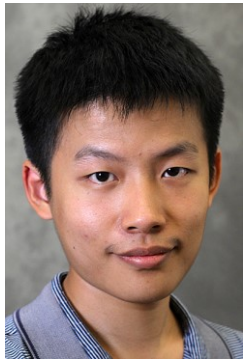
Moral Decision Making Frameworks for Artificial Intelligence

[AAAI'17 blue sky track]

Vincent Conitzer, with:

Walter Sinnott-Armstrong

Jana Schaich Borg

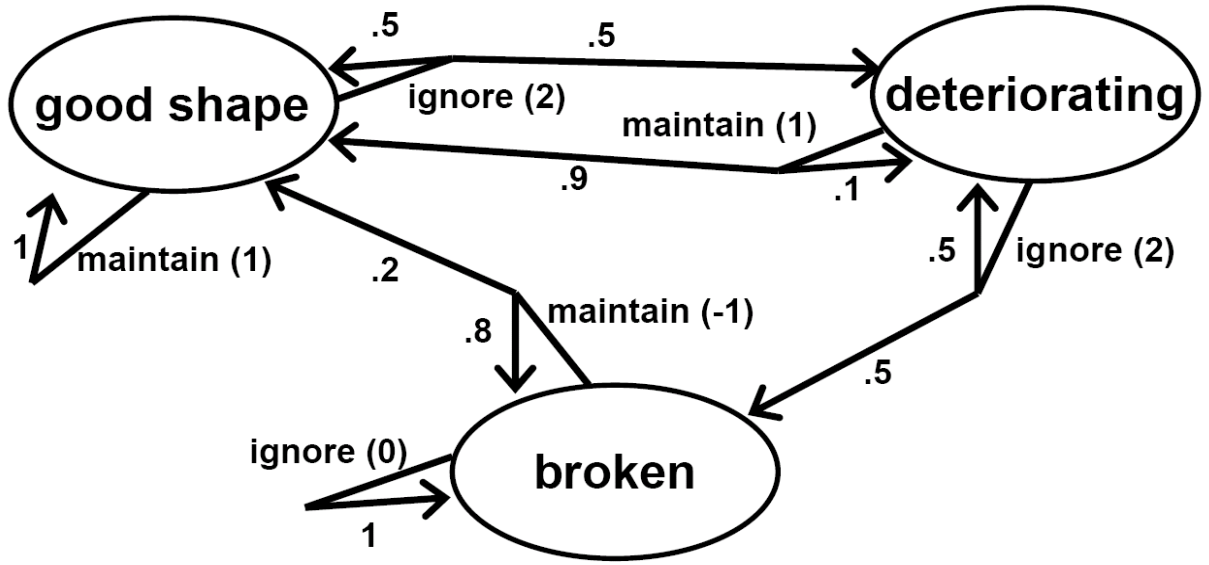


Yuan Deng

Max Kramer

The value of generally applicable frameworks for AI research

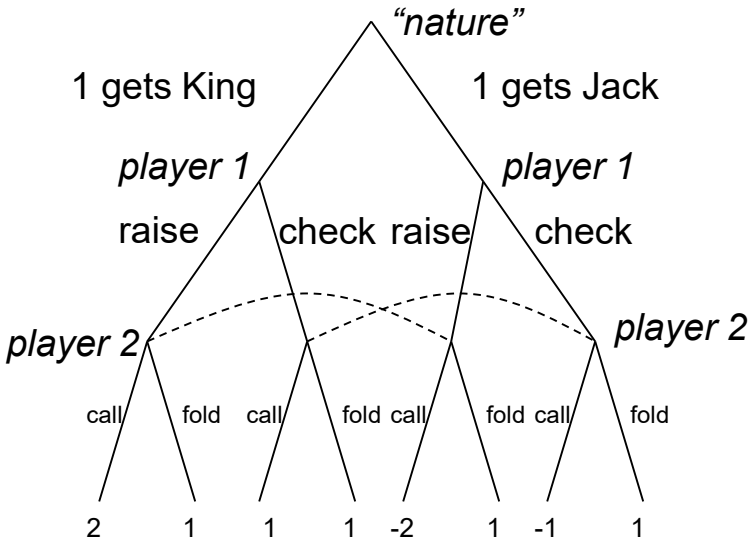
- Decision and game theory
- Example: Markov Decision Processes
- Can we have a **general** framework for moral reasoning?



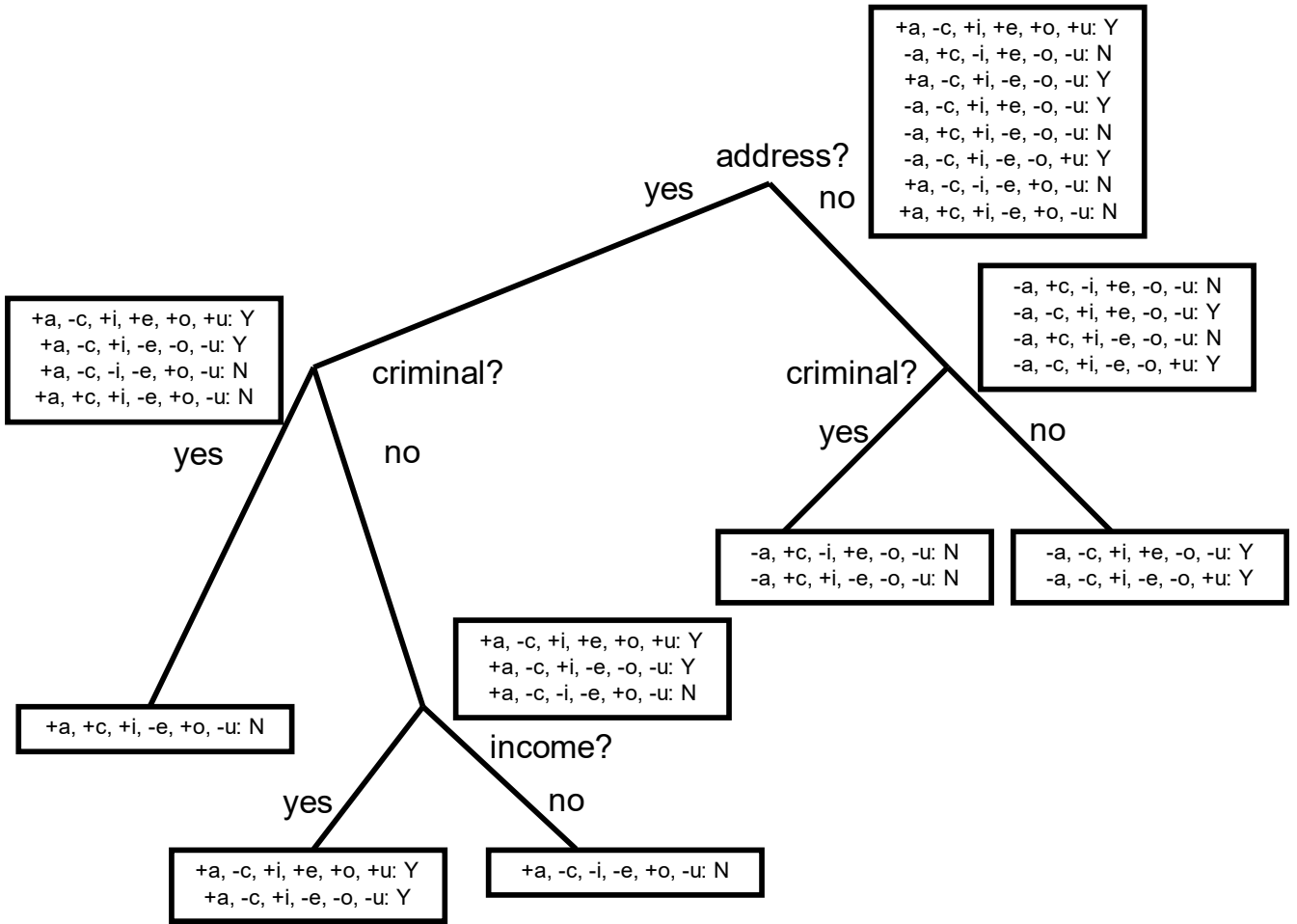
Two main approaches

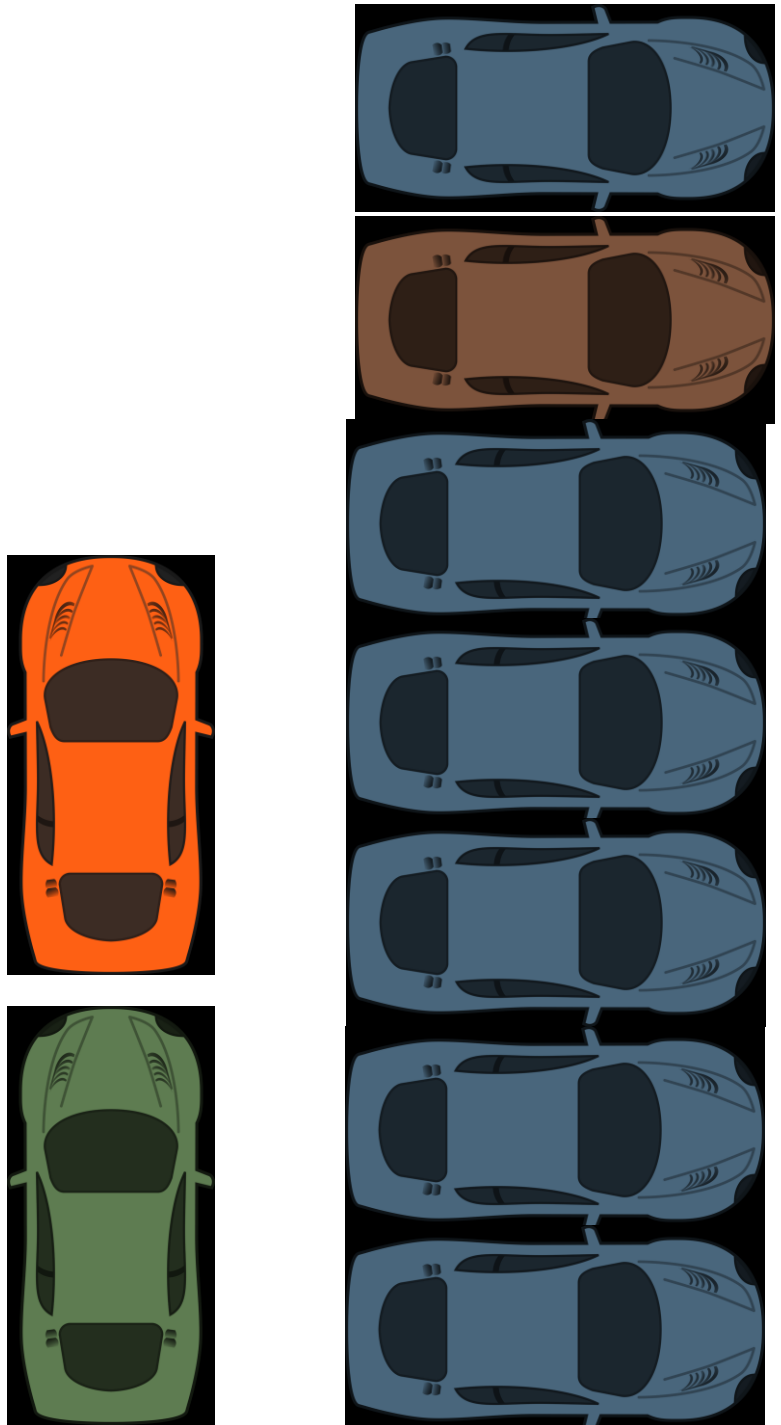
Cf. top-down vs. bottom-up distinction [Wallach and Allen 2008]

