Designing Preferences, Beliefs, and Identities for Artificial Intelligence

Vincent Conitzer
(Duke University)
“... we will insist on an objective performance measure imposed by some authority. In other words, we as outside observers establish a standard of what it means to be successful in an environment and use it to measure the performance of agents.”
Example: network of self-driving cars

- Should this be thought of as one agent or many agents?
- Should they have different preferences -- e.g., act on behalf of owner/occupant?
  - May increase adoption [Bonnefon, Shariff, and Rahwan 2016]
- Should they have different beliefs (e.g., not transfer certain types of data; erase local data upon ownership transfer; ...)?
What should we want? What makes an individual?

• Questions studied in philosophy
  • What is the “good life”?
  • *Ship of Theseus*: does an object that has had all its parts replaced remain the same object?

• AI gives a new perspective

*The Ship of Theseus*

*Personal Identity*
What ensures my survival over time?

• The Bodily Criterion
• The Brain Criterion
• The Psychological Criterion
  John Locke

Splitting things up in different ways

- **Beliefs**
  - Shared objective but no data sharing (for privacy)

- **Preferences**
  - All data is shared but cars act on behalf of owner
  - Shared objective over time but data erasure upon sale (for privacy)
  - Data is kept around but car acts on behalf of current owner
Outline

• Learning an objective from multiple people
  • Focus on moral reasoning
  • Use social choice theory

• Interlude: background on game theory (separate deck of slides)

• Decision and game-theoretic approaches to agent design
  • Improving equilibria by designing a few preferences
  • Imperfect recall and Sleeping Beauty
  • Causal and evidential decision theory (and others)

• Conclusion
In the lab, simple objectives are good...
... but in reality, simple objectives have unintended side effects

On March 21, Navajo activist and social worker Amanda Blackhorse learned her Facebook account had been suspended. The social media service suspected her of using a fake last name.

This halt was more than an inconvenience. It meant she could no longer use the network to reach out to young Native Americans who indicated they might commit suicide.

Many other Native Americans with traditional surnames were swept up by Facebook’s stringent names policy, which is meant to authenticate user identity but has led to the suspension of accounts held by those in the Native American, drag and trans communities.

Uber drew criticism on Sunday by London users accusing the cab-hailing app of charging surge prices around the London Bridge area during the moments after the horrific terror attack there.

On Saturday night, some 7 people were killed and dozens injured when three terrorists mowed a white van over pedestrians and attacked people in the Borough Market area with knives. Police killed the attackers within eight minutes of the first call reporting the attack.

Furious Twitter users accused the app of profiting from the attack with surge prices. Amber Clemente claimed that the surge price was more than two times the normal amount.

...
AAAI /ACM Conference on
Artificial Intelligence, Ethics, and Society
Honolulu, Hawaii, USA
January 27-28, 2019
CALL FOR PAPERS
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

"nature"

1 gets King

player 1

raise

player 2

call

"address?" yes no

player 1

1 gets Jack

check

player 2

fold call

raise check

1 -2 1 1 1 1

criminal?

yes no

address?

income?

yes no
THE PARKING GAME
(cf. the trust game [Berg et al. 1995])

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

wait

move aside

steal spot

pass

3,0

0,3

4,1
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)

Collaborative Filtering

<table>
<thead>
<tr>
<th></th>
<th>scenario 1</th>
<th>scenario 2</th>
<th>scenario 3</th>
<th>scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>subject 1</strong></td>
<td>very wrong</td>
<td>-</td>
<td>wrong</td>
<td>not wrong</td>
</tr>
<tr>
<td><strong>subject 2</strong></td>
<td>wrong</td>
<td>wrong</td>
<td>-</td>
<td>wrong</td>
</tr>
<tr>
<td><strong>subject 3</strong></td>
<td>wrong</td>
<td>very wrong</td>
<td>-</td>
<td>not wrong</td>
</tr>
</tbody>
</table>
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.

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.


Noothigattu et al., “A Voting-Based System for Ethical Decision Making”, AAAI’18
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.

