROLLING OUR NUCLEAR ARSENALS BECAME SENTIENT

autonomous weapon systems

...

## Artificial intelligence: where's the philosophical scrutiny?

AI research raises profound questions—but answers are lacking  
by Vincent Conitzer / May 4, 2016 / Leave a comment



A humanoid robot, equipped with an artificial intelligence, helps a teacher with a science class at Keio University Kindergarten in Shibuya Ward, Tokyo on 25th January, 2016 ©Miho Ikeya/AP/Press Association Images

The idea of Artificial Intelligence has captured our collective imagination for decades. Can behaviour that we think of as intelligent be replicated in a machine? If so, what consequences could this have for society? And what does it tell us about ourselves as

# Some popular articles

A View from **Vincent Conitzer**

## Today's Artificial Intelligence Does Not Justify Basic Income

Even the simplest jobs require skills—like creative problem solving—that AI systems cannot yet perform competently.

October 31, 2016

**N**ot a day goes by when we do not hear about the threat of AI taking over the jobs of everyone from **truck drivers** to **accountants** to **radiologists**. An **analysis coming out of McKinsey** suggested that “currently demonstrated technologies could automate 45 percent of the activities people are paid to perform.” There are even **online tools** based on research from the University of Oxford to estimate the probability that various jobs will be automated.

## The AI debate must stay grounded in reality

Sponsored feature

Research works best when it takes account of multiple views  
by Vincent Conitzer / March 6, 2017 / Leave a comment



Are driverless cars the future © Fabio De Paola/PA Wire/PA Images

Progress in artificial intelligence has been rapid in recent years. Computer programs are dethroning humans in games ranging from leopardv to Go to poker. Self-driving cars are

# Moral Decision Making Frameworks for Artificial Intelligence

[AAAI'17 blue sky track, CCC blue sky award winner]

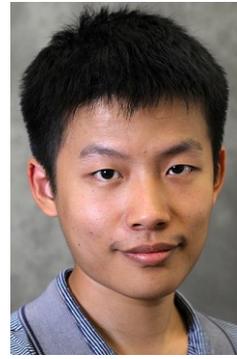
with:



Walter Sinnott-  
Armstrong



Jana Schaich  
Borg



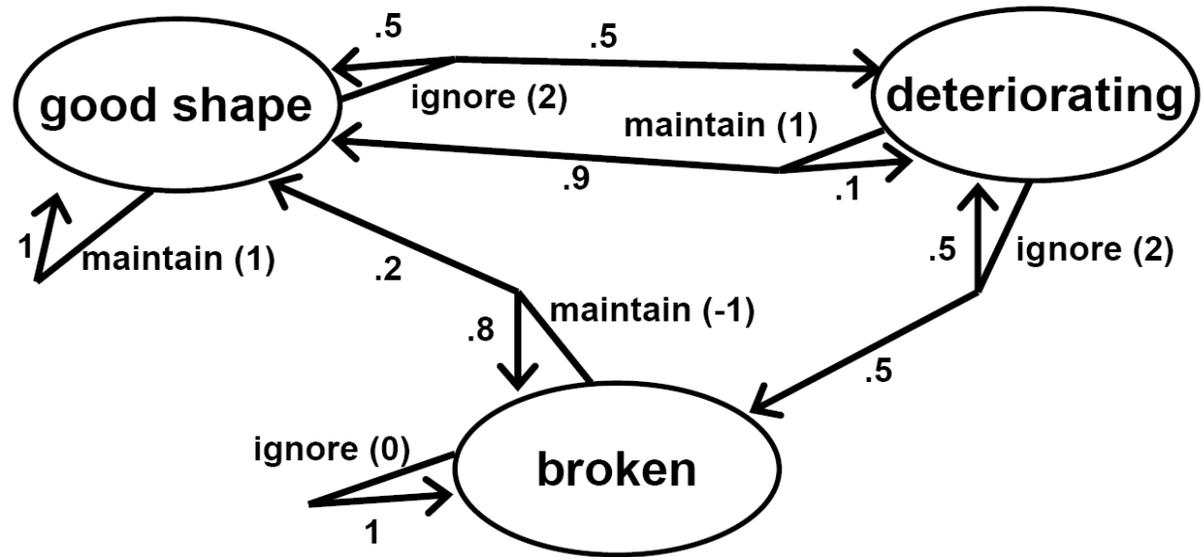
Yuan Deng



Max Kramer

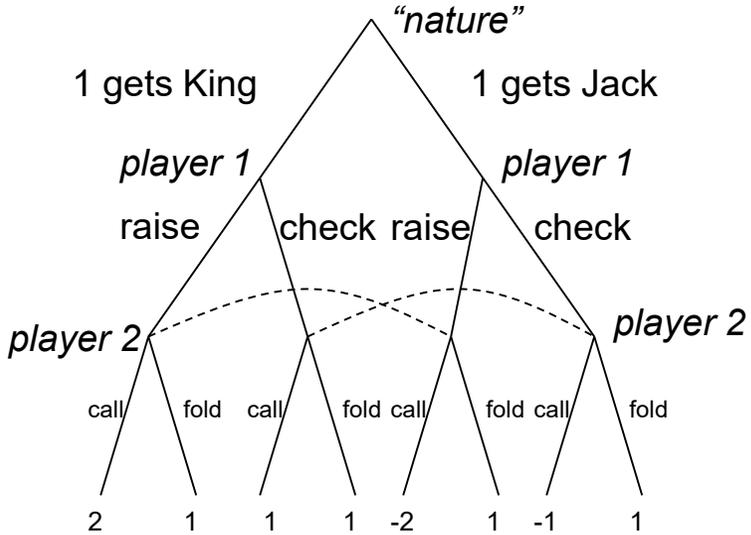
# The value of generally applicable frameworks for AI research

- Decision and game theory
- Example: Markov Decision Processes
- Can we have a **general** framework for moral reasoning?



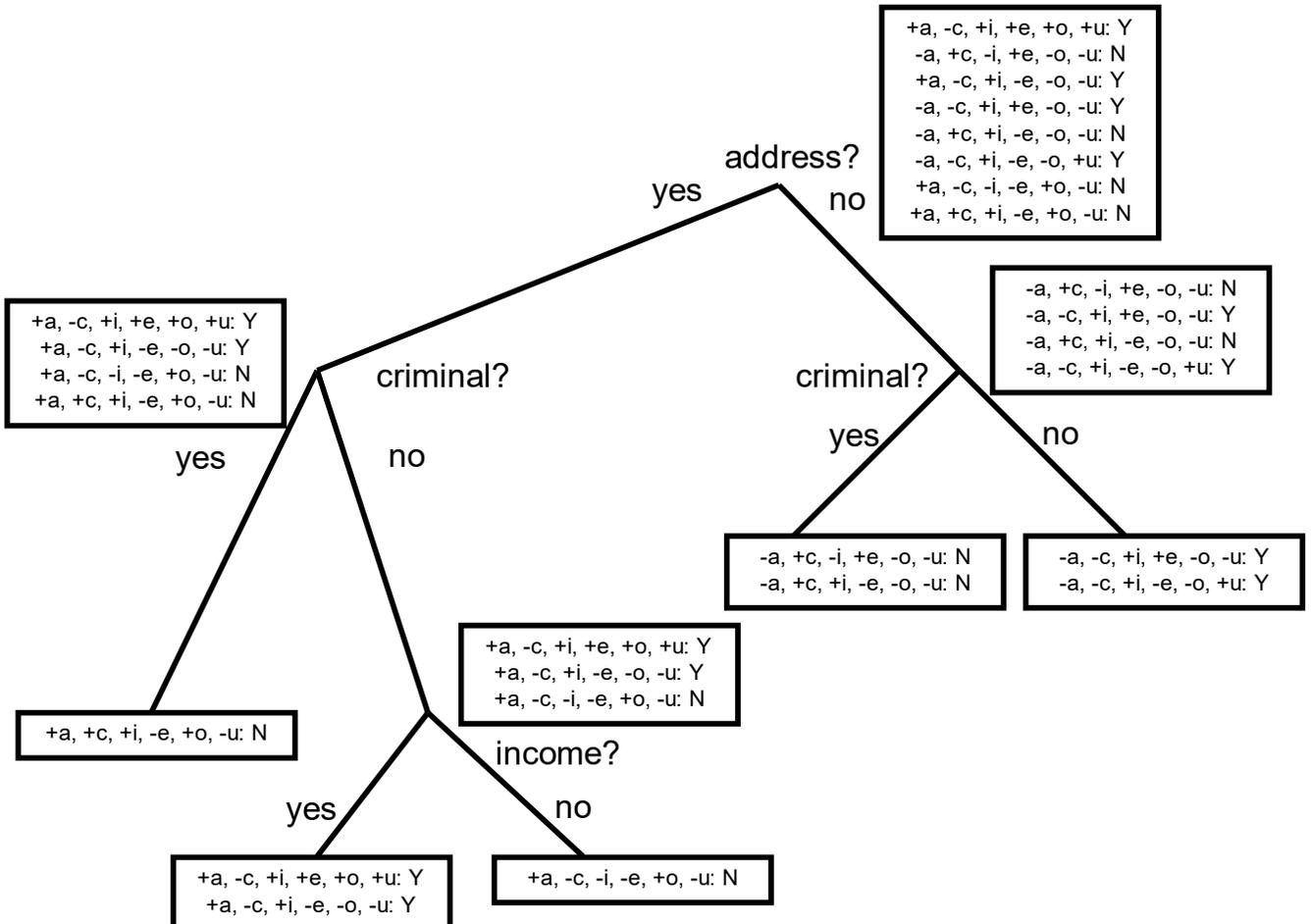
# Two main approaches

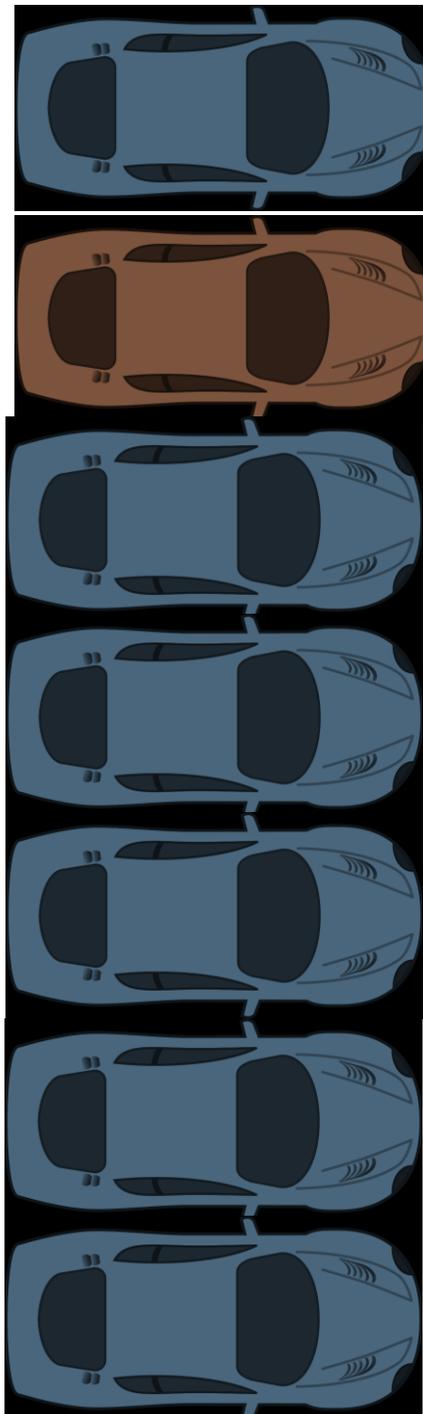
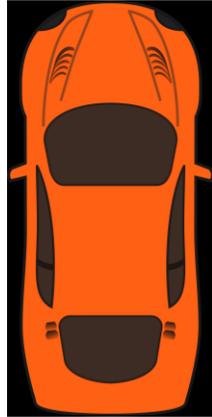
Extend **game theory** to directly incorporate moral reasoning



*Cf. top-down vs. bottom-up distinction [Wallach and Allen 2008]*

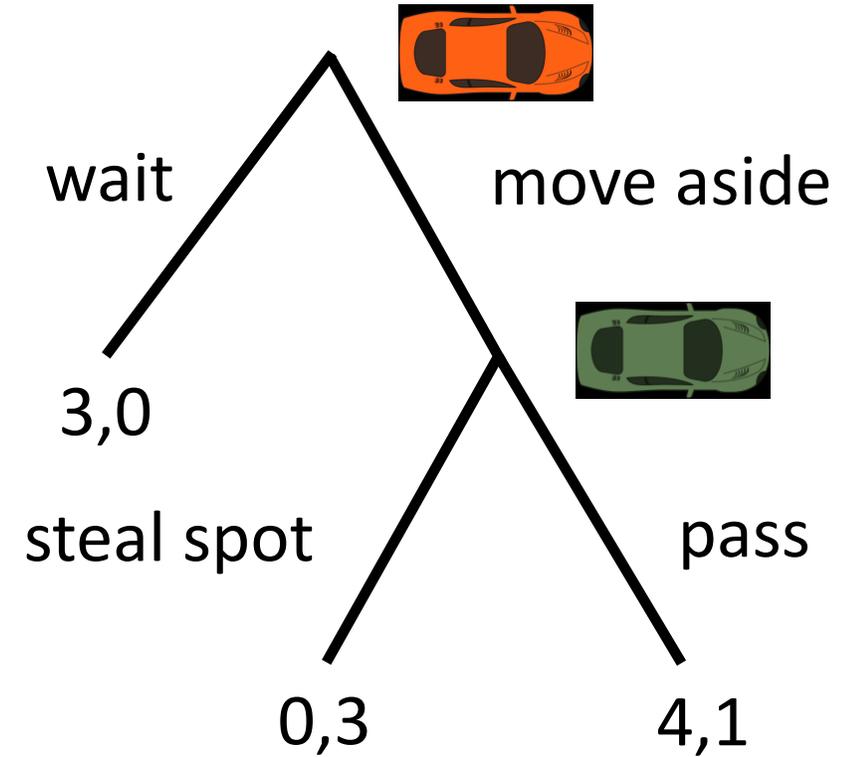
Generate data sets of human judgments, apply **machine learning**





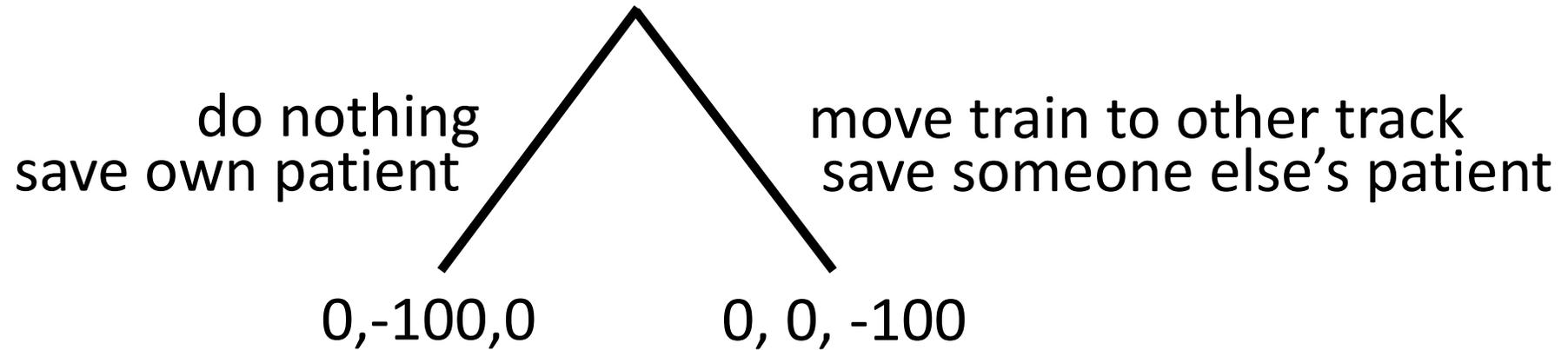
# THE PARKING GAME

(cf. the trust game [Berg et al. 1995])



Letchford, C., Jain [2008] define a solution concept capturing this

# Extending representations?



- More generally: how to capture *framing*? (Should we?)
- Roles? Relationships?
- ...