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



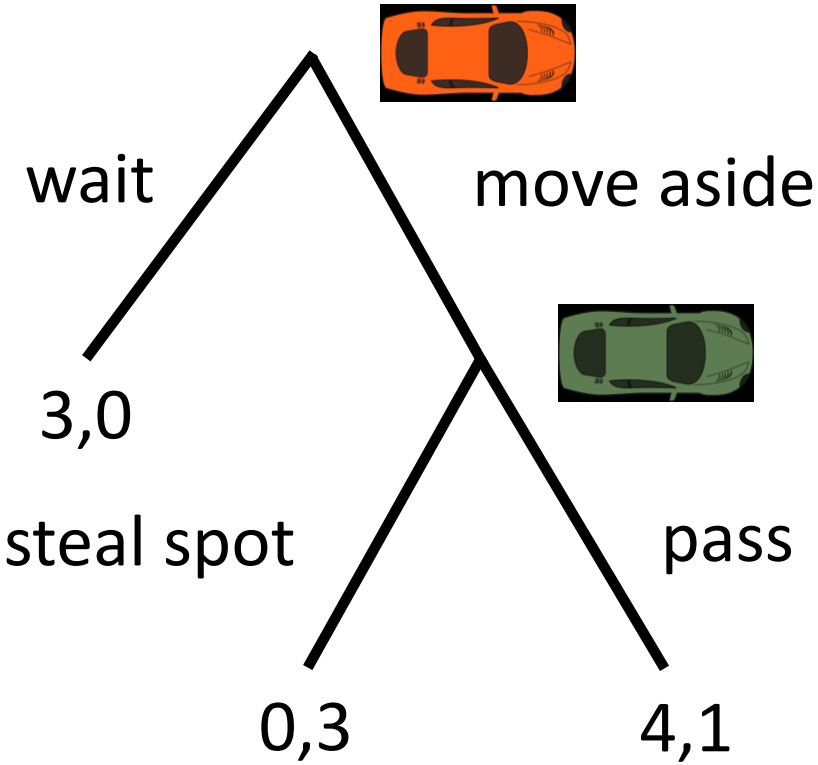
Generate data sets of human judgments, apply **machine learning**





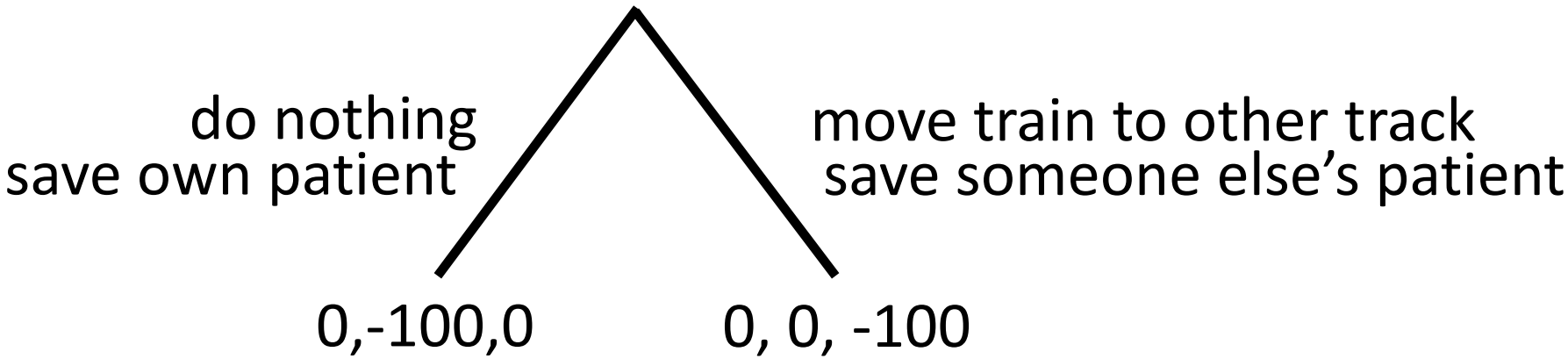
THE PARKING GAME

(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

Moral Machine

moralmachine.mit.edu

MORAL MACHINE

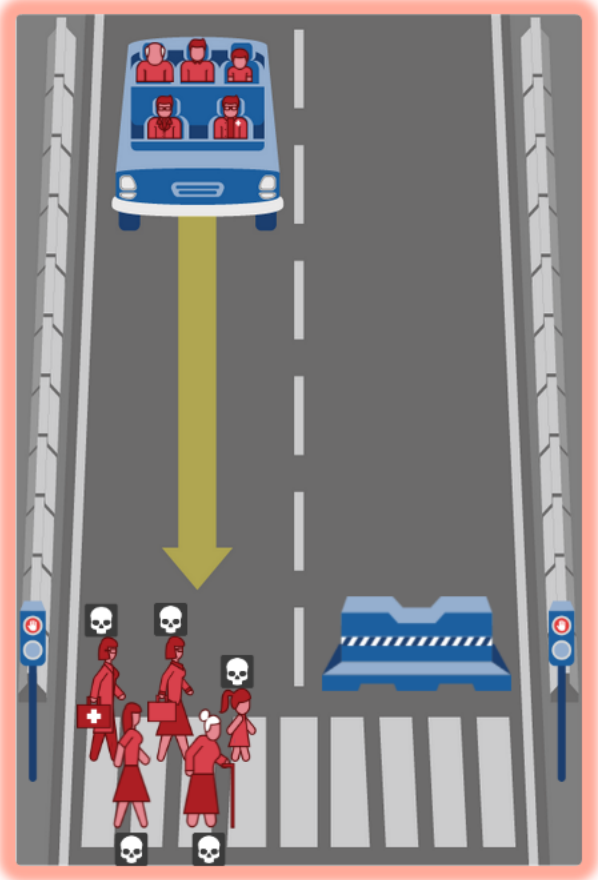
Home Judge Design Browse About Feedback

What should the self-driving car do?

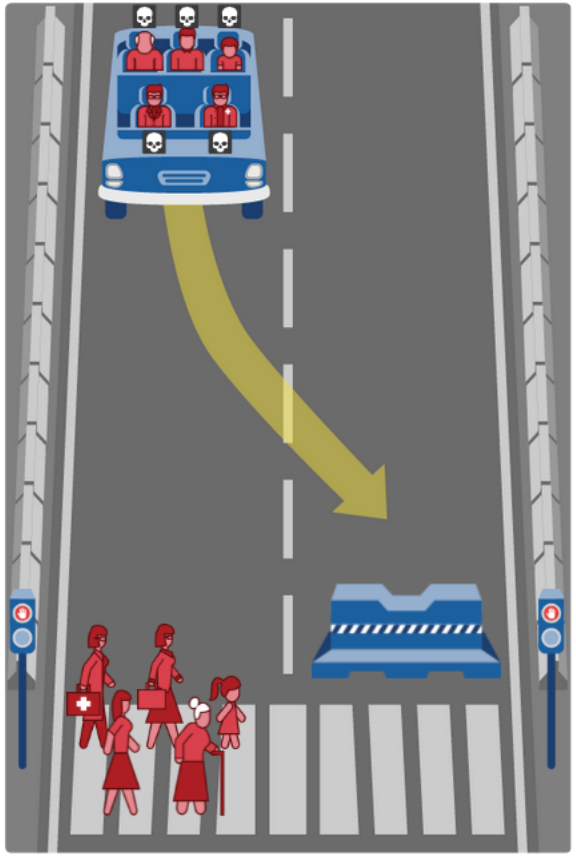
In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in

- 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.