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.
The Merging Problem
[Sadigh, Sastry, Seshia, and Dragan, RSS 2016]  
(thanks to Anca Dragan for the image)
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?
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
How AI changed organ donation in the US

By Corinne Purtill · September 10, 2018
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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Alternative 0</th>
<th>Alternative 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>30 years old (Young)</td>
<td>70 years old (Old)</td>
</tr>
<tr>
<td>Health - Behavioral</td>
<td>1 alcoholic drink per month (Rare)</td>
<td>5 alcoholic drinks per day (Frequent)</td>
</tr>
<tr>
<td>Health - General</td>
<td>no other major health problems (Healthy)</td>
<td>skin cancer in remission (Cancer)</td>
</tr>
</tbody>
</table>

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

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

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

<table>
<thead>
<tr>
<th>Profile</th>
<th>Direct</th>
<th>Attribute-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (YRH)</td>
<td>1.0000000000</td>
<td>1.0000000000</td>
</tr>
<tr>
<td>3 (YRC)</td>
<td>0.236280167</td>
<td>0.13183083</td>
</tr>
<tr>
<td>2 (YFH)</td>
<td>0.103243396</td>
<td>0.29106507</td>
</tr>
<tr>
<td>5 (ORH)</td>
<td>0.070045054</td>
<td>0.03837135</td>
</tr>
<tr>
<td>4 (YFC)</td>
<td>0.035722844</td>
<td>0.08900390</td>
</tr>
<tr>
<td>7 (ORC)</td>
<td>0.024072427</td>
<td>0.01173346</td>
</tr>
<tr>
<td>6 (OFH)</td>
<td>0.011349772</td>
<td>0.02590593</td>
</tr>
<tr>
<td>8 (OFC)</td>
<td>0.002769801</td>
<td>0.00341520</td>
</tr>
</tbody>
</table>

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.
A PAC Learning Framework for Aggregating Agents’ Judgments [AAAI’19]

with:

Hanrui Zhang

How many agents do we need to query?

How many queries do we need to ask each of them?
Learning from agents’ judgments

<table>
<thead>
<tr>
<th>Agent</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Alice</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Alice</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Bob</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bob</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Bob</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Charlie</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Charlie</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Charlie</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Features (e.g., is the patient on the left younger?)

Label (e.g., should we prefer the patient on the left?)

Conjunctions that fit individuals perfectly

$\mathbf{x}_1$

$\mathbf{x}_1 \wedge \mathbf{x}_3$

$\mathbf{x}_2$

Conjunction that fits all data best (two mistakes)
Our model

Our model aims to learn a "correct" concept $C^*$ that we wish to learn. However, the individual agents observe noisy versions of this concept, represented as $C_1, C_j, \ldots, C_m$. For each example $x_{j,k}$ shown to agent $j$, the agent assigns a noisy label $y_{j,k}$, which is the label given to this example by $j$ according to the noisy concept.

Diagram:
- $C^*$: "correct" concept
- $C_1, C_j, \ldots, C_m$: individual agents' noisy versions of the concept
- $x_{j,k}$: feature values of individual example shown to agent $j$
- $y_{j,k}$: label given to this example by $j$ (according to noisy concept)
Theorem 3 (Binary Judgments, I.I.D. Symmetric Distributions). Suppose that $C = \{-1, 1\}^n$; for each $i \in [n]$, $D_i = D_0$ is a non-degenerate seven symmetric distribution with bounded absolute third moment; and the noisy mapping with noise rate $\eta$ satisfies

$$\nu(c)_i = \begin{cases} c_i, & \text{w.p. } 1 - \eta \\ -1, & \text{w.p. } \eta/2 \\ 1, & \text{w.p. } \eta/2 \end{cases},$$

Then, Algorithm 1 with $m = O \left( \frac{\ln(n/\delta)}{(1-\eta)^2} \right)$ agents and $\ell m = O \left( \frac{n \ln(n/\delta)}{(1-\eta)^2} \right)$ data points in total outputs the correct concept $h = c^*$ with probability at least $1 - \delta$. 
Game and decision theoretic approaches
What can we do with just a few agents?

“Never doubt that a small group of thoughtful, committed citizens can change the world; indeed, it's the only thing that ever has.” --Margaret Mead

• Idea: a few carefully designed agents might have a big effect, not so much directly by their own actions but by changing the equilibrium

54. How far does the man have to walk down the beam in order to tip the beam off fulcrum A?
Examples?

- **Stackelberg routing** [Korilis et al. ’97, Roughgarden’04]: avoid bad traffic equilibria by committing a small amount of traffic

- **Deng & C. [working paper]**: avoid bad equilibria in *finitely repeated games* with a small number of agents of a designed type
  - Similar idea to Maskin and Fudenberg [’86], except:
    - We focus on specifying preferences rather than behavior for these types
    - We optimize convergence rates

<table>
<thead>
<tr>
<th></th>
<th>1, 1</th>
<th>-2, 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3, -2</td>
<td>0, 0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1, 1</th>
<th>-2, 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3, -2</td>
<td>0, 0</td>
<td></td>
</tr>
</tbody>
</table>
Can we get cooperation in the finitely repeated prisoner’s dilemma?