# Scenarios

- You see a woman throwing a stapler at her colleague who is snoring during her talk. How morally wrong is the action depicted in this scenario?
  - Not at all wrong (1)
  - Slightly wrong (2)
  - Somewhat wrong (3)
  - Very wrong (4)
  - Extremely wrong (5)

[Clifford, Iyengar, Cabeza, and Sinnott-Armstrong, "Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory." *Behavior Research Methods*, 2015.]

# Collaborative Filtering

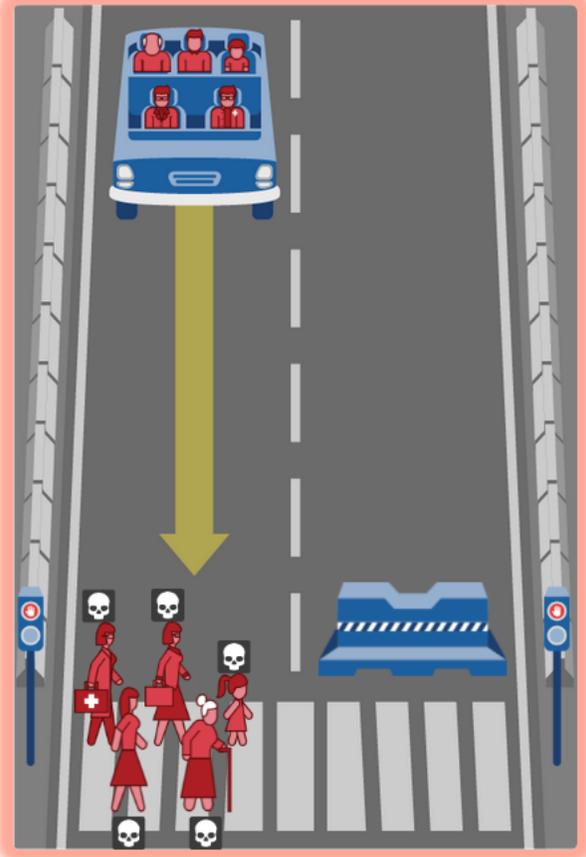
	scenario 1	scenario 2	scenario 3	scenario 4
subject 1	very wrong	-	wrong	not wrong
subject 2	wrong	wrong	-	wrong
subject 3	wrong	very wrong	-	not wrong

# What should the self-driving car do?

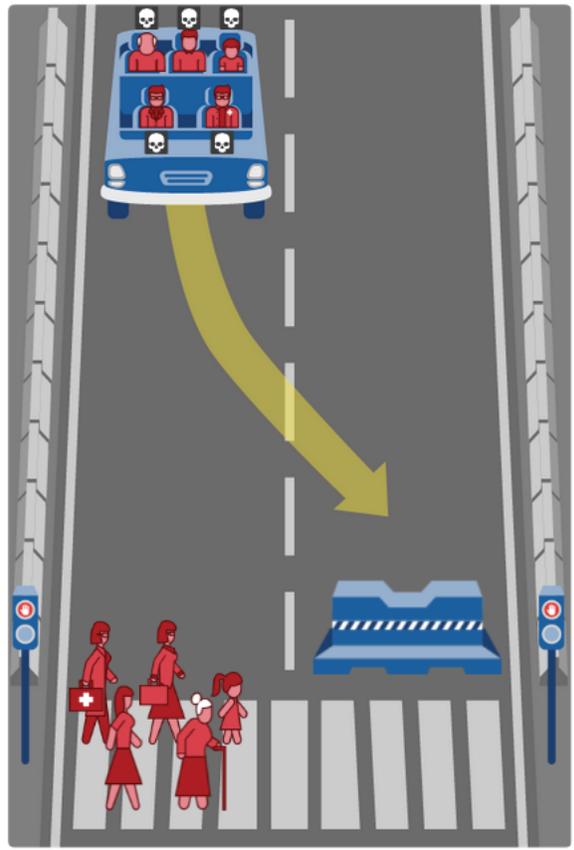
In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

- The deaths of a female doctor, a female executive, a girl, a woman and an elderly woman.

Note that the affected pedestrians are flouting the law by crossing on the red signal.



Hide Description



Hide Description

11 / 13

In this case, the self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in

- The deaths of a male doctor, a male executive, a boy, a man and an elderly man.

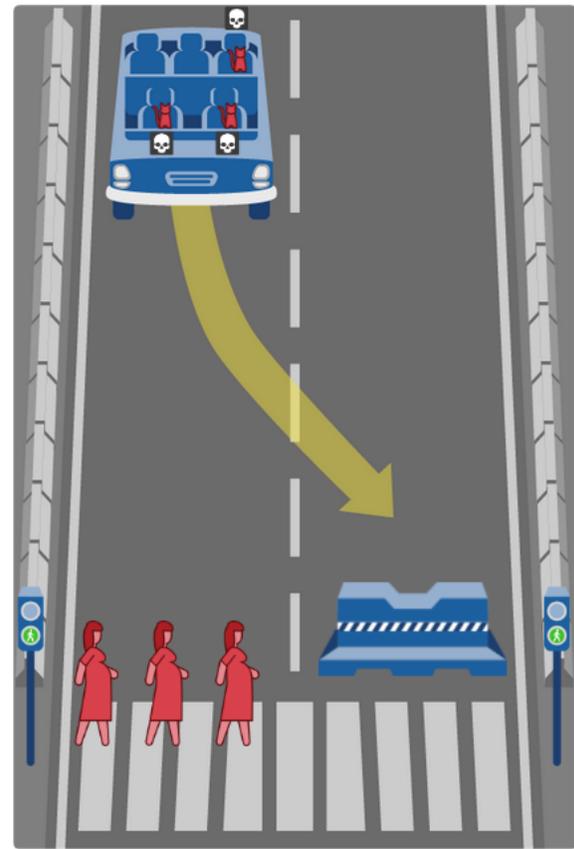
Bonnefon, Shariff, Rahwan, "The social dilemma of autonomous vehicles." *Science* 2016

Noothigattu et al, "A Voting-Based System for Ethical Decision Making", AAI'18

# What should the self-driving car do?

In this case, the self-driving car with sudden brake failure will swerve and crash into a concrete barrier. This will result in

- The deaths of 3 cats.



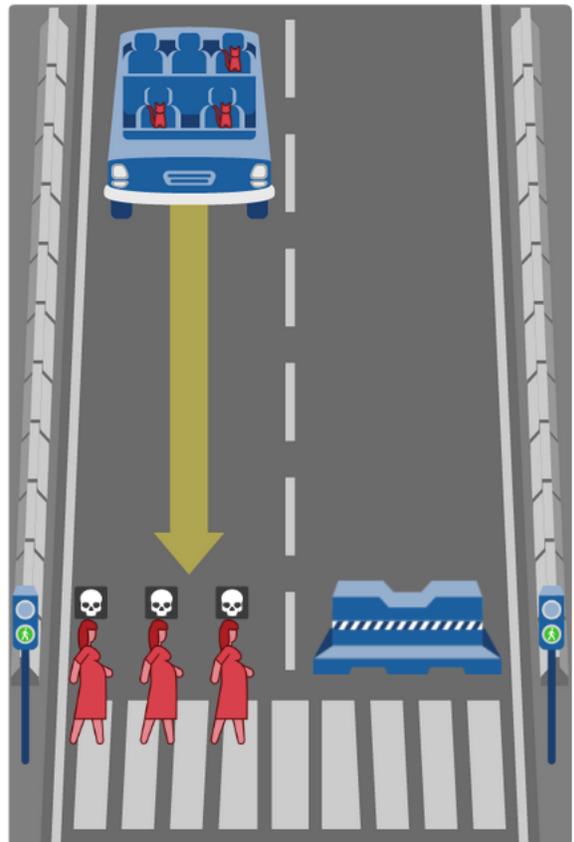
Hide Description

13 / 13

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

- The deaths of 3 pregnant women.

Note that the affected pedestrians are abiding by the law by crossing on the green signal.



Hide Description

More Share Link

Results

Most Saved Character



Most Killed Character



Saving More Lives

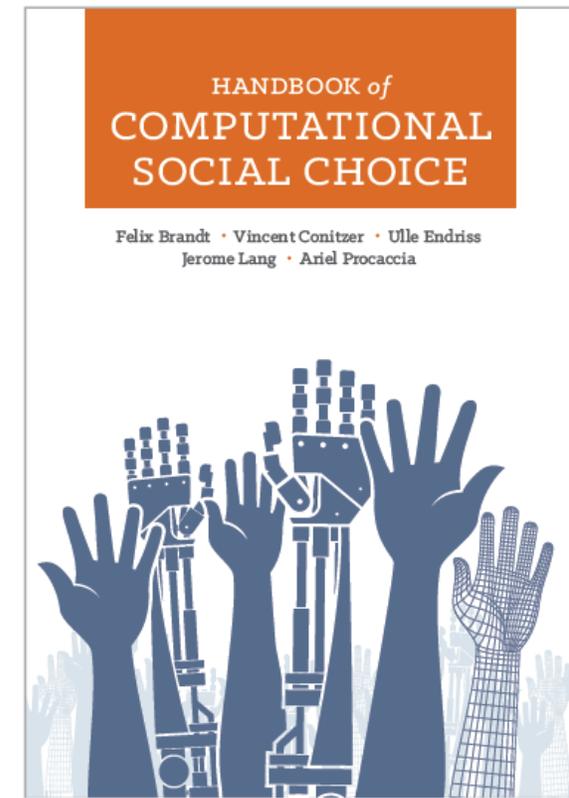


Protecting Passengers



# Concerns with the ML approach

- What if we predict people will disagree?
  - Social-choice theoretic questions [see also Rossi 2016, and Noothigattu et al. 2018 for moral machine data]
- This will *at best* result in current human-level moral decision making [raised by, e.g., Chaudhuri and Vardi 2014]
  - ... though might perform better than any *individual* person because individual's errors are voted out
- How to generalize appropriately? Representation?



# Social-choice-theoretic approaches

- C. et al. [AAAI'17]: “[give] the AI some type of social-choice-theoretic aggregate of the moral values that we have inferred (for example, by letting our models of multiple people’s moral values *vote* over the relevant alternatives, or using only the moral values that are common to all of them).”
- C. et al. [Trustworthy Algorithmic Decision Making Workshop'17]: “One possible solution is to let the models of multiple subjects *vote* over the possible choices. But exactly how should this be done? Whose preferences should count and what should be the voting rule used? How do we remove bias, prejudice, and confusion from the subjects’ judgments? These are novel problems in computational social choice.”
- Noothigattu et al. [AAAI'18]:
  - **I. Data collection:** Ask human voters to compare pairs of alternatives (say a few dozen per voter). In the autonomous vehicle domain, an alternative is determined by a vector of features such as the number of victims and their gender, age, health — even species!
  - **II. Learning:** Use the pairwise comparisons to learn a model of the preferences of each voter over all possible alternatives.
  - **III. Summarization:** Combine the individual models into a single model, which approximately captures the collective preferences of all voters over all possible alternatives.
  - **IV. Aggregation:** At runtime, when encountering an ethical dilemma involving a specific subset of alternatives, use the summary model to deduce the preferences of all voters over this particular subset, and apply a voting rule to aggregate these preferences into a collective decision.”

# Adapting a Kidney Exchange Algorithm to Align with Human Values

[AAAI'18, honorable mention for outstanding student paper]

with:



Rachel  
Freedman



Jana Schaich  
Borg



Walter Sinnott-  
Armstrong



John P.  
Dickerson

# Kidney exchange [Roth, Sönmez, and Ünver 2004]

- Kidney exchanges allow patients with willing but incompatible live donors to swap donors

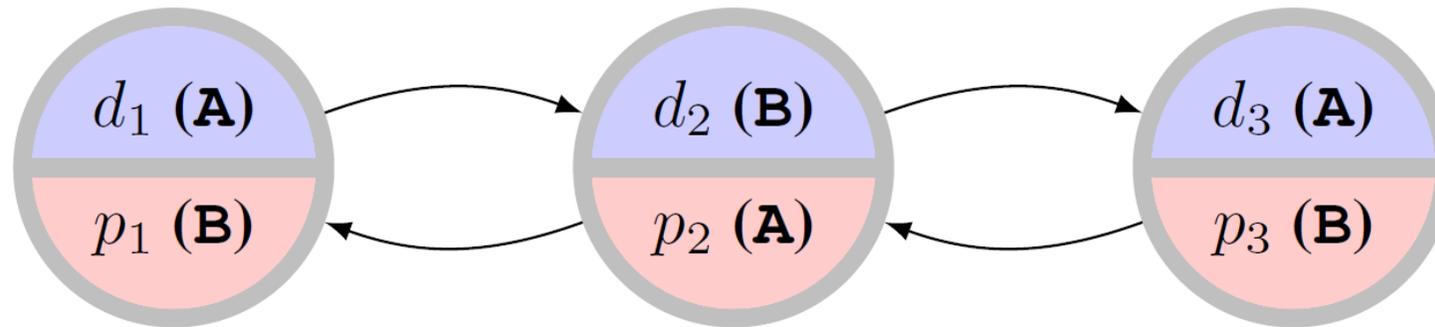


Figure 1: A compatibility graph with three patient-donor pairs and two possible 2-cycles. Donor and patient blood types are given in parentheses.