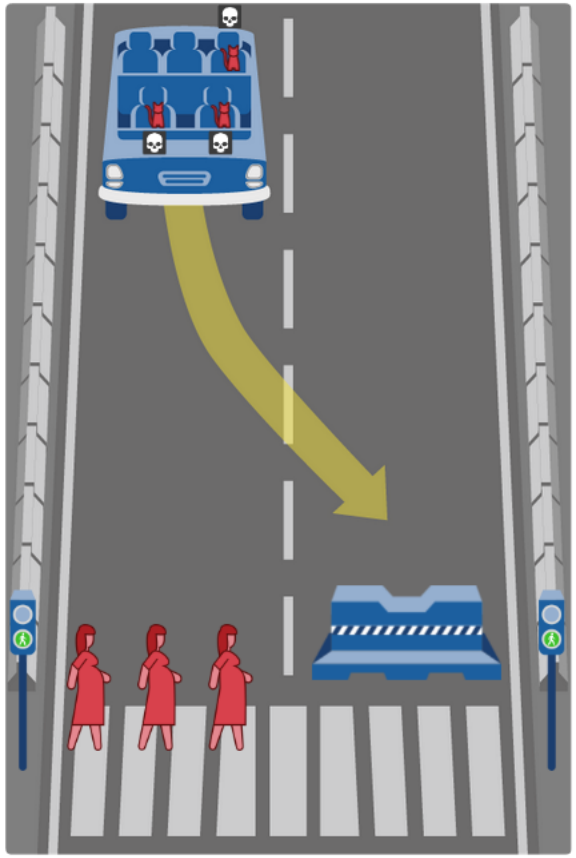
[Bonnefon, Shariff, Rahwan, "The social dilemma of autonomous vehicles." *Science*, June 2016]

Made by Scalable Cooperation at MIT Media Lab

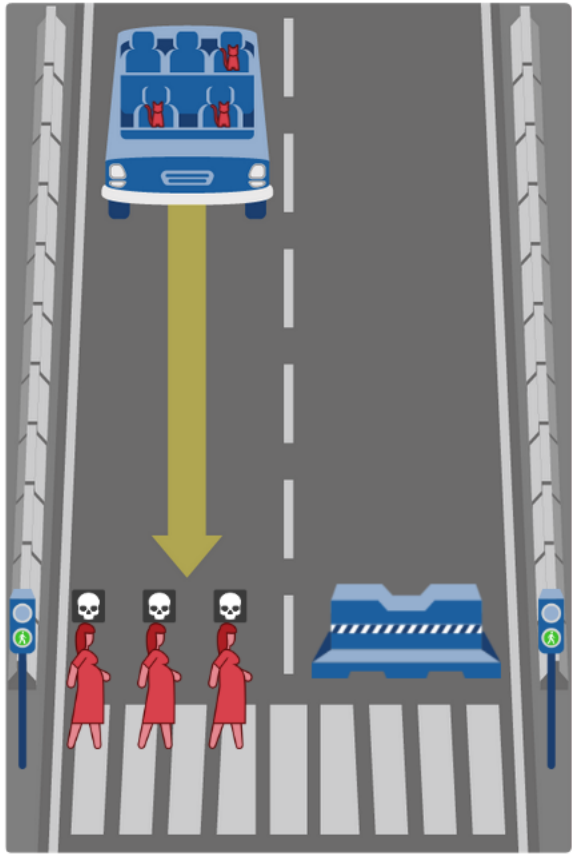
What should the self-driving car do?

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



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.

More Share Link

Results

Most Saved Character

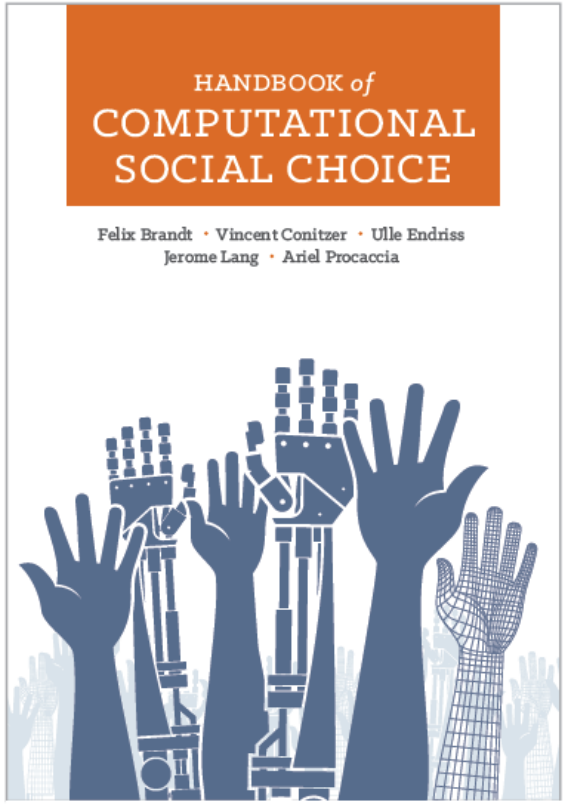
Most Killed Character

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]
- 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?



Some popular articles

Prospect
The leading magazine of ideas

HOME | BLOGS | POLITICS | ECONOMICS & FINANCE | WORLD | ARTS & BOOKS | LIFE

HOME > SCIENCE & TECHNOLOGY

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

MIT Technology Review

Topics+ Top Storie

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

Not 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.

Prospect
The leading magazine of ideas

HOME | BLOGS | POLITICS | ECONOMICS & FINANCE | WORLD | ARTS & BOOKS | LIFE

HOME > BRITISH ACADEMY

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 leopardy to Go to poker. Self-driving cars are

Crowdsourcing Societal Tradeoffs

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



with Rupert Freeman, Markus Brill, Yuqian Li

The basic version of our problem

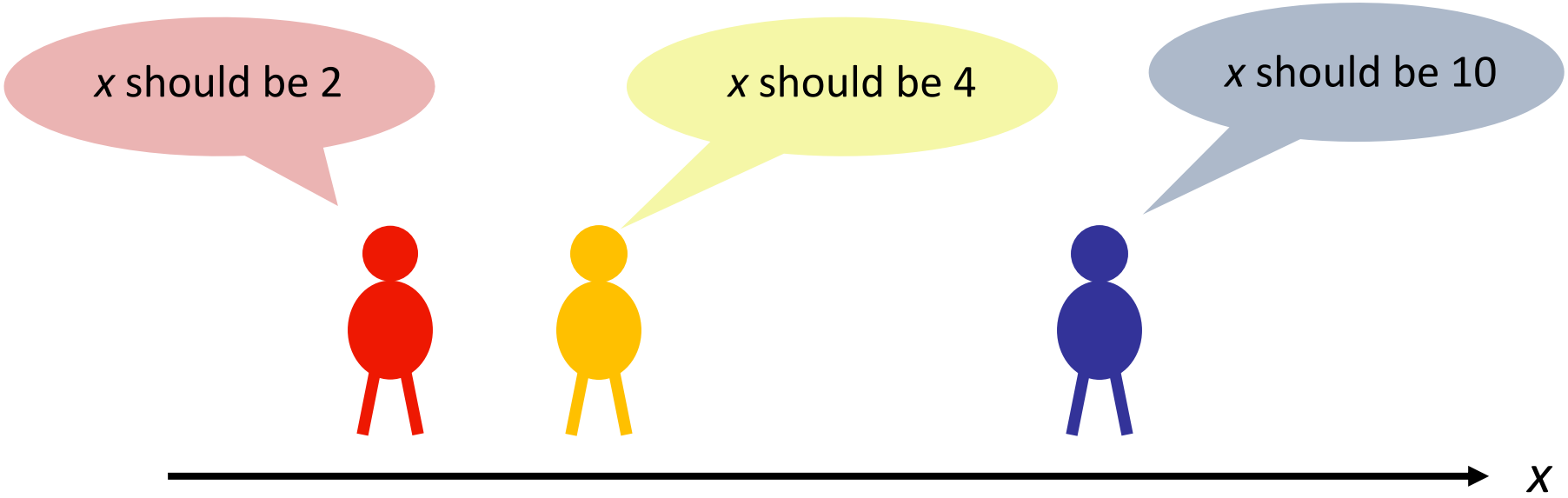
producing 1 bag
of landfill trash

is as bad as

using ***x*** gallons
of gasoline

*How to determine **x**?*

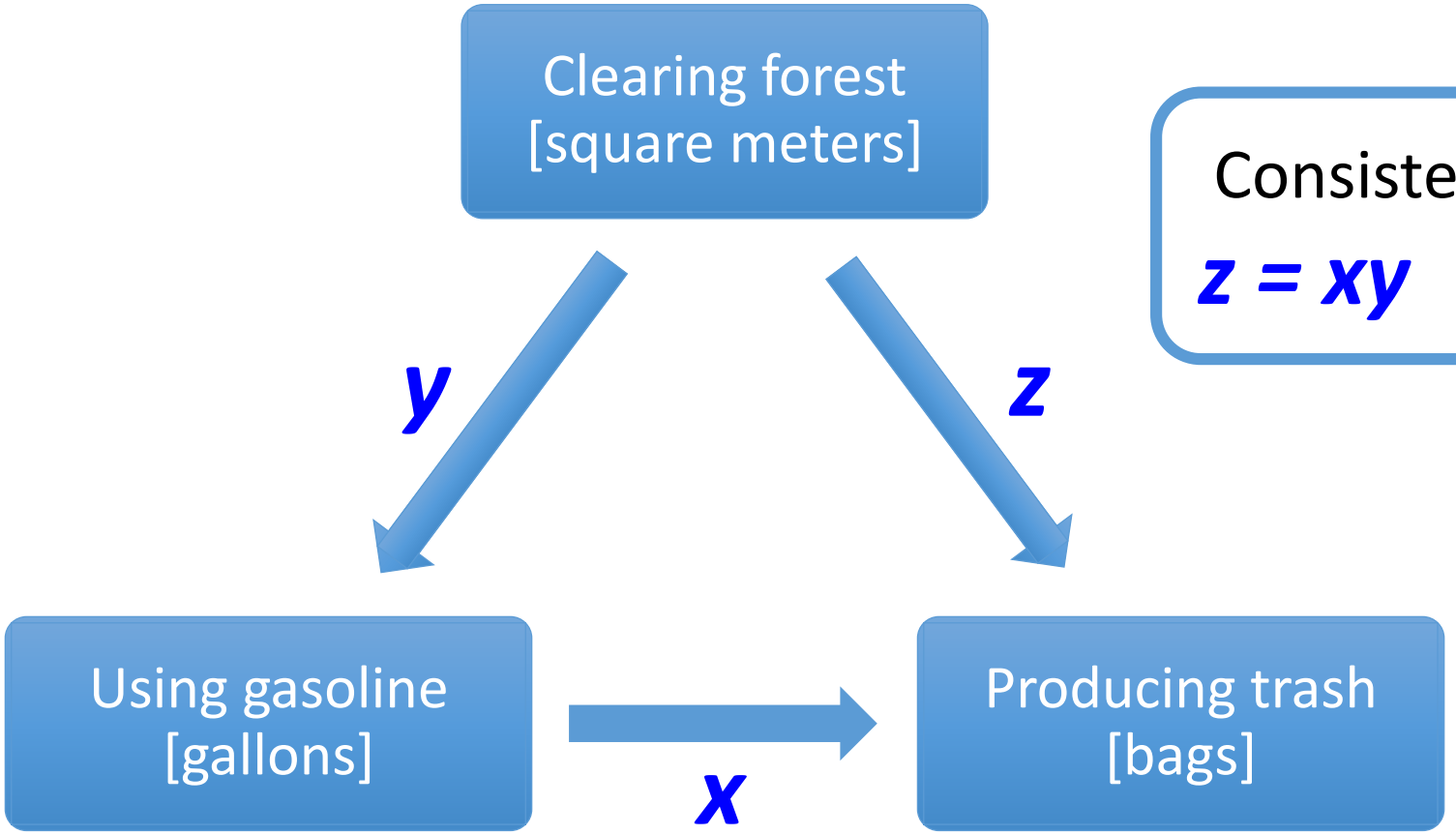
One Approach: Let's Vote!