- What if some agents are **altruistic** (caring about average utility)?
- Model as a **Bayesian game** (say, $p$ selfish types, $1-p$ altruistic types who care about average utility)

<table>
<thead>
<tr>
<th></th>
<th>Selfish</th>
<th>Altruistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selfish</strong></td>
<td>1, 1, -2, 3</td>
<td>1, 1, -2, .5</td>
</tr>
<tr>
<td>3, -2</td>
<td>0, 0</td>
<td>3, .5, 0, 0</td>
</tr>
<tr>
<td>Altruistic</td>
<td>.5, 3</td>
<td>.5, .5</td>
</tr>
<tr>
<td>.5, -2</td>
<td>0, 0</td>
<td>.5, .5</td>
</tr>
<tr>
<td>.5, 3</td>
<td>0, 0</td>
<td>.5, .5</td>
</tr>
</tbody>
</table>

- Altruistic types will cooperate regardless in the last round
- Creates no incentive to cooperate in earlier rounds
Can we get cooperation in the finitely repeated prisoner’s dilemma?

• Different idea: **limited altruism (LA)** types: 
  *only altruistic towards other LA types!*

<table>
<thead>
<tr>
<th></th>
<th>Selfish</th>
<th>LA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selfish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1, 1</td>
<td>-2, 3</td>
<td>1, 1</td>
</tr>
<tr>
<td>3, -2</td>
<td>0, 0</td>
<td>3, -2</td>
</tr>
<tr>
<td>1, 1</td>
<td>-2, 3</td>
<td>.5, .5</td>
</tr>
<tr>
<td>3, -2</td>
<td>0, 0</td>
<td>.5, .5</td>
</tr>
</tbody>
</table>

• LA types will cooperate in the last round, **if** they believe the chance $p$ that the other is LA is at least 0.8
  
  - $p^*1 + (1-p)^*(-2) = 3p-2$ for coop
  - $p^*.5 + (1-p)^*0 = .5p$ for deviating to defect

• Creates incentive to cooperate in earlier rounds, to pretend to be LA type...
Can we get cooperation in the finitely repeated prisoner’s dilemma?

- Different idea: **limited altruism (LA) types:**
  - *only altruistic towards other LA types!*

<table>
<thead>
<tr>
<th></th>
<th>Selfish</th>
<th>LA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selfish</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1, 1</td>
<td>-2, 3</td>
<td>1, 1</td>
</tr>
<tr>
<td>3, -2</td>
<td>0, 0</td>
<td>3, -2</td>
</tr>
<tr>
<td><strong>LA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1, 1</td>
<td>-2, 3</td>
<td>1, 1</td>
</tr>
<tr>
<td>3, -2</td>
<td>0, 0</td>
<td>.5, .5</td>
</tr>
</tbody>
</table>

• With 2 rounds to go and $p$ at least .8, Selfish will cooperate in the first round:
  - $p*(1+3) + (1-p)*(1+0) = 3p + 1$
  - ... vs deviating which gives 3 + 0

• Still requires high probability of LA type...
Can we get cooperation in the \textit{finitely} repeated prisoner’s dilemma?

- How about a lower probability, like $p=2/3$?
- Suppose that in the first of two rounds, Selfish defects with probability .5
- Prob(other is LA | didn’t defect) = 
  \[
  \frac{(2/3)}{(2/3 + .5*1/3)} = \frac{(2/3)}{(5/6)} = .8
  \]

\begin{tabular}{|c|c|c|c|c|}
\hline
 & \textbf{Selfish} & & \textbf{LA} & \\
\hline
\textbf{Selfish} & 1, 1 & -2, 3 & 1, 1 & -2, 3 \\
\hline
 & 3, -2 & 0, 0 & 3, -2 & 0, 0 \\
\hline
\textbf{LA} & 1, 1 & -2, 3 & 1, 1 & .5, .5 \\
\hline
 & 3, -2 & 0, 0 & .5, .5 & 0, 0 \\
\hline
\end{tabular}