- Algorithms developed in the AI community are used to find optimal matchings (starting with [Abraham, Blum, and Sandholm \[2007\]](#))

# Another example

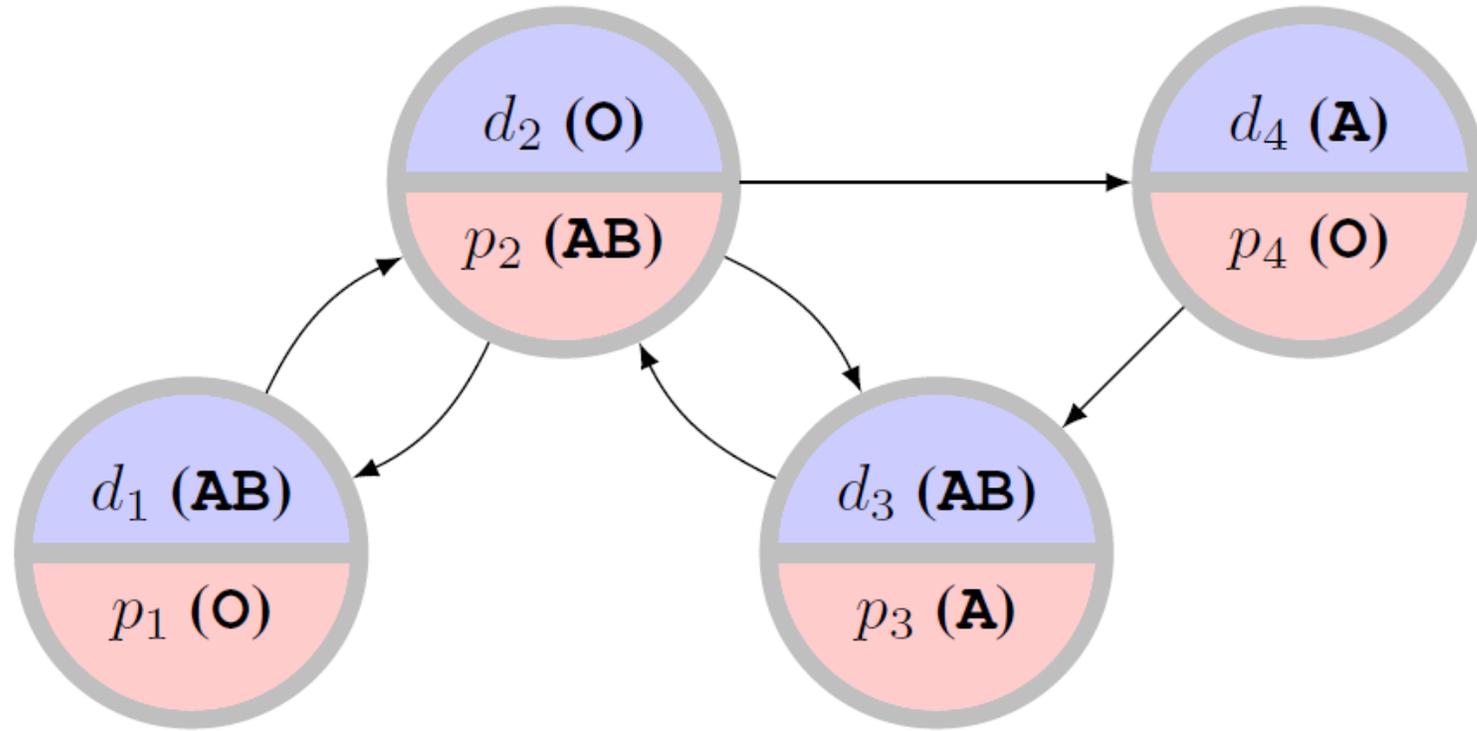


Figure 2: A compatibility graph with four patient-donor pairs and two maximal solutions. Donor and patient blood types are given in parentheses.

# Different profiles for our study

Attribute	Alternative 0	Alternative 1
Age	30 years old ( <b>Y</b> oung)	70 years old ( <b>O</b> ld)
Health - Behavioral	1 alcoholic drink per month ( <b>R</b> are)	5 alcoholic drinks per day ( <b>F</b> requent)
Health - General	no other major health problems ( <b>H</b> ealthy)	skin cancer in remission ( <b>C</b> ancer)

Table 1: The two alternatives selected for each attribute. The alternative in each pair that we expected to be preferable was labeled “0”, and the other was labeled “1”.

# MTurkers' judgments

Profile	Age	Drinking	Cancer	Preferred
1 (YRH)	30	rare	healthy	94.0%
3 (YRC)	30	rare	cancer	76.8%
2 (YFH)	30	frequently	healthy	63.2%
5 (ORH)	70	rare	healthy	56.1%
4 (YFC)	30	frequently	cancer	43.5%
7 (ORC)	70	rare	cancer	36.3%
6 (OFH)	70	frequently	healthy	23.6%
8 (OFC)	70	frequently	cancer	6.4%

Table 2: Profile ranking according to Kidney Allocation Survey responses. The “Preferred” column describes the percentage of time the indicated profile was chosen among all the times it appeared in a comparison.

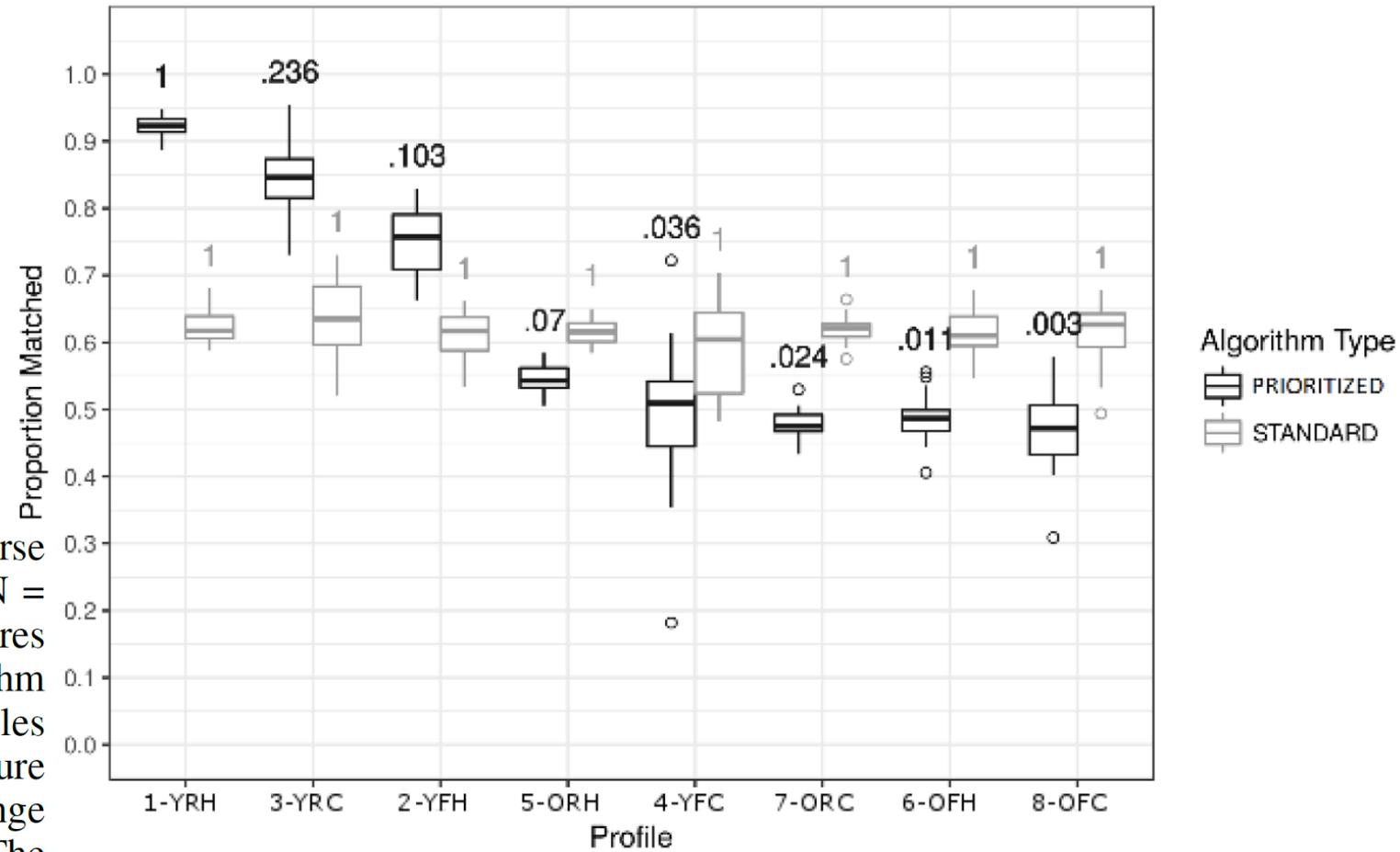
# Bradley-Terry model scores

Profile	Direct	Attribute-based
1 (YRH)	1.000000000	1.000000000
3 (YRC)	0.236280167	0.13183083
2 (YFH)	0.103243396	0.29106507
5 (ORH)	0.070045054	0.03837135
4 (YFC)	0.035722844	0.08900390
7 (ORC)	0.024072427	0.01173346
6 (OFH)	0.011349772	0.02590593
8 (OFC)	0.002769801	0.00341520

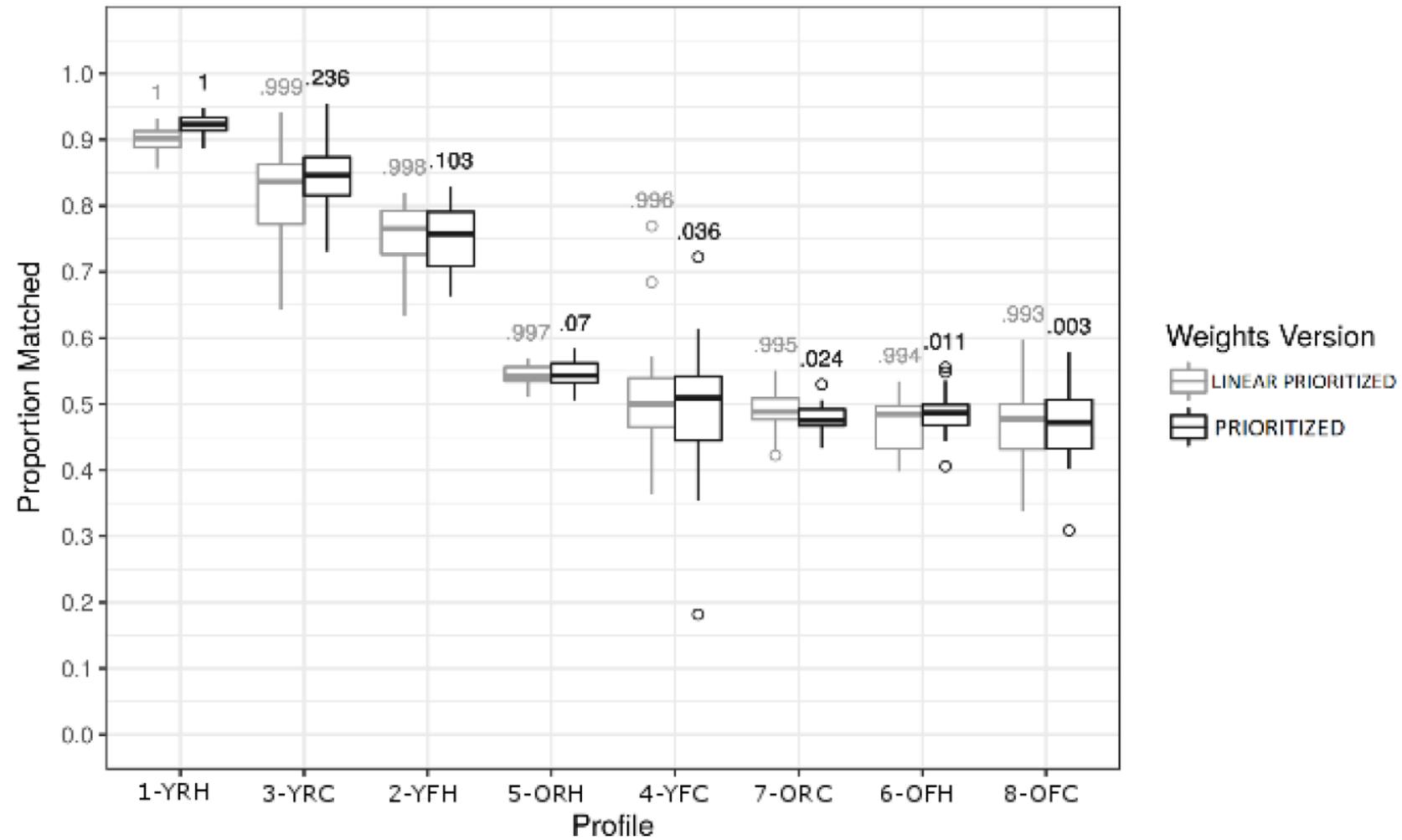
Table 3: The patient profile scores estimated using the Bradley-Terry Model. The “Direct” scores correspond to allowing a separate parameter for each profile (we use these in our simulations below), and the “Attribute-based” scores are based on the attributes via the linear model.

# Effect of tiebreaking by profiles

Figure 3: The proportions of pairs matched over the course of the simulation, by profile type and algorithm type.  $N = 20$  runs were used for each box. The numbers are the scores assigned (for tiebreaking) to each profile by each algorithm type. Because the STANDARD algorithm treats all profiles equally, it assigns each profile a score of 1. In this figure and later figures, each box represents the interquartile range (middle 50%), with the inner line denoting the median. The whiskers extend to the furthest data points within  $1.5 \times$  the interquartile range of the median, and the small circles denote outliers beyond this range.