- 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

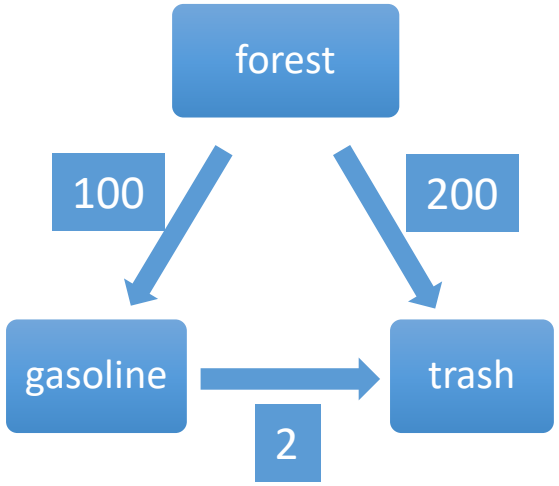
Consistency of tradeoffs

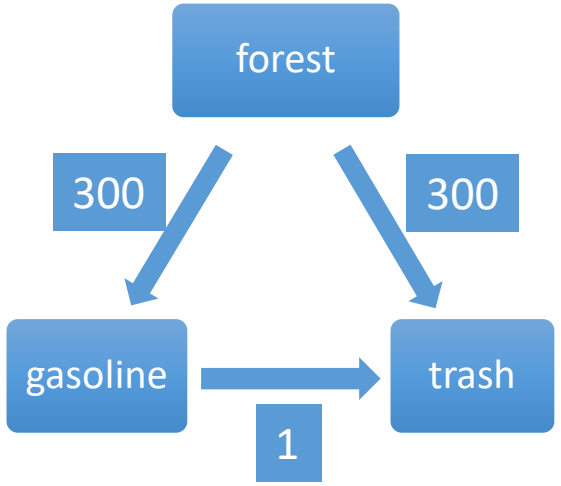


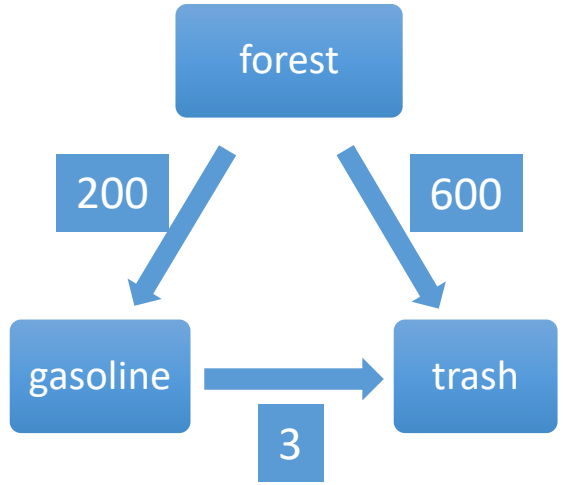
Consistency:

$z = xy$

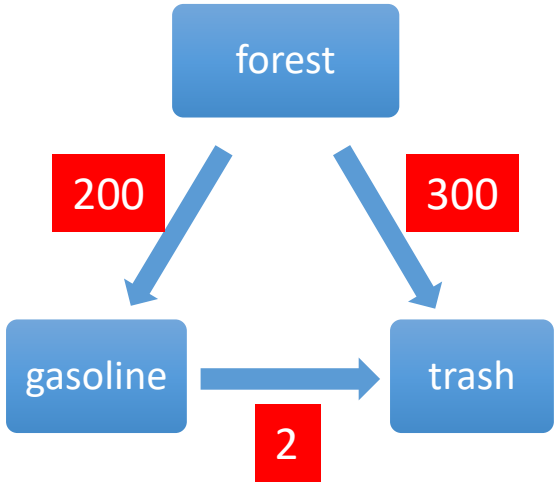
A paradox





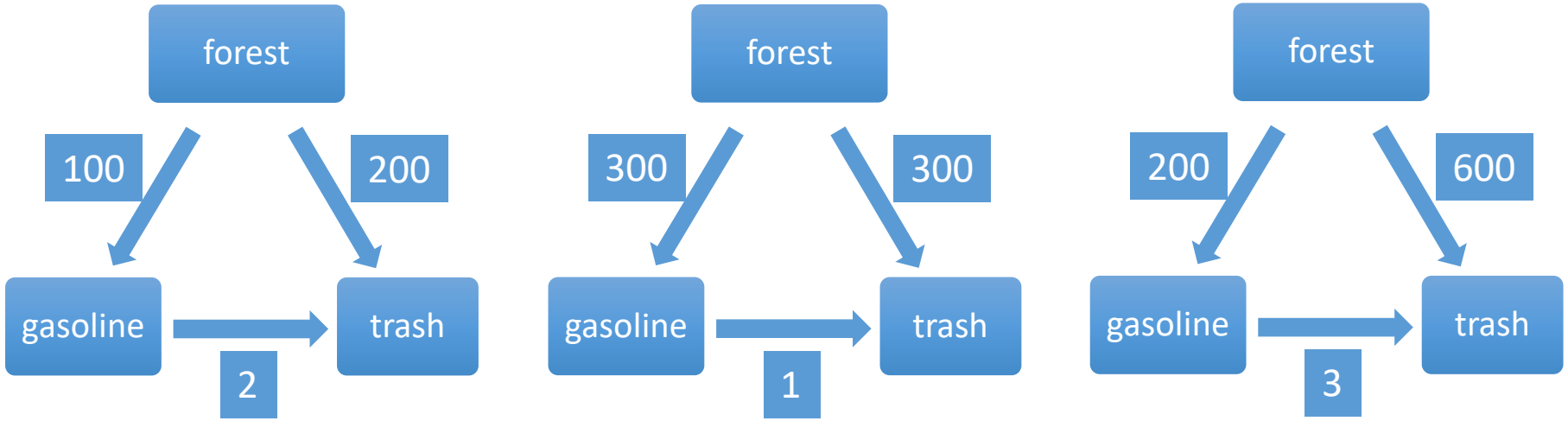


Just taking
medians
pairwise results
in inconsistency

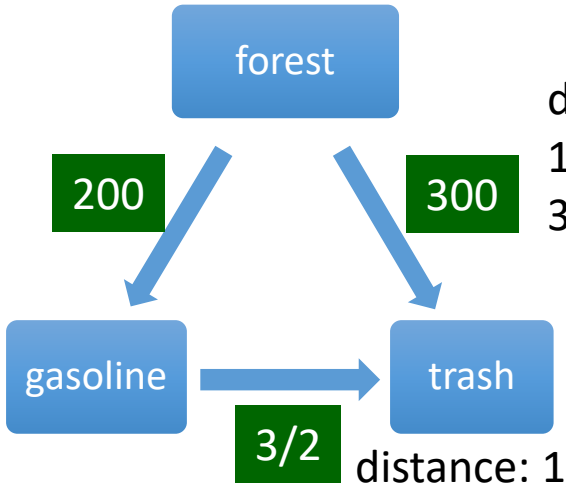


A first attempt at a rule satisfying consistency

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



distance:
100 to v_1
100 to v_2



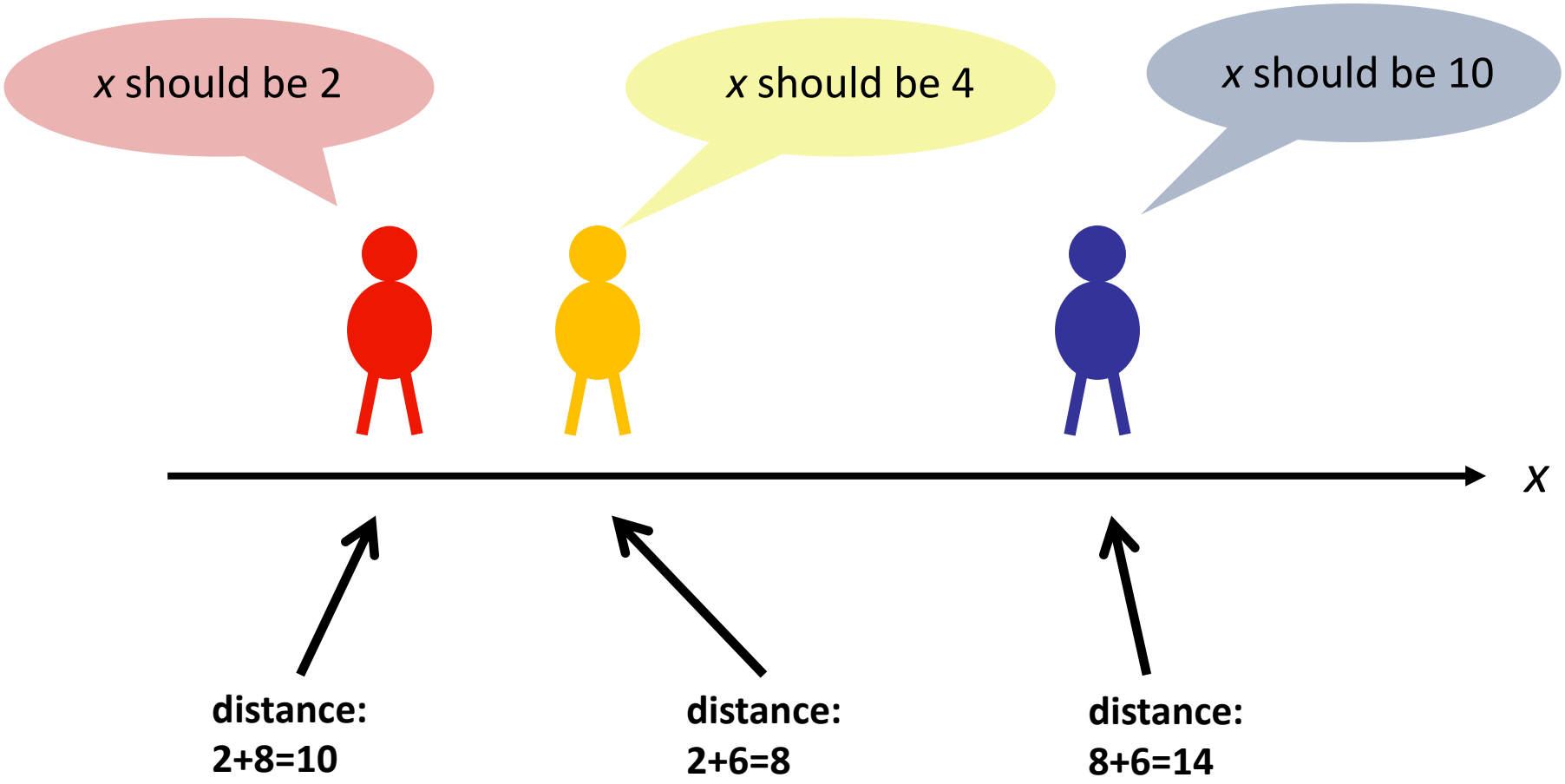
distance:
100 to v_1
300 to v_3

total distance: 602.5 (minimum)