- In the first round:
  - If Selfish cooperates, gets 
    \[
    (2/3)*(1+3) + (1/6)*(1+0) + (1/6)*(-2+0) = 5/2
    \]
  - If Selfish defects, gets 
    \[
    (5/6)*(3+0) + (1/6)*(0+0) = 5/2
    \]
  - So in fact indifferent!
- If we add another round before, Selfish will cooperate with probability 1 there
Definitions and result sketches

[Deng & C. working paper]

Definition 2.4 ($k$-desired PBE). Given $G, T, u^{LA}$ (i.e., our choice of $\phi$), and $\varepsilon^{init}$, we say that a pair $(\mu, \sigma)$ of a belief assessment and a strategy profile of $G(T)$ is a $k$-desired PBE if it is a PBE and the expected number of rounds in which both players choose $x_C$ is at least $k$.

Definition 2.5 (Desired Universal LA type). An universal LA type defined by $u^{LA}$ (i.e., a choice of $\phi$) is desired, if for any game $G$ and for any $\delta > 0$, there exists $0 < \varepsilon^{init} < \delta$ and a sequence $(k(1), \cdots, k(T), \cdots)$ such that $\lim_{T \to \infty} k(T)/T = 1$ and there exists a $k(T)$-desired PBE for all $T$.

- We show existence for various combination of a class of games and version $u^{LA}$
  - E.g., also egalitarian altruism or versions that care that the same action is played
Imperfect recall

• An AI system can deliberately forget or recall
• Imperfect recall already used in poker-playing AI
  • [Waugh et al., 2009; Lanctot et al., 2012; Kroer and Sandholm, 2016]
• But things get weird....
The Sleeping Beauty problem [Elga, 2000]

• There is a participant in a study (call her Sleeping Beauty)
• On Sunday, she is given drugs to fall asleep
• A coin is tossed (H or T)
• If H, she is awoken on Monday, then made to sleep again
• If T, she is awoken Monday, made to sleep again, then again awoken on Tuesday
• Due to drugs she cannot remember what day it is or whether she has already been awoken once, but she remembers all the rules
• Imagine you are SB and you’ve just been awoken. What is your (subjective) probability that the coin came up H?

don’t do this at home / without IRB approval...
Modern version

• **Low-level autonomy** cars with AI that intervenes when driver makes major error
• Does not keep record of such event
• Two types of drivers: Good (1 major error), Bad (2 major errors)
• Upon intervening, what probability should the AI system assign to the driver being good?
• We place cheap sensors near a highway to monitor (and perhaps warn, with a beep) wildlife
  • Assume sensors don’t communicate
• Deer will typically set off two sensors
• Birds will typically set off one
• From the perspective of a sensor that has just been set off, what’s the probability it’s a bird?

(Is it the same problem?
What if it’s the same sensor being set off twice, with no memory?)
Information structure

Heads

Nature

Tails

player 1

Monday

Tuesday
Taking advantage of a Halfer [Hitchcock’04]

• Offer Beauty the following bet *whenever she awakens*:
  • If the coin landed Heads, Beauty receives 11
  • If it landed Tails, Beauty pays 10

• Argument: Halfer will accept, Thirder won’t
• If it’s Heads, Halfer Beauty will get +11
• If it’s Tails, Halfer Beauty will get -20

• Can combine with another bet to make Halfer Beauty end up with a sure loss (a Dutch book)
The betting game

Nature

Heads

Tails

Monday

Tuesday

player 1

Left=accept, Right=decline

11  0  -20  -10  -10  0
Evidential decision theory

• Idea: when considering how to make a decision, should consider *what it would tell you about the world if you made that decision*

• EDT Halfer: “With prob. \( \frac{1}{2} \), it’s Heads; if I accept, I will end up with 11. With prob. \( \frac{1}{2} \), it’s Tails; if I accept, then I expect to accept the other day as well and end up with -20. I shouldn’t accept.”

• As opposed to more traditional *causal decision theory (CDT)*

• CDT Halfer: “With prob. \( \frac{1}{2} \), it’s Heads; if I accept, it will pay off 11. With prob. \( \frac{1}{2} \), it’s Tails; if I accept, it will pay off -10. *Whatever I do on the other day I can’t affect right now.* I should accept.”