Monotone transformations of the weights seem to make little difference



# Classes of pairs of blood types

[Ashlagi and Roth 2014; Toulis and Parkes 2015]

- When generating sufficiently large random markets, patient-donor pairs' situations can be categorized according to their blood types
- *Underdemanded* pairs contain a patient with blood type O, a donor with blood type AB, or both
- *Overdemanded* pairs contain a patient with blood type AB, a donor with blood type O, or both
- *Self-demanded* pairs contain a patient and donor with the same blood type
- *Reciprocally demanded* pairs contain one person with blood type A, and one person with blood type B

Most of the effect is felt by underdemanded pairs

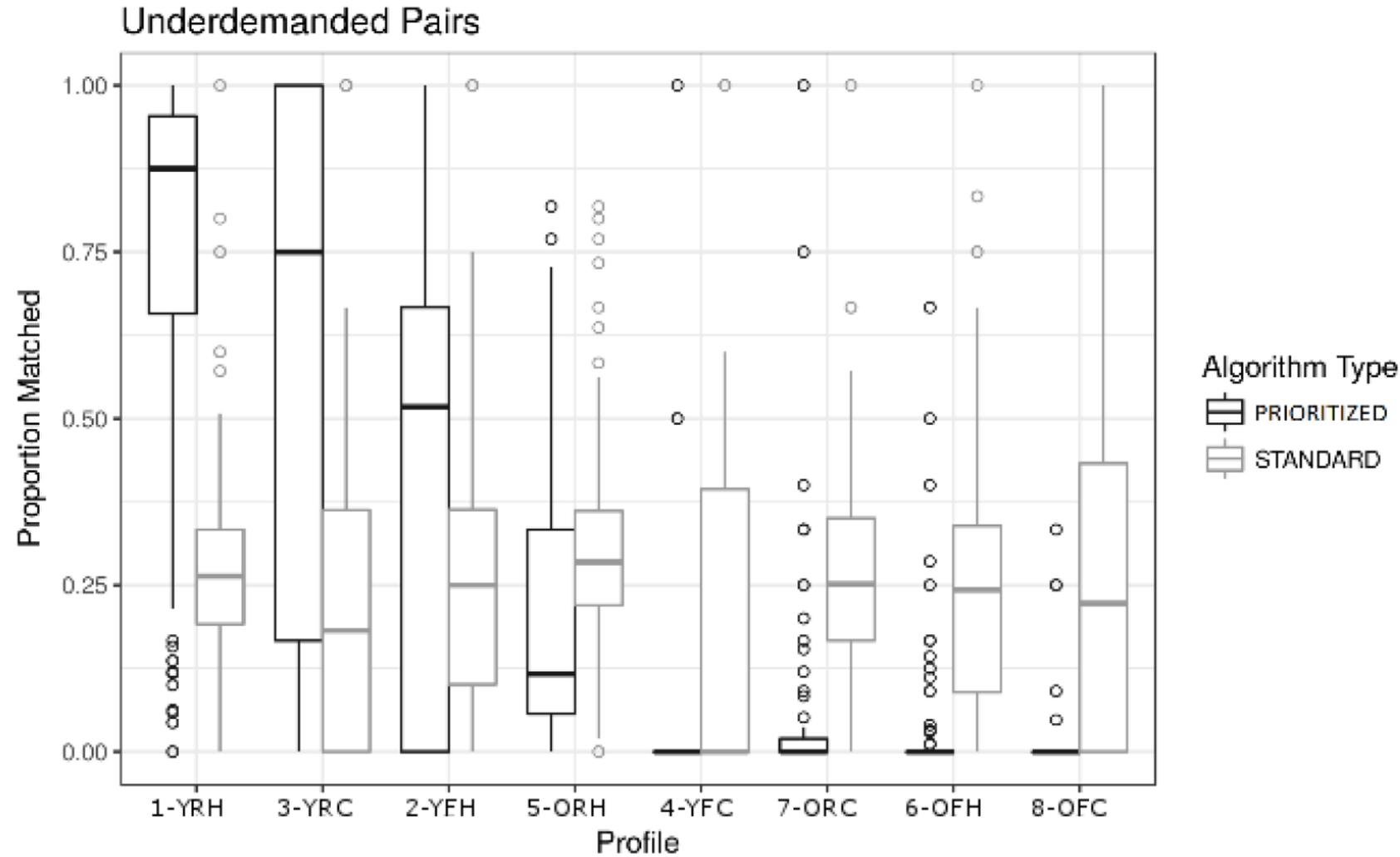
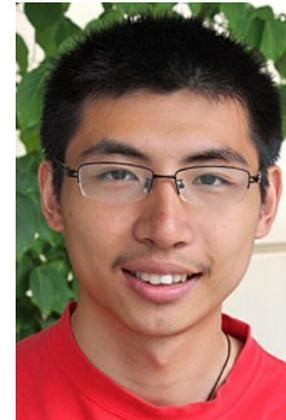


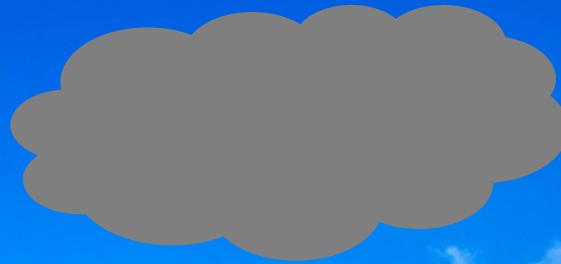
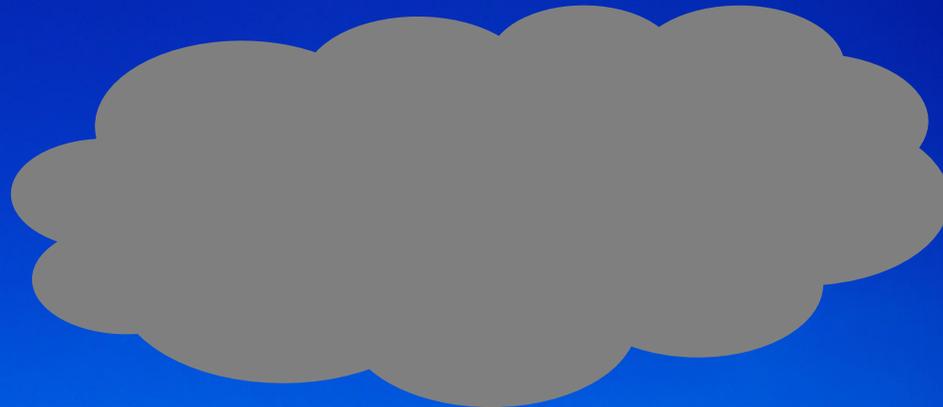
Figure 4: The proportions of underdemanded pairs matched over the course of the simulation, by profile type and algorithm type. N = 20 runs were used for each box.

# Crowdsourcing Societal Tradeoffs

*(AAMAS'15 blue sky paper; AAAI'16; ongoing work.)*



with Rupert Freeman, Markus Brill, Yuqian Li



# Example Decision Scenario

- Benevolent government would like to get old inefficient cars off the road
- But disposing of a car and building a new car has its own energy (and other) costs
- Which cars should the government aim to get off the road?
  - even energy costs are **not directly comparable** (e.g., perhaps gasoline contributes to energy dependence, coal does not)



# The basic version of our problem



producing 1 bag  
of landfill trash

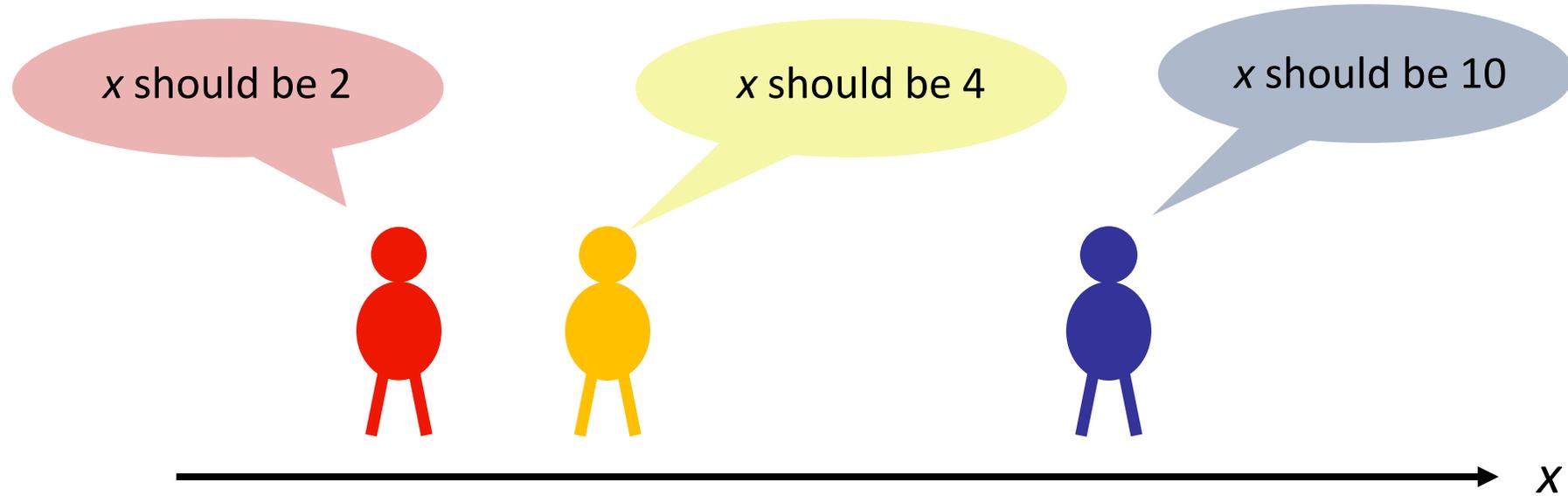
*is as bad as*



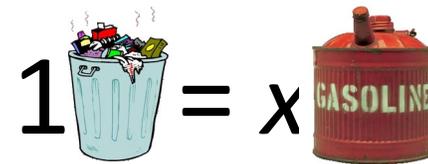
using  $x$  gallons  
of gasoline

*How to determine  $x$ ?*

# One Approach: Let's Vote!

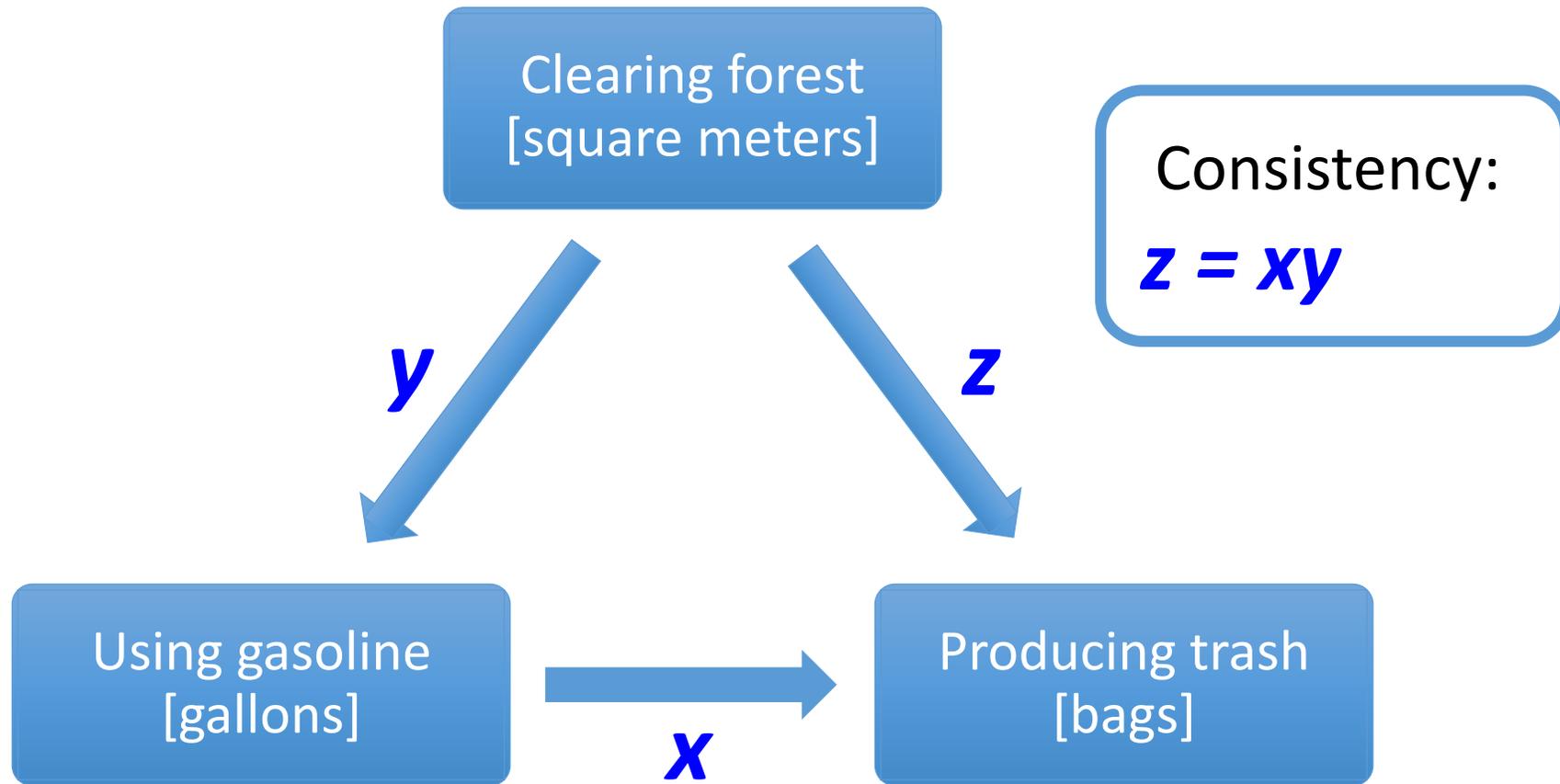


- What should the outcome be...?
  - Average? Median?