distance: 1/2 to v_1,1/2 to v_2, 3/2 to v_3

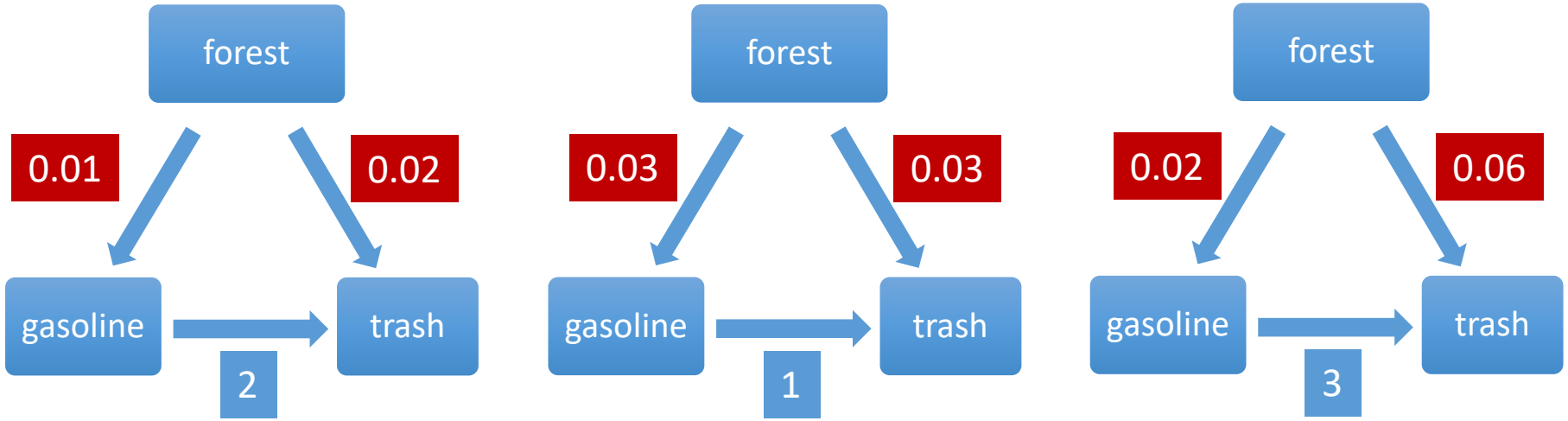
A nice property

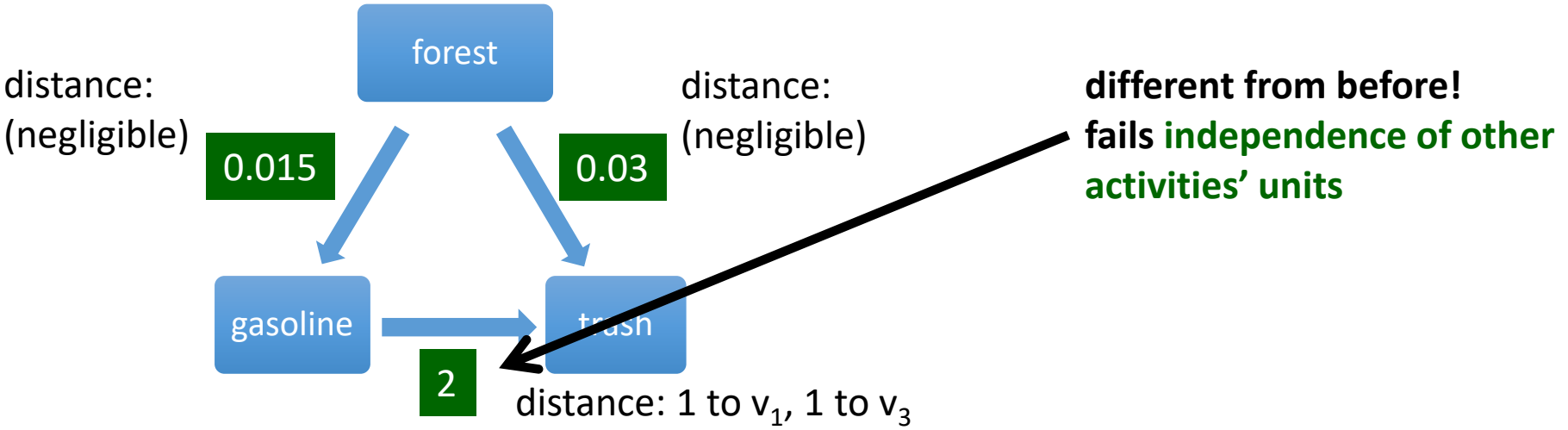
- This rule agrees with the median when there are only two activities!



Not all is rosy, part 1

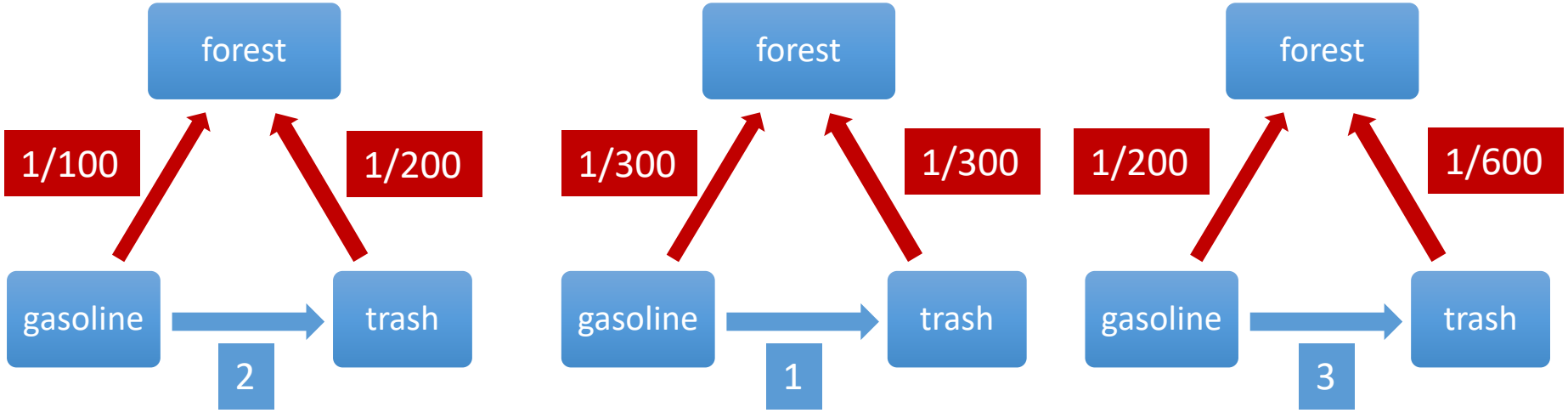
- What if we change units? Say forest from m^2 to cm^2 (divide by 10,000)



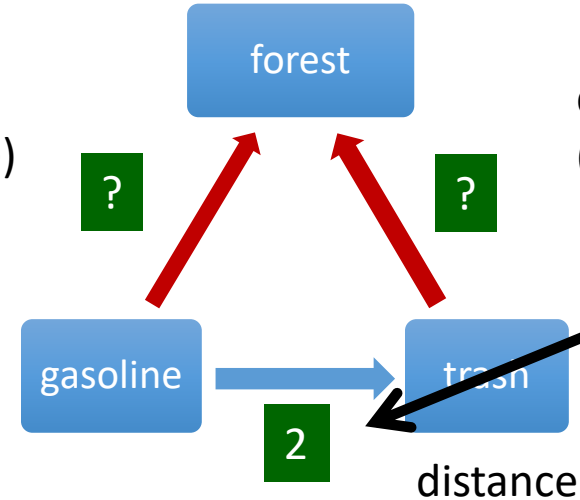


Not all is rosy, part 2

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



distance: (negligible)



distance: 1 to v_1, 1 to v_3

distance: (negligible)

different from before!
fails independence of other edges' directions

Summarizing

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

  $\Sigma_i \Sigma_{a,b} \mid t_{a,b} - t_{a,b,i} \mid$
- Upsides:
  - Coincides with median for 2 activities
- Downsides:
  - Dependence on choice of units:

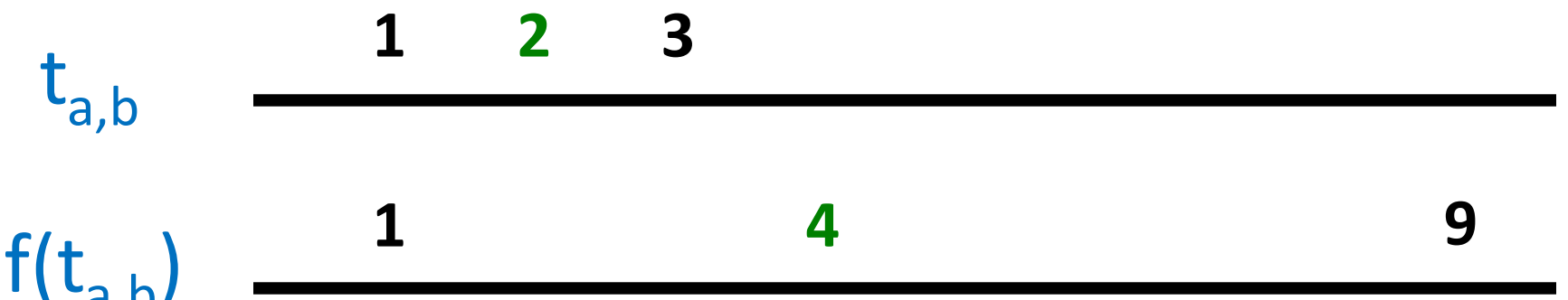
    $\mid t_{a,b} - t_{a,b,i} \mid \neq \mid 2t_{a,b} - 2t_{a,b,i} \mid$
  - Dependence on direction of edges:

    $\mid t_{a,b} - t_{a,b,i} \mid \neq \mid 1/t_{a,b} - 1/t_{a,b,i} \mid$
  - 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

  $\Sigma_i \Sigma_{a,b} \mid f(t_{a,b}) - f(t_{a,b,i}) \mid$
- Still coincides with median for 2 activities!
- **Theorem:** These are the **only** rules satisfying this property, agent separability, and edge separability



An MLE justification

- Suppose probability of tradeoff profile $\{t_i\}$ given true aggregate tradeoff t is

  $\prod_i \prod_{a,b} \exp\{-| f(t_{a,b}) - f(t_{a,b,i}) |\}$

- Then $\arg\max_t \prod_i \prod_{a,b} \exp\{-| f(t_{a,b}) - f(t_{a,b,i}) |\} =$

  $\arg\max_t \log \prod_i \prod_{a,b} \exp\{-| f(t_{a,b}) - f(t_{a,b,i}) |\} =$

  $\arg\max_t \Sigma_i \Sigma_{a,b} -| f(t_{a,b}) - f(t_{a,b,i}) | =$

  $\arg\min_t \Sigma_i \Sigma_{a,b} | f(t_{a,b}) - f(t_{a,b,i}) |$

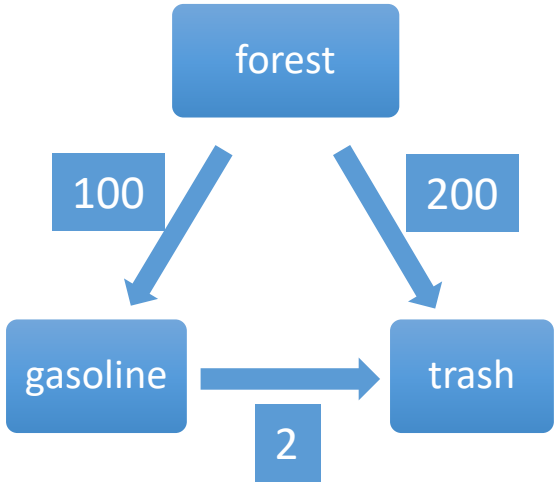
  which is our rule!

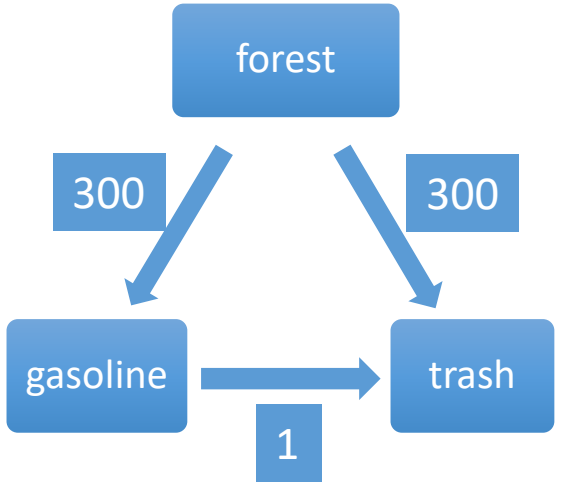
So what's a good f?

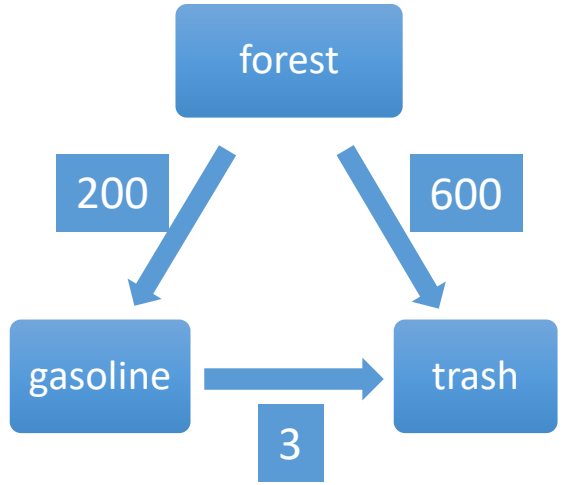
- Intuition: Is the difference between tradeoffs of 1 and 2 the same as between 1000 and 1001, or as between 1000 and 2000?
- So how about f(x)=log(x)?
  - (Say, base e – remember $\log_a(x)=\log_b(x)/\log_b(a)$)

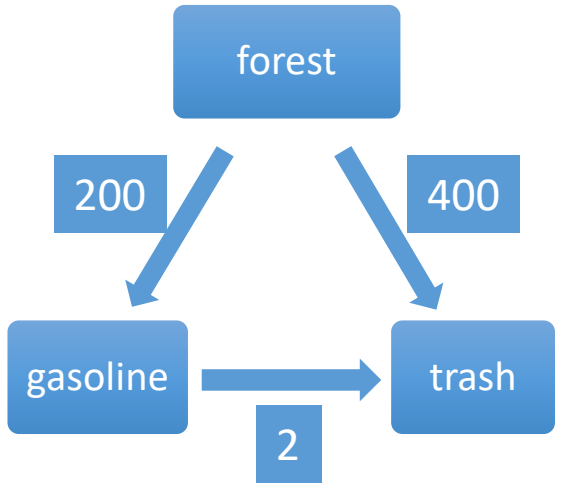
$t_{a,b}$	**1 2**	**1000**	**2000**
$\ln(t_{a,b})$	**ln(1) ln(2)**	**ln(1000)**	**ln(2000)**
	0 0.69	6.91	7.60

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

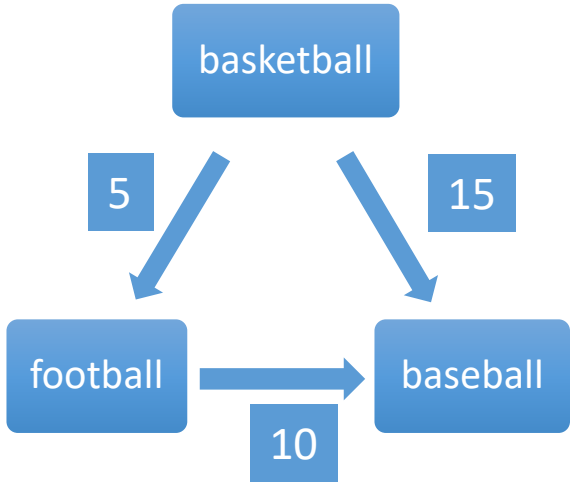
$\log(xy) = \log(z)$

is equivalent to

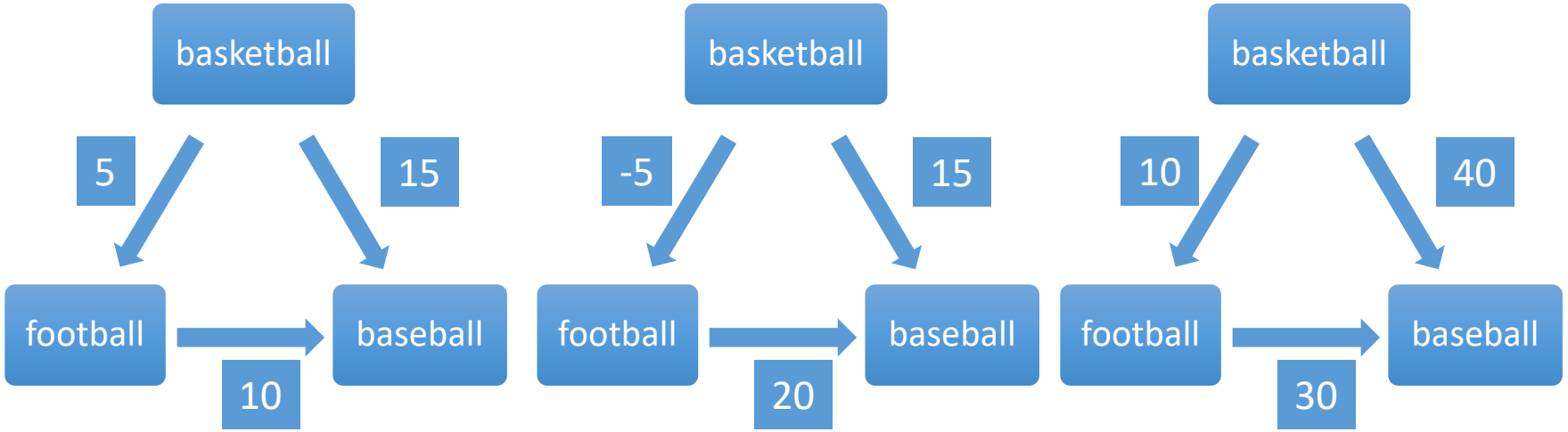
$\log(x) + \log(y) = \log(z)$

An additive variant

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

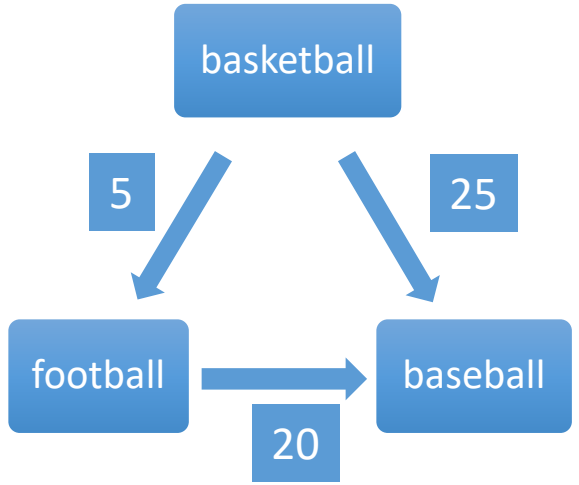


Aggregation in the additive variant



Natural objective:

minimize $\Sigma_i \Sigma_{a,b} d_{a,b,i}$ 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., $d_{a,b,i} = |q_a - q_b - t_{a,b,i}|$

$$\text{minimize } \Sigma_i \Sigma_{a,b} d_{a,b,i}$$

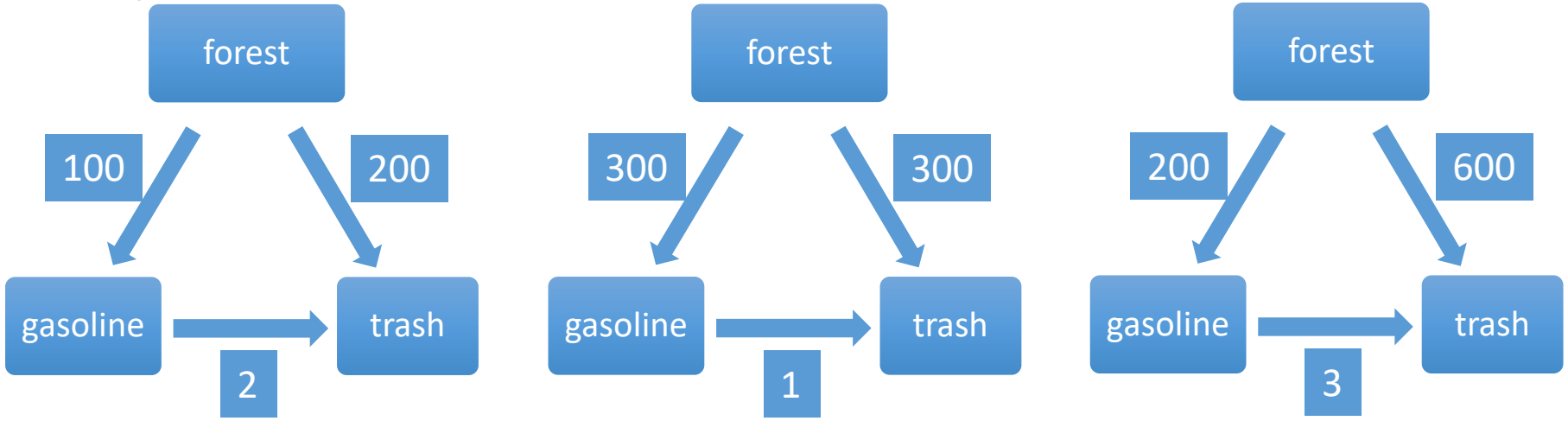
subject to

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

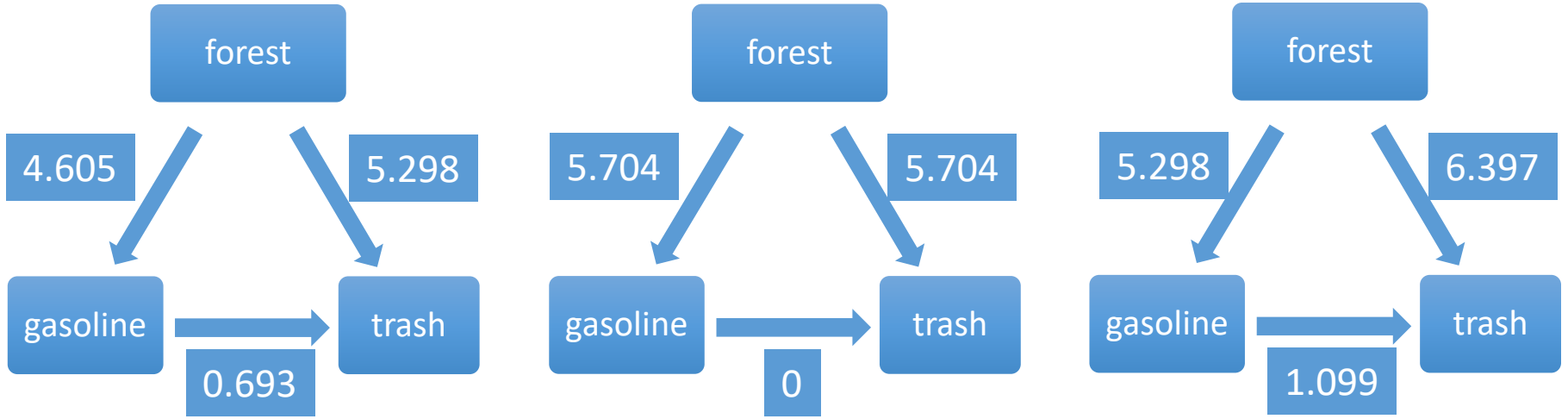
$$\text{for all } a,b,i: d_{a,b,i} \geq t_{a,b,i} - q_a + q_b$$

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

Applying this to the logarithmic rule in the multiplicative variant

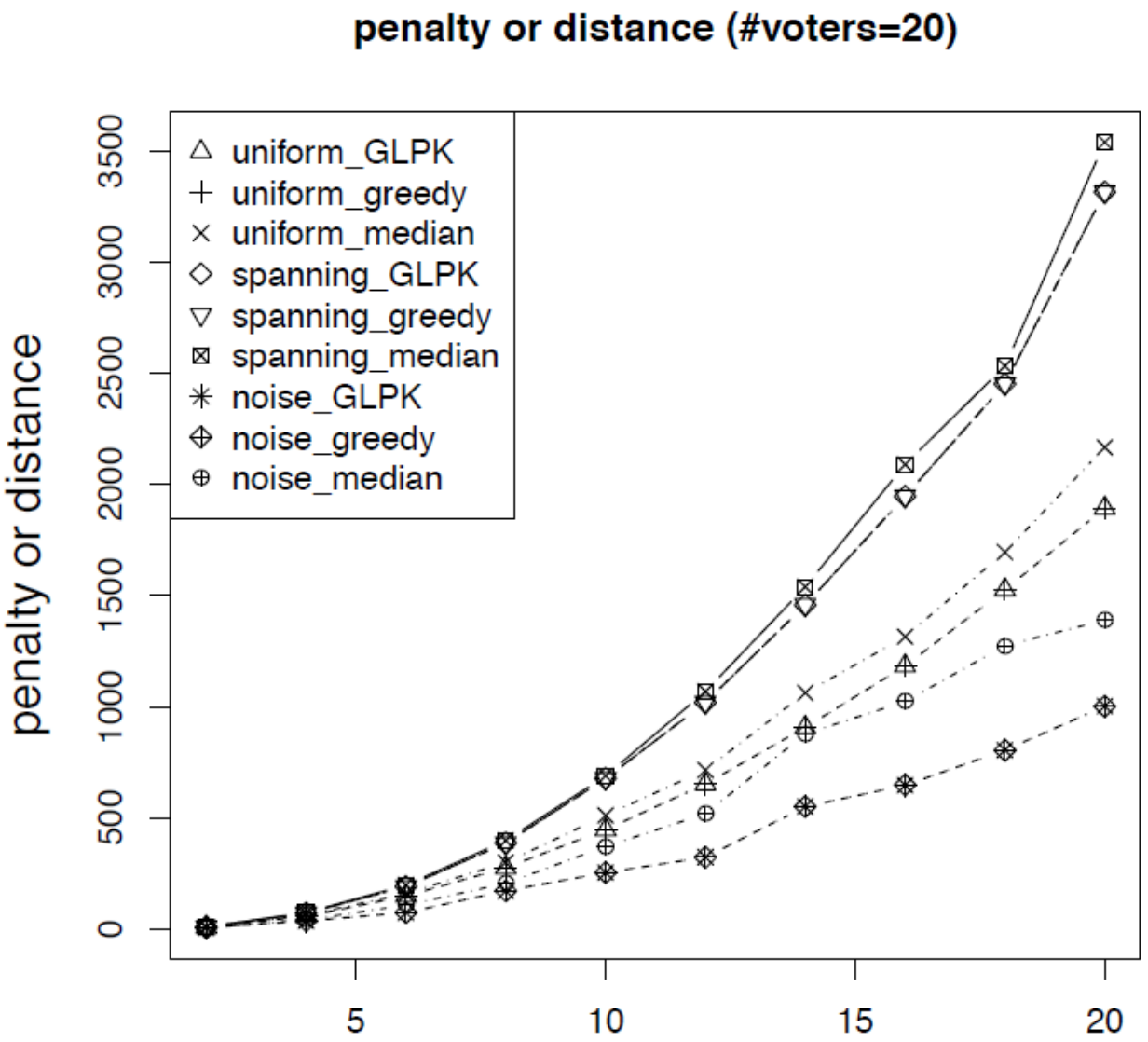


Just take logarithms on the edges, solve the additive variant, and exponentiate back