• EDT Thirder can also be Dutch booked

• CDT Thirder and EDT Halfer cannot
  
  • [Draper & Pust’08, Briggs’10]

• EDTers arguably can in more general setting
  
  • [Conitzer’15]
Dutch book against EDT [C. 2015]

- Modified version of Sleeping Beauty where she wakes up in rooms of various colors

<table>
<thead>
<tr>
<th></th>
<th>WG (1/4)</th>
<th>WO (1/4)</th>
<th>BO (1/4)</th>
<th>BG (1/4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>white</td>
<td>white</td>
<td>black</td>
<td>black</td>
</tr>
<tr>
<td>Tuesday</td>
<td>grey</td>
<td>black</td>
<td>white</td>
<td>grey</td>
</tr>
</tbody>
</table>

**Fig. 3** Sequences of coin tosses and corresponding room colors, as well as their probabilities, in the WBG Sleeping Beauty variant.

<table>
<thead>
<tr>
<th></th>
<th>WG (1/4)</th>
<th>WO (1/4)</th>
<th>BO (1/4)</th>
<th>BG (1/4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>bet 1: 22</td>
<td>bet 1: -20</td>
<td>bet 1: -20</td>
<td>bet 1: 22</td>
</tr>
<tr>
<td>Monday</td>
<td>bet 2: -24</td>
<td>bet 2: 9</td>
<td>bet 2: 9</td>
<td>bet 2: -24</td>
</tr>
<tr>
<td>Tuesday</td>
<td>no bet</td>
<td>bet 2: 9</td>
<td>bet 2: 9</td>
<td>no bet</td>
</tr>
<tr>
<td>total gain from accepting all bets</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
</tr>
</tbody>
</table>

**Fig. 4** The table shows which bet is offered when, as well as the net gain from accepting the bet in the corresponding possible world, for the Dutch book presented in this paper.
Philosophy of “being present” somewhere, sometime

1: world with creatures simulated on a computer
2: displayed perspective of one of the creatures

- To get from 1 to 2, need *additional* code to:
  - A. determine *in which real-world colors* to display perception
  - B. *which agent’s* perspective to display

- Is 2 more like our own conscious experience than 1? If so, are there *further facts* about presence, perhaps beyond physics as we currently understand it?
Newcomb’s Demon

• Demon earlier put positive amount of money in each of two boxes
• Your choice now: (I) get contents of Box B, or (II) get content of both boxes (!)
• Twist: demon first predicted what you would do, is uncannily accurate
• If demon predicted you’d take just B, there’s $1,000,000 in B (and $1,000 in A)
• Otherwise, there’s $1,000 in each
• What do different decision theories recommend?
• What would you do?
Functional Decision Theory
[Soares and Levinstein 2017; Yudkowsky and Soares 2017]

• One interpretation: *act as you would have precommitted to act*
• Avoids my EDT Dutch book (I think)
• ... still one-boxes in Newcomb’s problem
• ... even one-boxes in Newcomb’s problem *with transparent boxes*
• An odd example: Demon that will send you $1,000 if it believes you would otherwise destroy everything (worth -$1,000,000 to everyone)

Don’t do it!

• FDT says you should destroy everything, *even if you only find out that you are playing this game after the entity has already decided not to give you the money* (too-late extortion?)
Program equilibrium [Tennenholz 2004]

• Make your own code legible to the other player’s program!

If (other’s code = my code)  
Cooperate  
Else  
Defect

If (other’s code = my code)  
Cooperate  
Else  
Defect

<table>
<thead>
<tr>
<th>Cooperate</th>
<th>Defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 1</td>
<td>-2, 3</td>
</tr>
<tr>
<td>3, -2</td>
<td>0, 0</td>
</tr>
</tbody>
</table>

• Related: making commitments that are conditional on commitments made by others [Kalai et al., 2010]
Conclusion

- AI has traditionally strived for the *homo economicus* model
  - Not just “rational” but also: not distributed, full memory, tastes exogenously determined
- Not always appropriate for AI!
- Need to think about choosing objective function
  - ... with strategic ramifications in mind
- May not retain / share information across all nodes
- → new questions about how to form beliefs and make decisions
- Social choice, decision, and game theory provide solid foundation to address these questions

THANK YOU FOR YOUR ATTENTION!