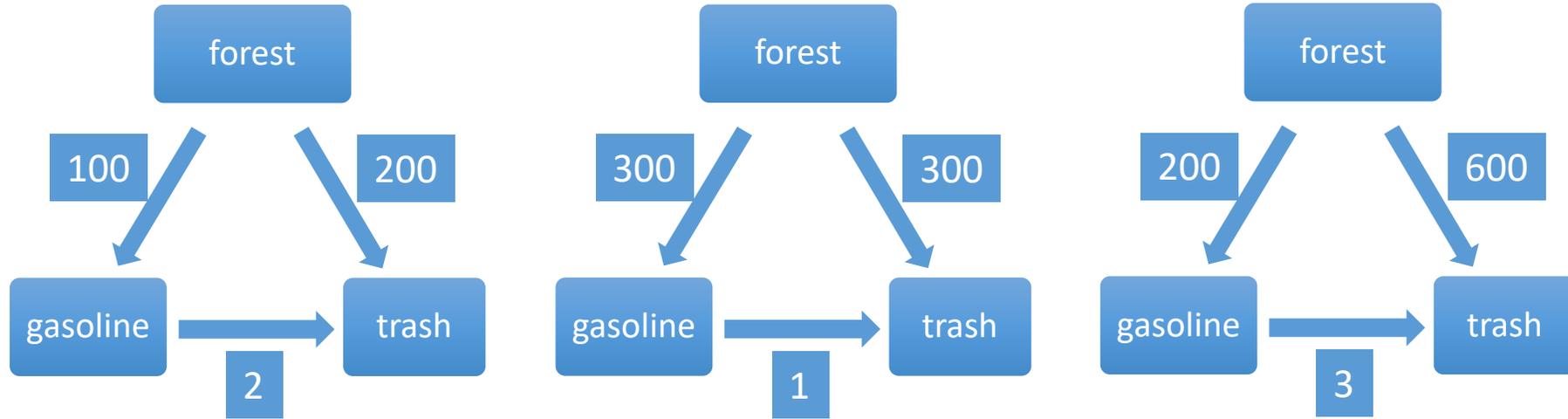


- Assuming that preferences are single-peaked, selecting the **median** is strategy-proof and has other desirable social choice-theoretic properties

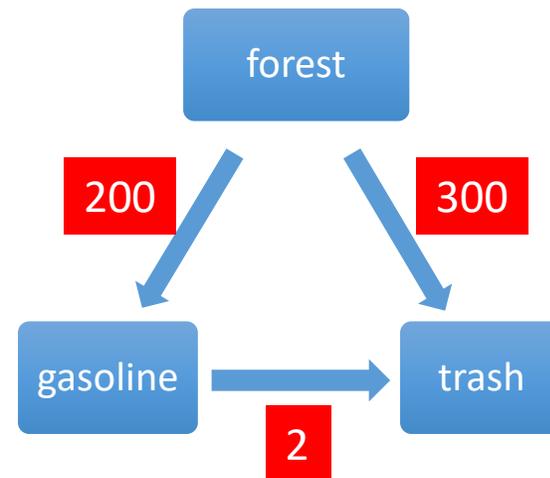
# Consistency of tradeoffs



# A paradox

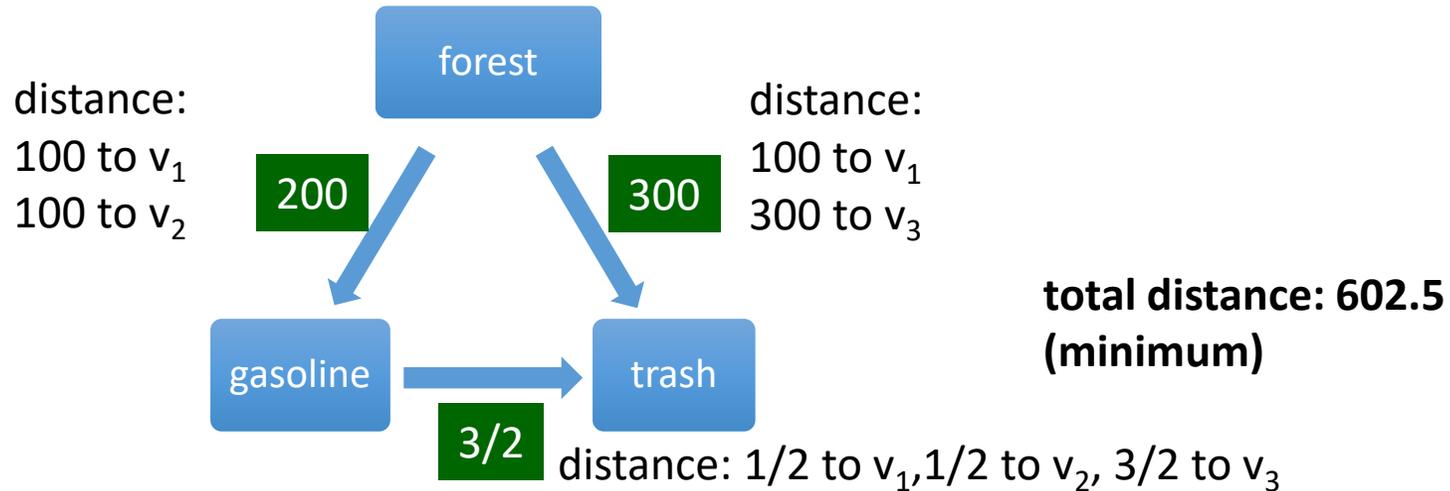
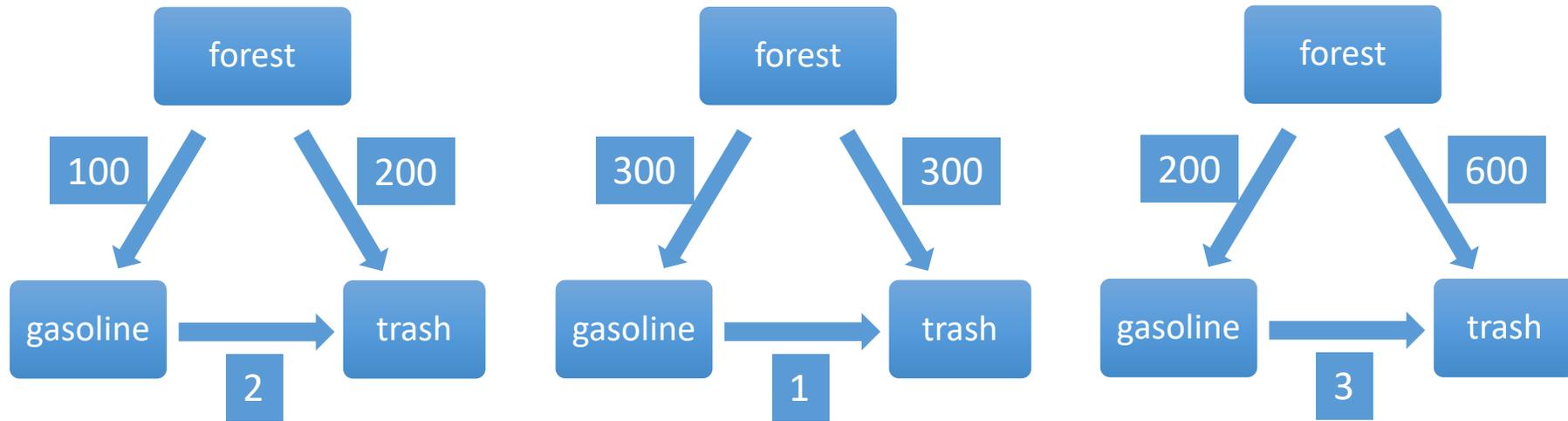


Just taking  
medians  
pairwise results  
in inconsistency



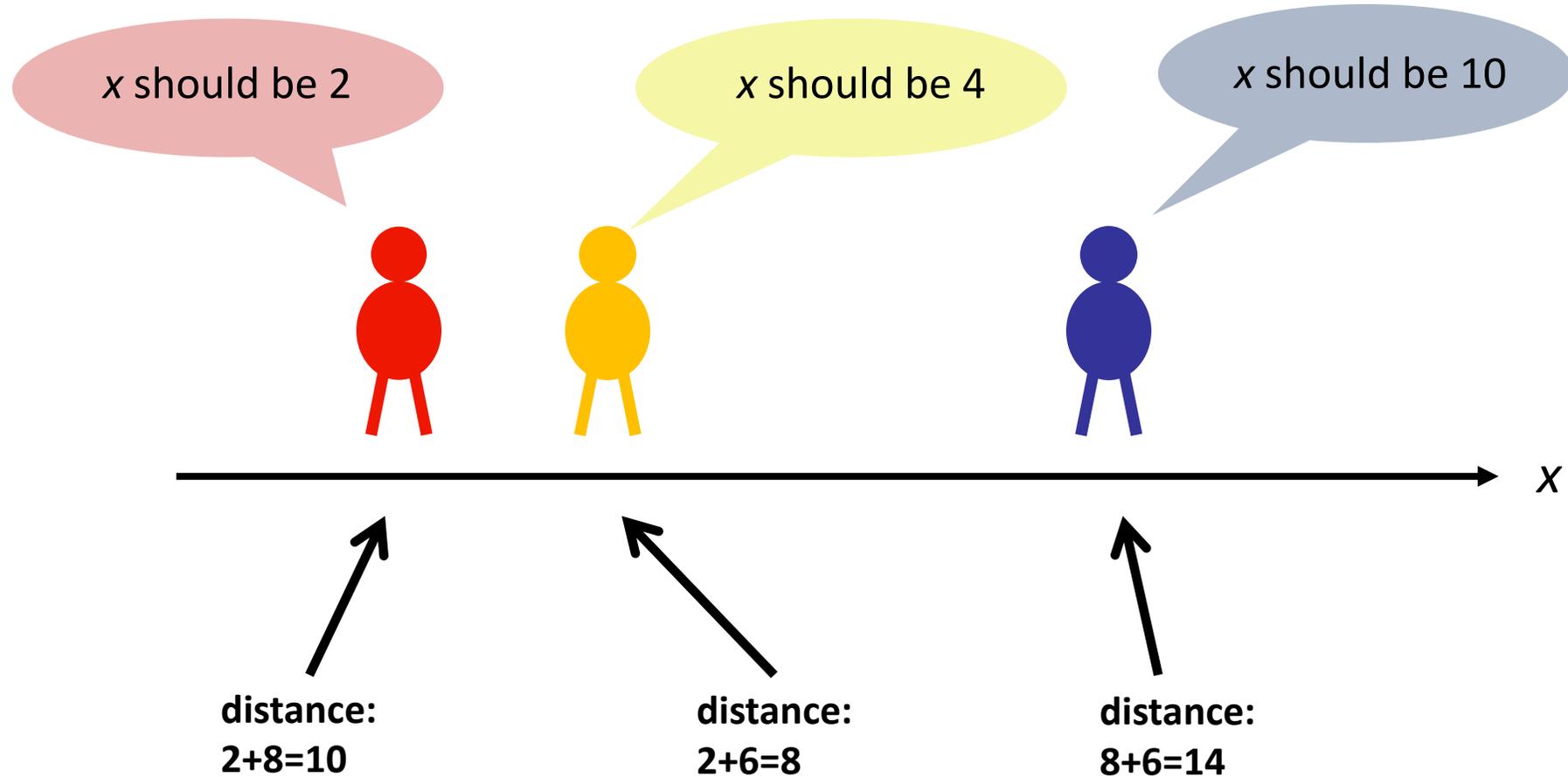
# A first attempt at a rule satisfying consistency

- Let  $t_{a,b,i}$  be voter  $i$ 's tradeoff between  $a$  and  $b$
- Aggregate tradeoff  $t$  has score  $\sum_i \sum_{a,b} |t_{a,b} - t_{a,b,i}|$



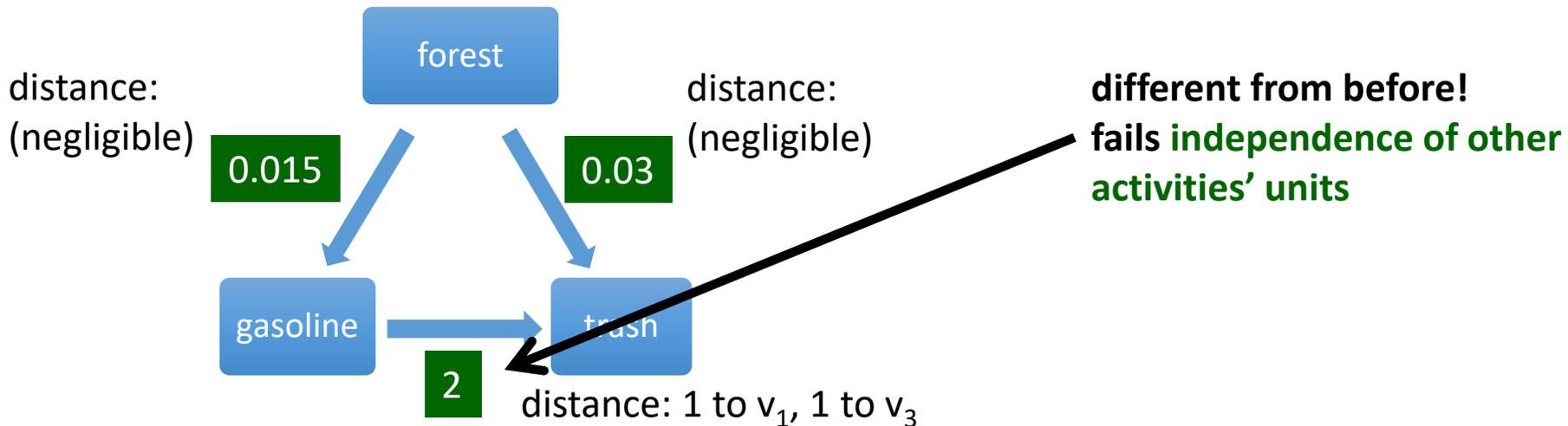
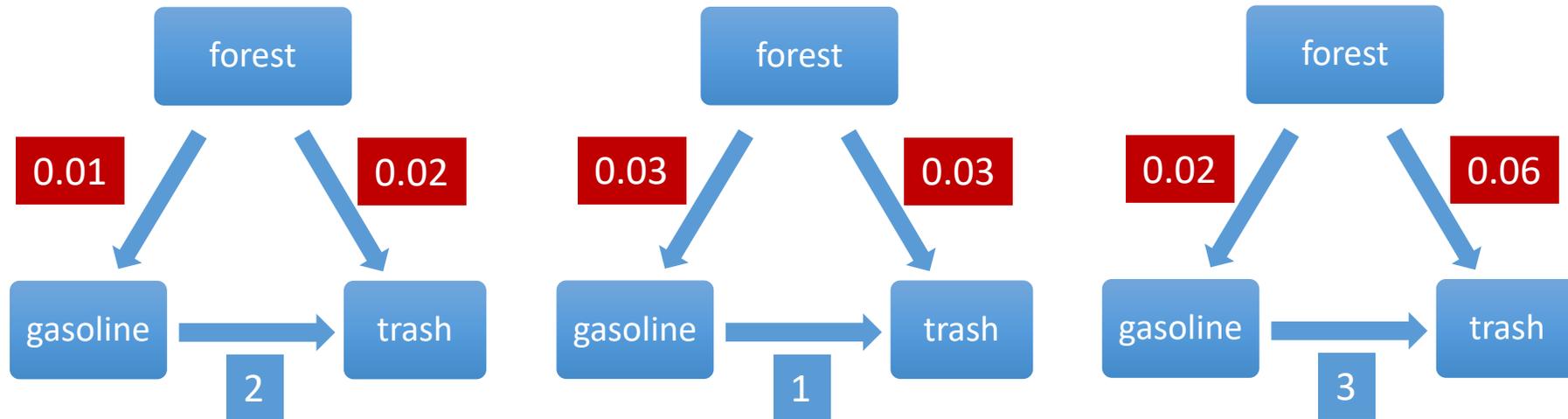
# A nice property

- This rule **agrees with the median** when there are only two activities!