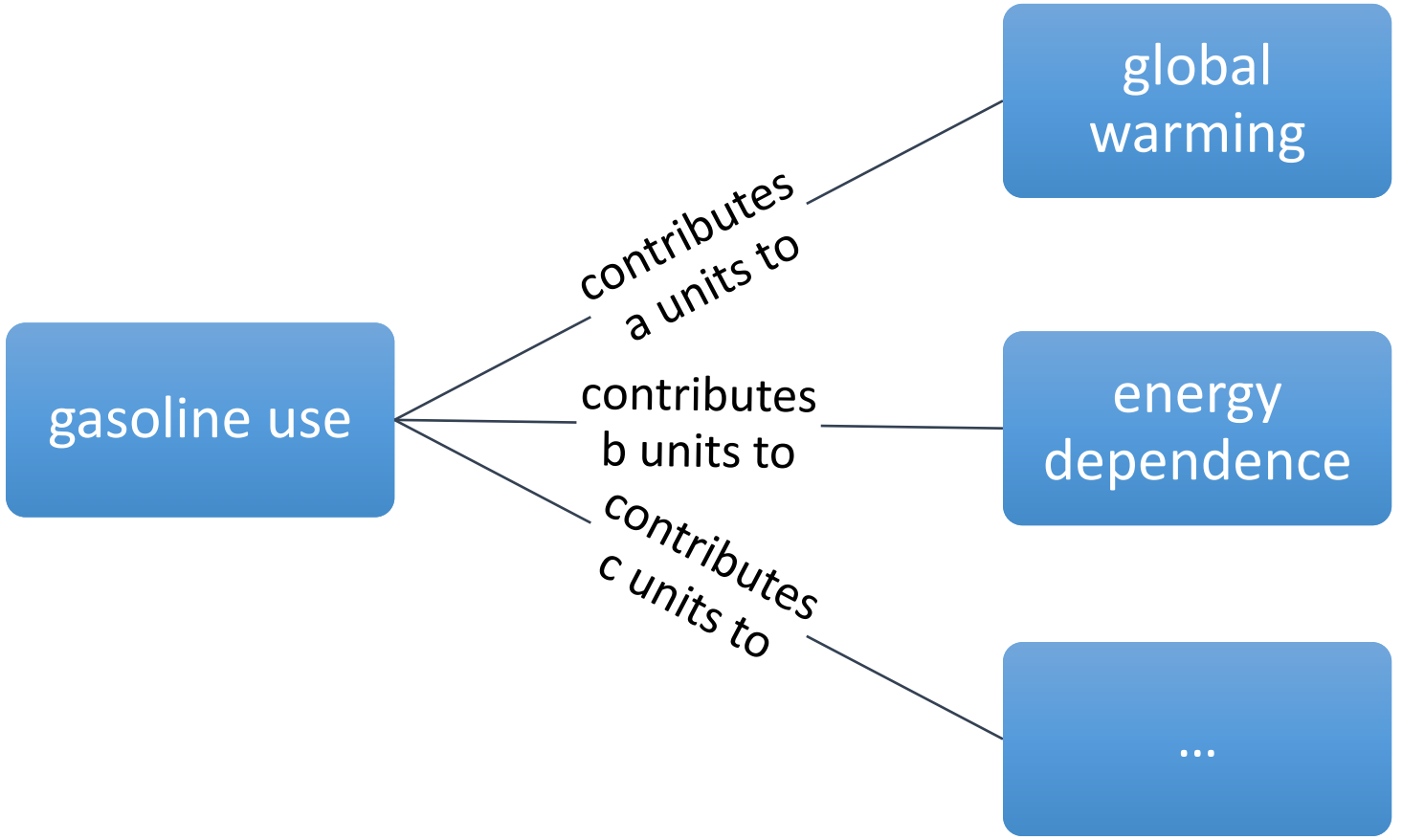
A simpler algorithm (hill climbing / greedy)

- 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 (#voters)*(#activities-1) implied votes on it
- Continue until convergence (possibly to local optimum)

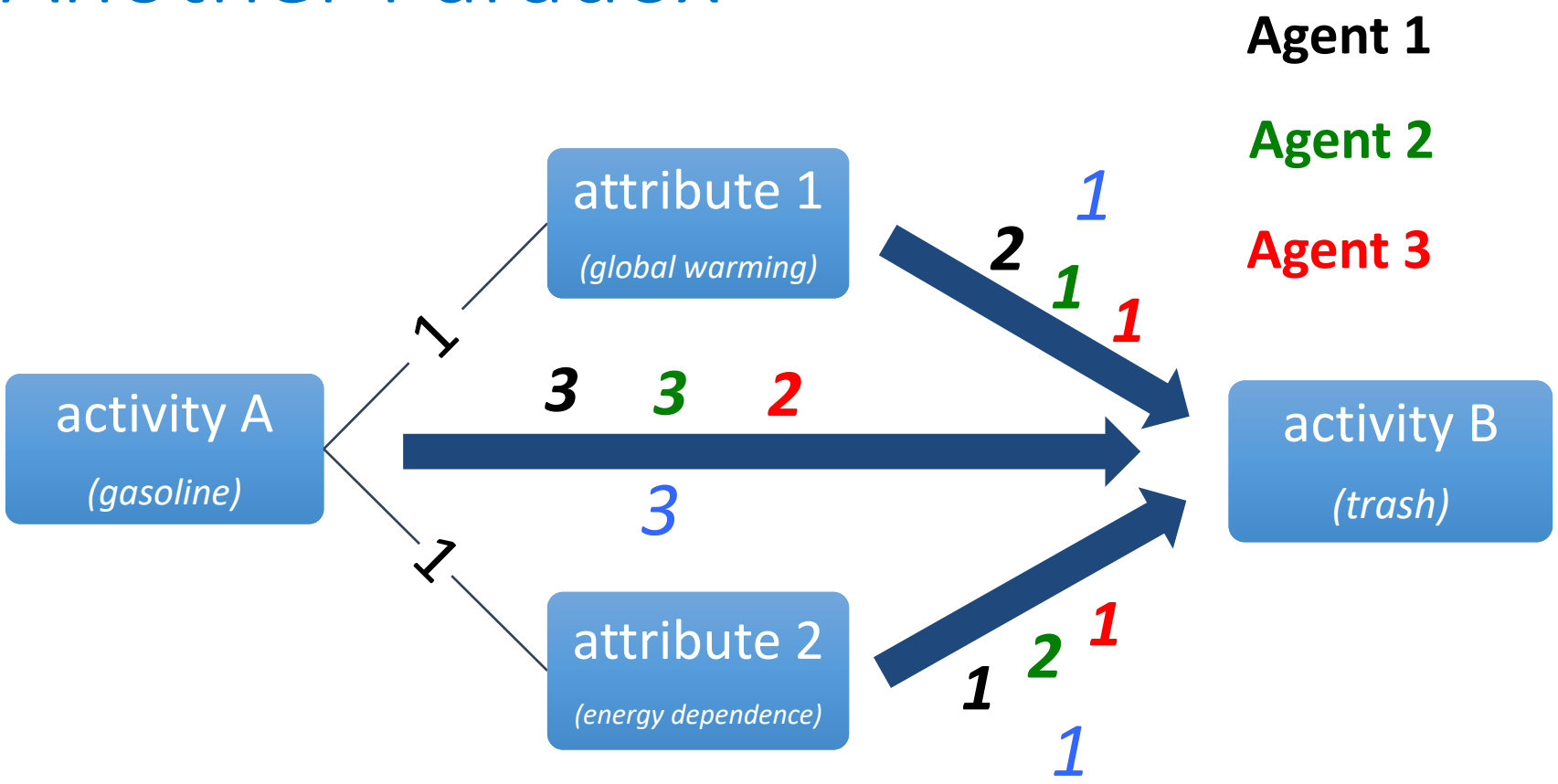


Decomposition

- Idea: Break down activities to relevant attributes



Another Paradox



aggregation on attribute level ≠ 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 to1 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

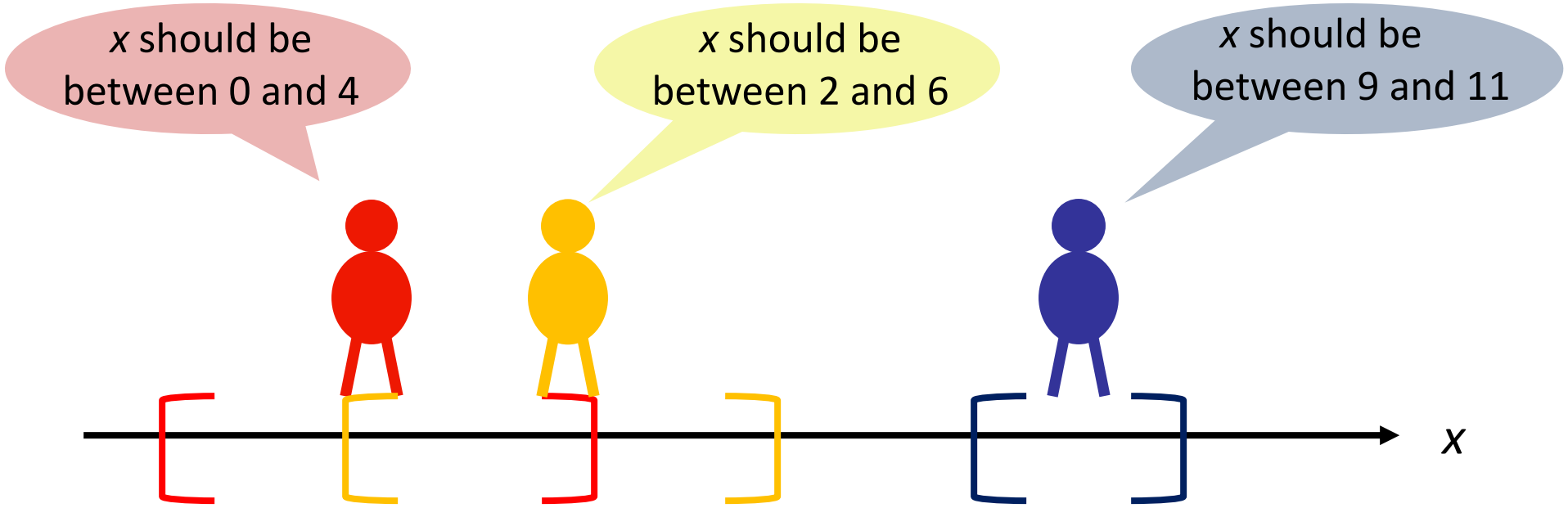
- *Definition:* Let agent a's most-preferred value be p_a.

  Let p and p' satisfy:

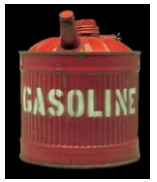
  - $p' \leq p \leq p_a$, or $p_a \leq p \leq p'$

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

Perhaps more reasonable…



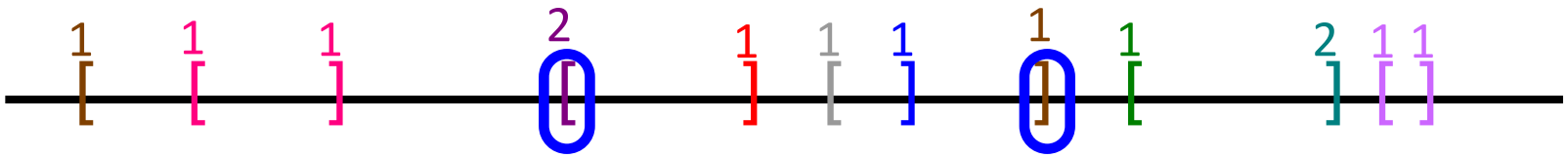
- E.g., due to missing information or plain uncertainty

1  = *x* 

- 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'