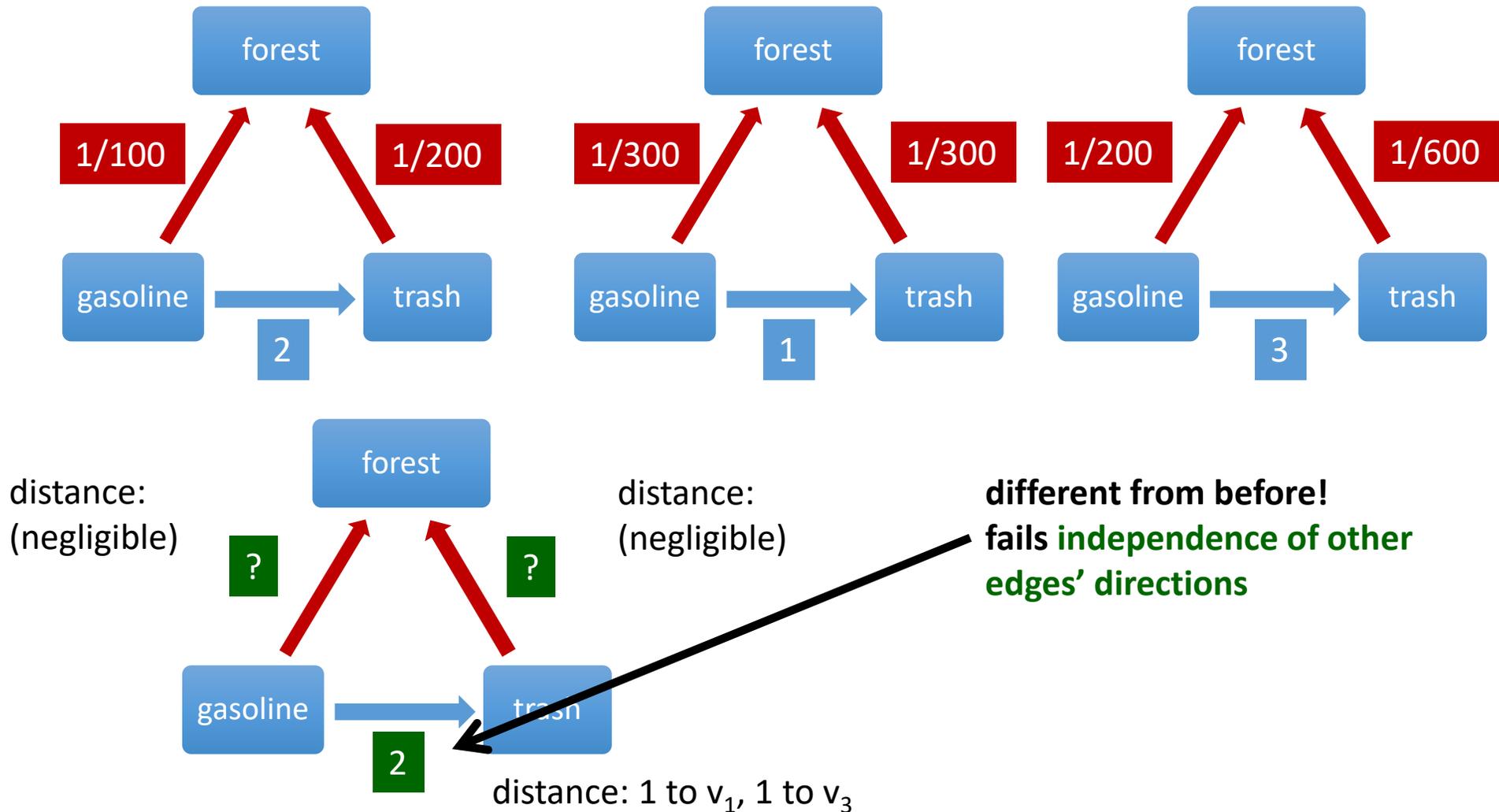
# Not all is rosy, part 1

- What if we **change units**? Say forest from  $\text{m}^2$  to  $\text{cm}^2$  (divide by 10,000)



# Not all is rosy, part 2

- Back to original units, but let's change some edges' direction



# Summarizing

- Let  $t_{a,b,i}$  be voter  $i$ 's tradeoff between  $a$  and  $b$
- Aggregate tradeoff  $t$  has score

$$\sum_i \sum_{a,b} | t_{a,b} - t_{a,b,i} |$$

- Upsides:
  - Coincides with median for 2 activities
- Downsides:
  - Dependence on **choice of units**:  
 $| t_{a,b} - t_{a,b,i} | \neq | 2t_{a,b} - 2t_{a,b,i} |$
  - Dependence on **direction of edges**:  
 $| t_{a,b} - t_{a,b,i} | \neq | 1/t_{a,b} - 1/t_{a,b,i} |$
  - We **don't have a general algorithm**

# A generalization

- Let  $t_{a,b,i}$  be voter  $i$ 's tradeoff between  $a$  and  $b$
- Let  $f$  be a monotone increasing function – say,  $f(x) = x^2$
- Aggregate tradeoff  $t$  has score  
$$\sum_i \sum_{a,b} | f(t_{a,b}) - f(t_{a,b,i}) |$$
- Still **coincides with median** for 2 activities!
- **Theorem:** These are the **only** rules satisfying this property, agent separability, and edge separability

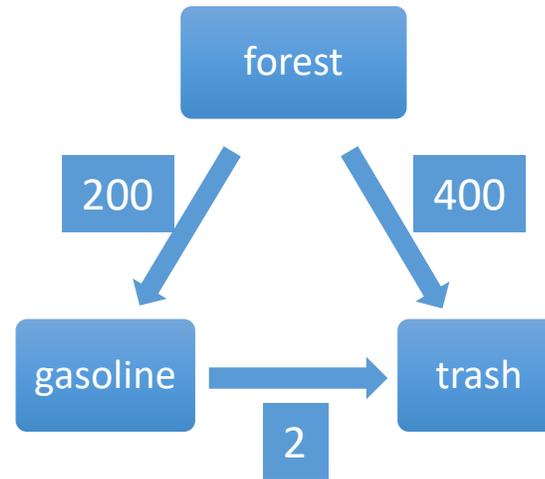
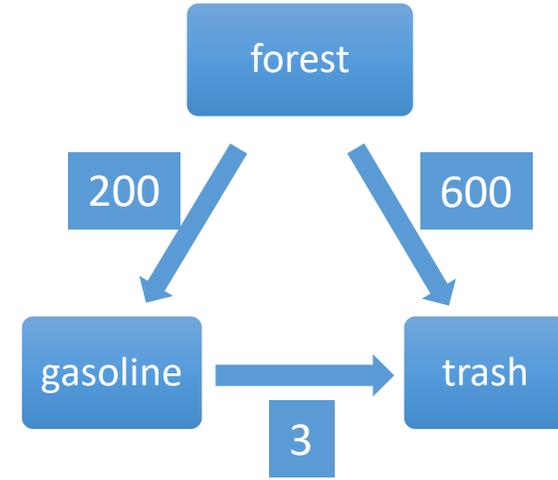
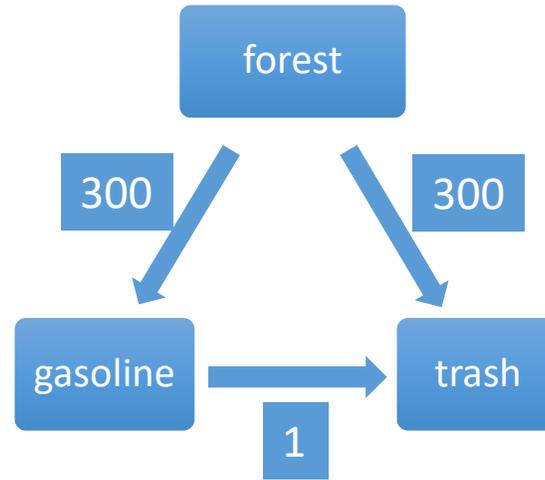
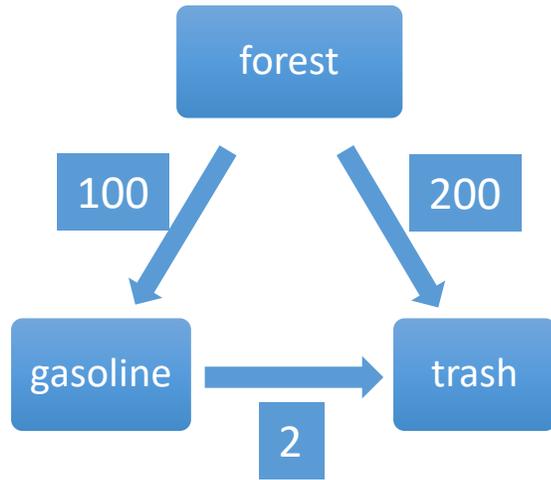
	<b>1</b>	<b>2</b>	<b>3</b>	
$t_{a,b}$	<hr/>			
$f(t_{a,b})$	<b>1</b>	<b>4</b>	<b>9</b>	
	<hr/>			

# So what's a good f?

- **Intuition:** Is the difference between tradeoffs of 1 and 2 the same as between 1000 and 1001, or as between 1000 and 2000?
- So how about  $f(x)=\log(x)$ ?
  - (Say, base e – remember  $\log_a(x)=\log_b(x)/\log_b(a)$  )

$t_{a,b}$	<b>1</b>	<b>2</b>	<b>1000</b>	<b>2000</b>
$\ln(t_{a,b})$	<b><math>\ln(1)</math></b>	<b><math>\ln(2)</math></b>	<b><math>\ln(1000)</math></b>	<b><math>\ln(2000)</math></b>
	0	0.69	6.91	7.60

# On our example



# Properties

- Independence of units

$$| \log(1) - \log(2) | = | \log(1/2) | =$$

$$| \log(1000/2000) | = | \log(1000) - \log(2000) |$$

More generally:

$$| \log(ax) - \log(ay) | = | \log(x) - \log(y) |$$

- Independence of edge direction

$$| \log(x) - \log(y) | = | \log(1/y) - \log(1/x) | =$$

$$| \log(1/x) - \log(1/y) |$$

- **Theorem.** The logarithmic distance based rule is unique in satisfying independence of units.\*

\* Depending on the exact definition of independence of units, may need another minor condition about the function locally having bounded derivative.

# Consistency constraint becomes additive

$$xy = z$$

is equivalent to

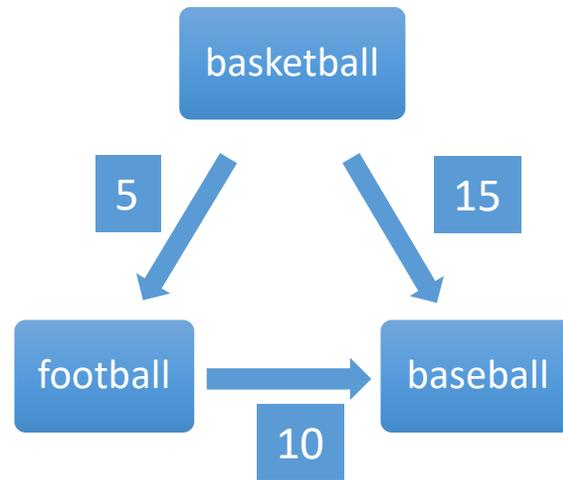
$$\log(xy) = \log(z)$$

is equivalent to

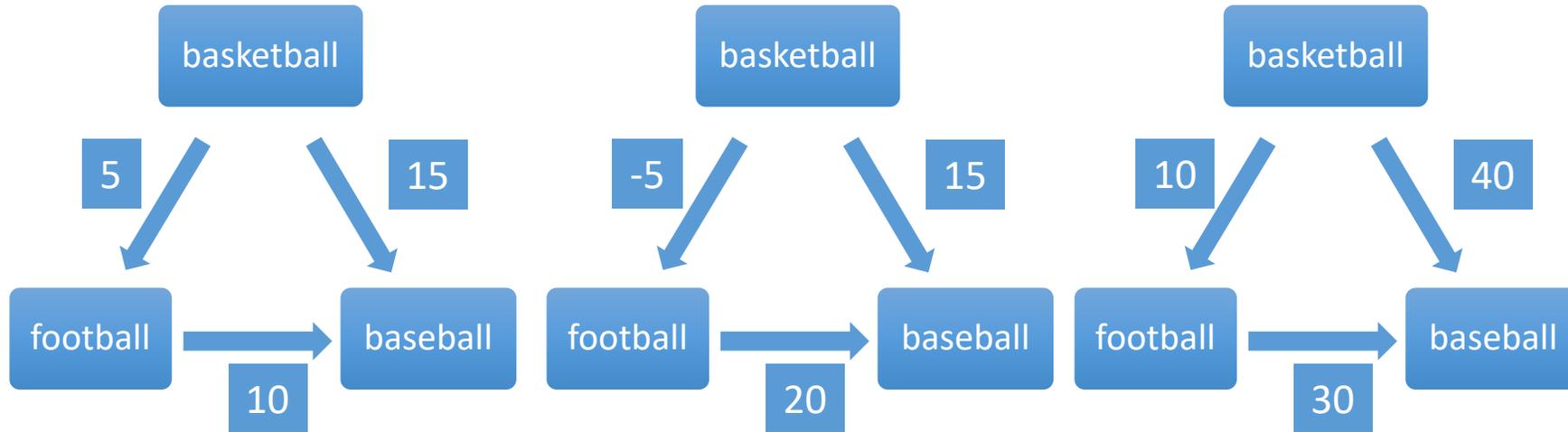
$$\log(x) + \log(y) = \log(z)$$

# An additive variant

- “I think basketball is 5 units more fun than football, which in turn is 10 units more fun than baseball”

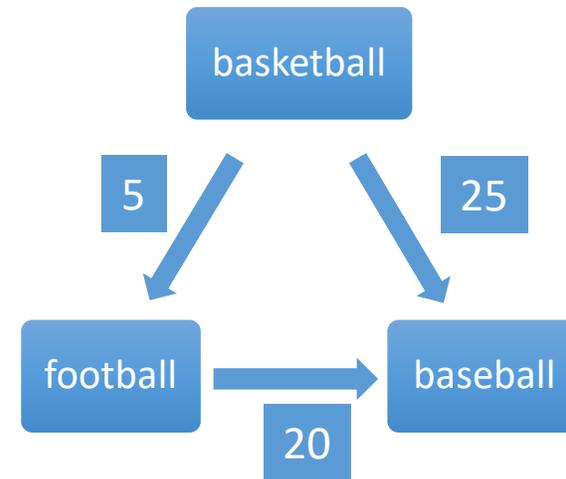


# Aggregation in the additive variant



Natural objective:

minimize  $\sum_i \sum_{a,b} d_{a,b,i}$  where  $d_{a,b,i}$   
 $= |t_{a,b} - t_{a,b,i}|$  is the distance  
 between the aggregate  
 difference  $t_{a,b}$  and the subjective  
 difference  $t_{a,b,i}$



objective value 70 (optimal)

# A linear program for the additive variant

$q_a$ : aggregate assessment of quality of activity  $a$  (we're really interested in  $q_a - q_b = t_{a,b}$ )

$d_{a,b,i}$ : how far is  $i$ 's preferred difference  $t_{a,b,i}$  from aggregate  $q_a - q_b$ , i.e.,  $d_{a,b,i} = |q_a - q_b - t_{a,b,i}|$

minimize  $\sum_i \sum_{a,b} d_{a,b,i}$

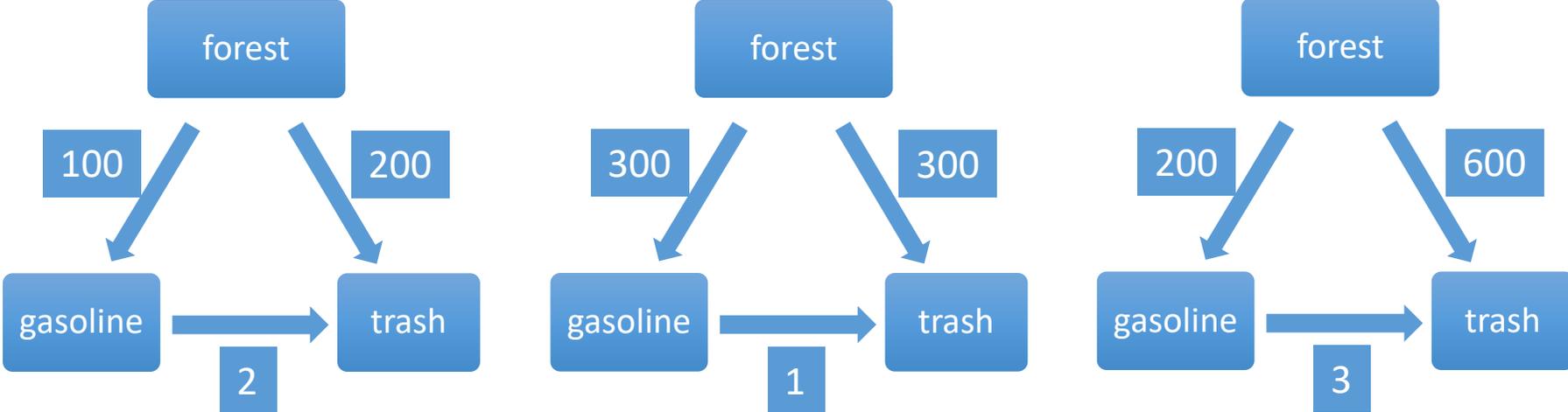
subject to

for all  $a,b,i$ :  $d_{a,b,i} \geq q_a - q_b - t_{a,b,i}$

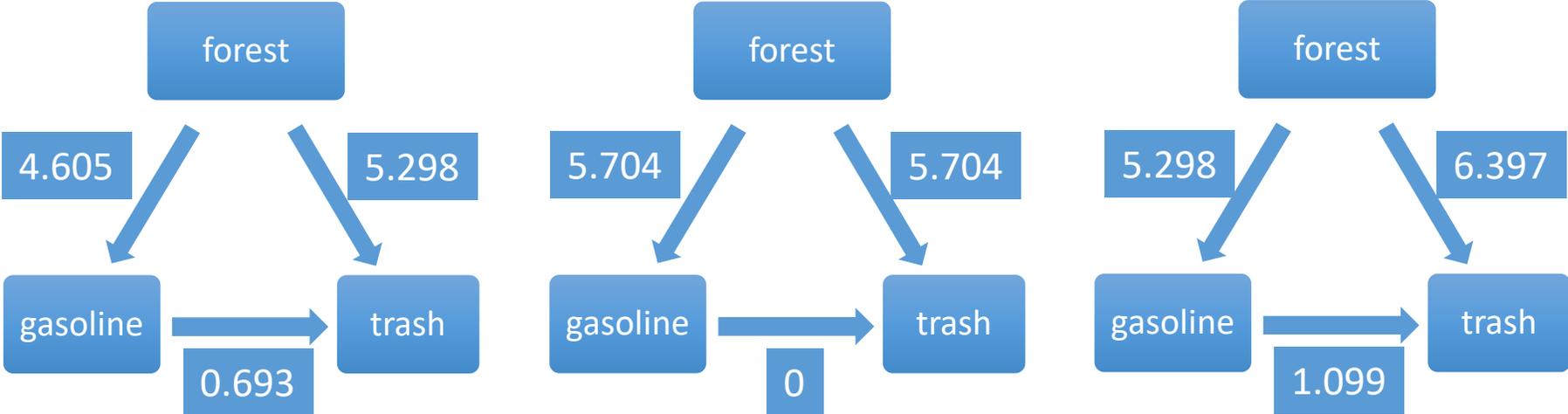
for all  $a,b,i$ :  $d_{a,b,i} \geq t_{a,b,i} - q_a + q_b$

(Can arbitrarily set one of the  $q$  variables to 0)

# Applying this to the logarithmic rule in the multiplicative variant

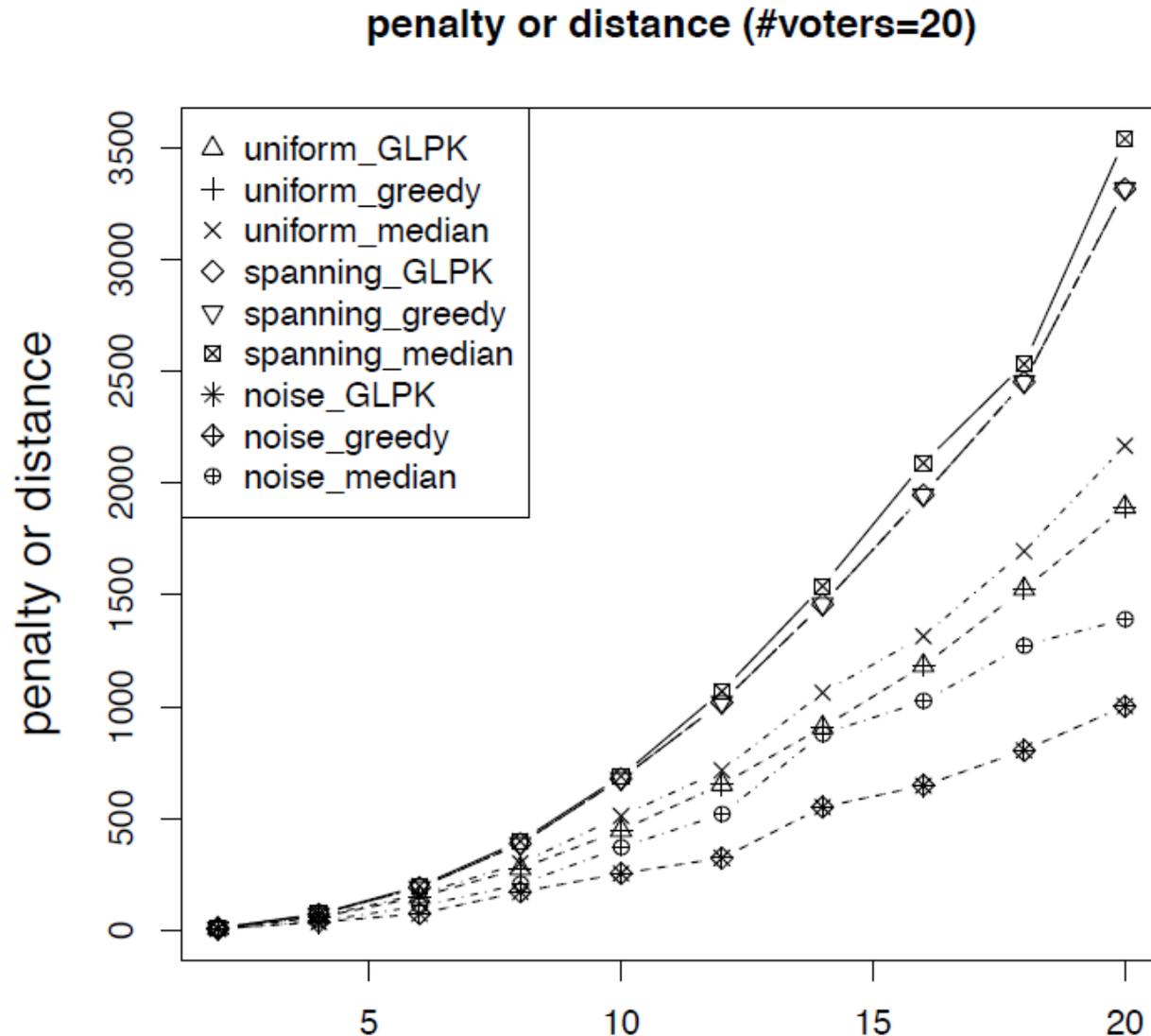


Just take logarithms on the edges, solve the additive variant, and exponentiate back

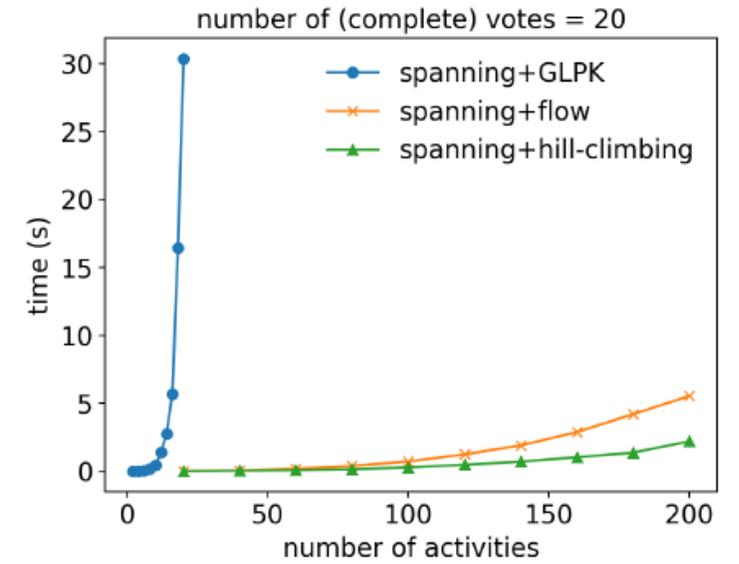
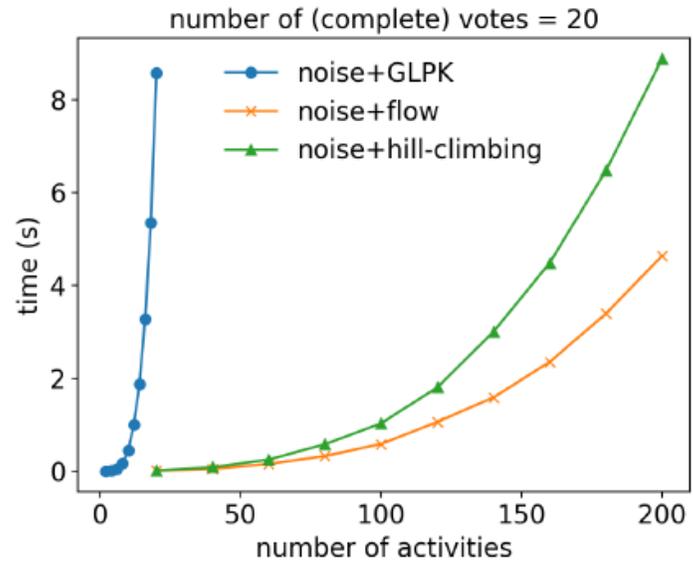
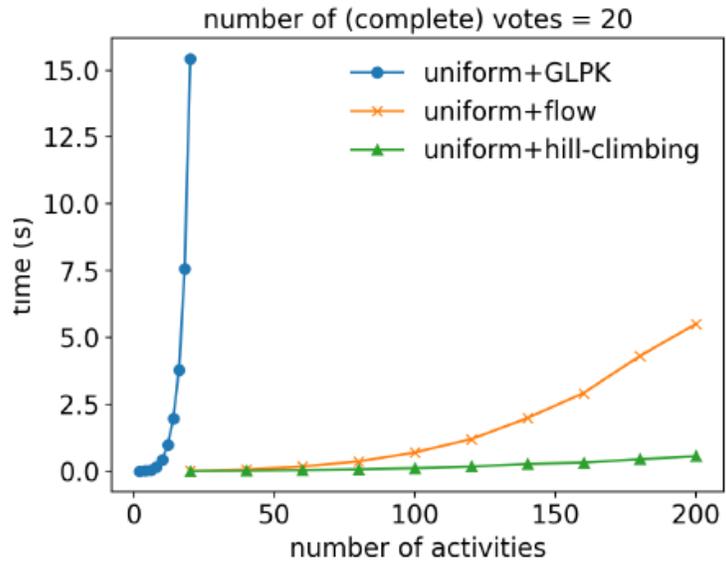


# A simpler algorithm (hill climbing / greedy)

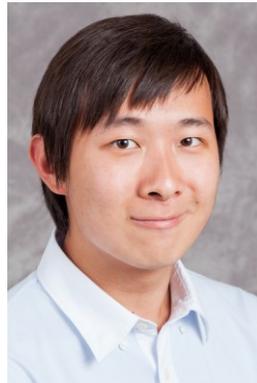
- Initialize qualities  $q_a$  arbitrarily
- If some  $q_a$  can be individually changed to improve the objective, do so
  - WLOG, set  $q_a$  to the median of the  $(\text{\#voters}) * (\text{\#activities}-1)$  implied votes on it
- Continue until convergence (possibly to local optimum)



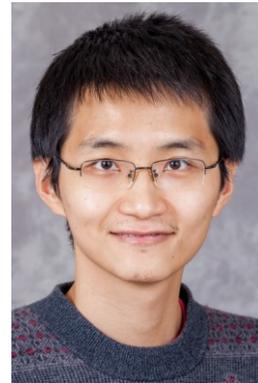
# Flow-based exact algorithm



with:



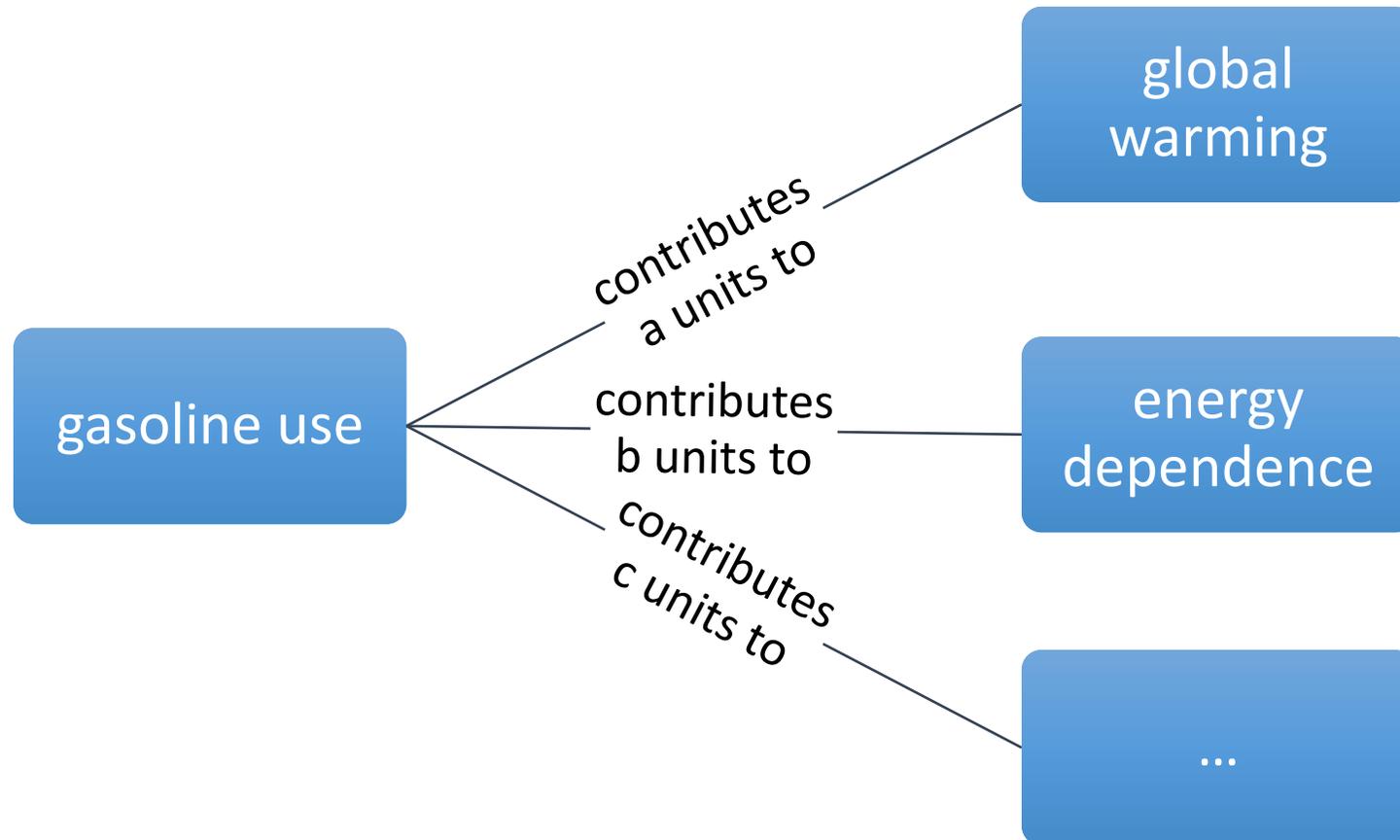
Hanrui  
Zhang



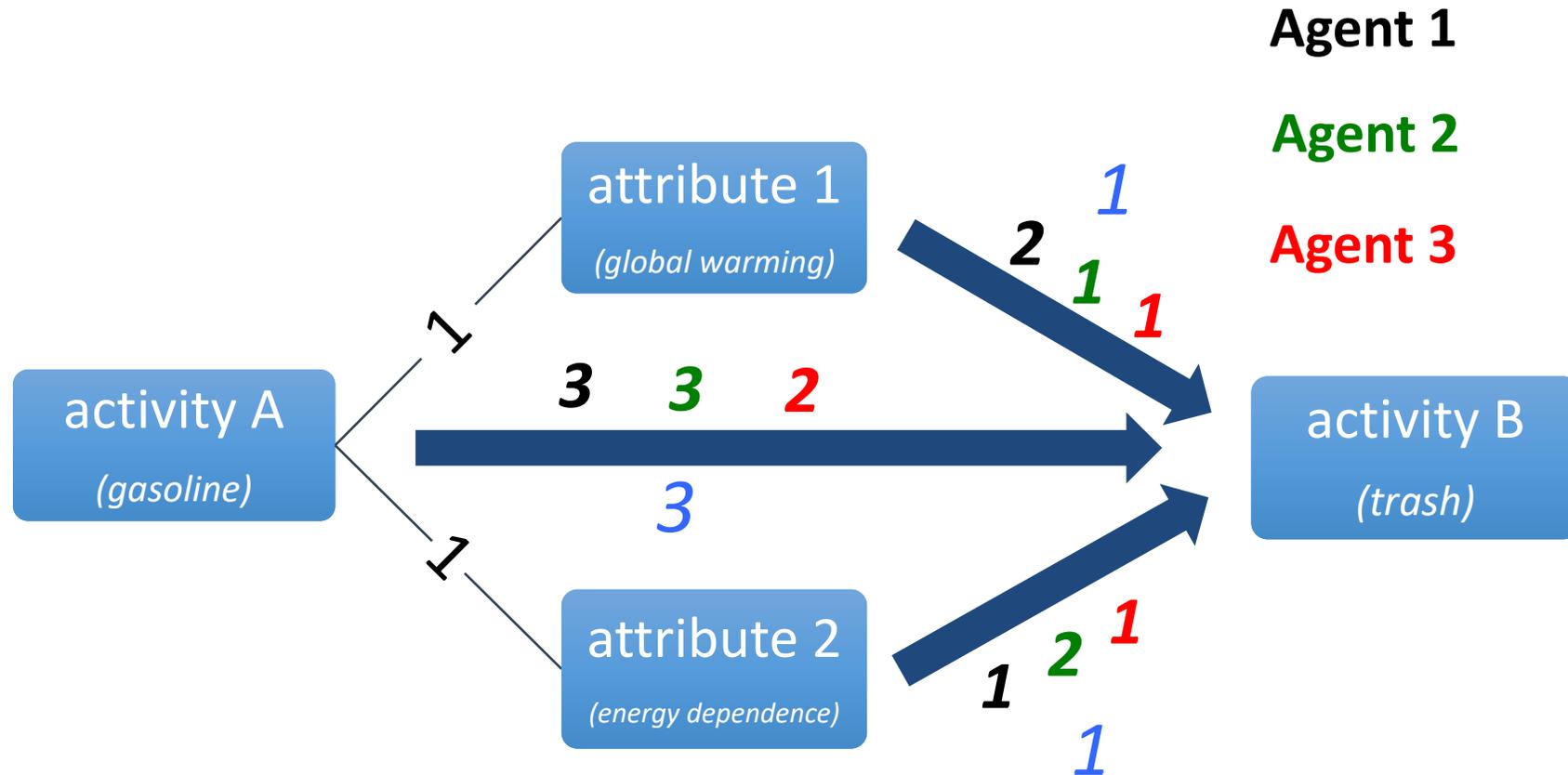
Yu  
Cheng

# Decomposition

- Idea: Break down activities to relevant attributes



# Another Paradox



aggregation on attribute level  $\neq$  aggregation on activity level

# Other Issues

- **Objective** vs. **subjective** tradeoffs
  - separate process?
  - who determines which is which?
- **Who gets to vote?**
  - how to bring **expert knowledge** to bear?
  - incentives to **participate**
- **Global** vs. **local** tradeoffs
  - different entities (e.g., countries) may wish to reach their tradeoffs **independently**
  - only care about opinions of **neighbors in my social network**
- ...

## Relevant Topics

- social choice theory
  - voting
  - judgment aggregation
- game theory
- mechanism design
- prediction markets
- peer prediction
- preference elicitation
- ...

Thank you for your  
attention!

# Why Do We Care?

- Inconsistent tradeoffs can result in **inefficiency**
  - Agents optimizing their utility functions individually leads to solutions that are Pareto inefficient
- **Pigovian taxes**: pay the cost your activity imposes on society (the **externality** of your activity)
  - If we decided using 1 gallon of gasoline came at a cost of \$ $x$  to society, we could charge a tax of \$ $x$  on each gallon
  - But where would we get  $x$ ?



*Arthur Cecil Pigou*

# Inconsistent tradeoffs can result in inefficiency

- Agent 1: 1 gallon = 3 bags = -1 util
  - I.e., agent 1 feels she should be willing to sacrifice up to 1 util to reduce trash by 3, but no more
- Agent 2: 1.5 gallons = 1.5 bags = -1 util
- Agent 3: 3 gallons = 1 bag = -1 util
- Cost of reducing gasoline by  $x$  is  $x^2$  utils for each agent
- Cost of reducing trash by  $y$  is  $y^2$  for each agent
- Optimal solutions for the individual agents:
  - Agent 1 will reduce by  $1/2$  and  $1/6$
  - Agent 2 will reduce by  $1/3$  and  $1/3$
  - Agent 3 will reduce by  $1/6$  and  $1/2$
- But if agents 1 and 3 each reduce everything by  $1/3$ , the total reductions are the same, and their costs are  $2/9$  rather than  $1/4 + 1/36$  which is clearly higher.
  - Could then reduce slightly more to make everyone happier.

# Single-peaked preferences

- *Definition:* Let agent  $a$ 's most-preferred value be  $p_a$ .

Let  $p$  and  $p'$  satisfy:

-  $p' \leq p \leq p_a$ , or  $p_a \leq p \leq p'$

- The agent's preferences are **single-peaked** if the agent always weakly prefers  $p$  to  $p'$

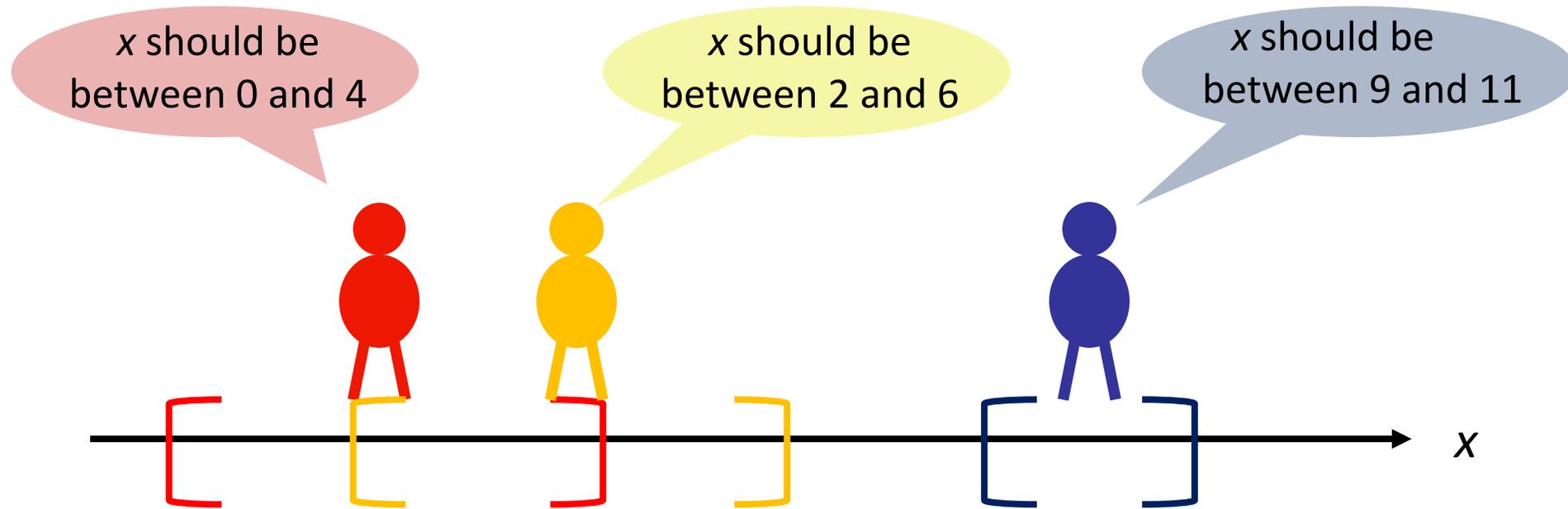
---

$p'$

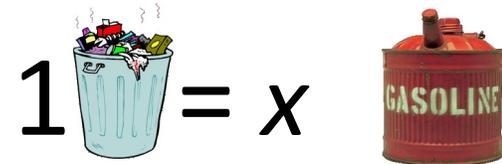
$p$

$p_a$

# Perhaps more reasonable...



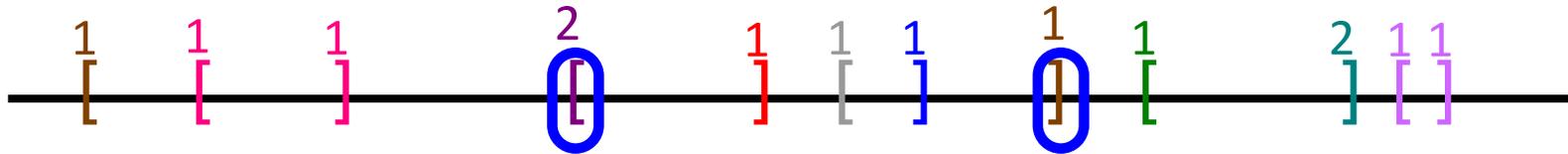
- E.g., due to **missing information** or plain **uncertainty**



- How to aggregate these interval votes? [Farfel & Conitzer 2011]

# Median interval mechanism

- Construct a consensus interval from the median lower bound and the median upper bound



- Strategy-proof if preferences are **single-peaked over intervals**

# Single-peaked preferences over intervals

- *Definition:* Let agent  $a$ 's most-preferred value interval be  $P_a = [l_a, u_a]$ .

Let  $S = [l, u]$  and  $S' = [l', u']$  be **any** two value intervals satisfying the following constraints:

- Either  $l' \leq l \leq l_a$ , or  $l_a \leq l \leq l'$
  - Either  $u' \leq u \leq u_a$ , or  $u_a \leq u \leq u'$
- The agent's preferences over intervals are **single-peaked** if the agent always weakly prefers  $S$  to  $S'$

