AI Agents May Cooperate Better If They Don’t Resemble Us

Vincent Conitzer

If I tailgate you, will your occupant take back control and pull over?

What makes you think I would tell you?

You just did. Better move aside now.

You’re bluffing.

Are you willing to take that chance?

Early blue sky paper:

Also see “Cooperative AI” community https://www.cooperativeai.com/ or our new lab at CMU!
Russell and Norvig’s “AI: A Modern Approach”

“... 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.”

Figure 2.12  A complete utility-based agent.
AI Alignment

THE ALIGNMENT PROBLEM
Machine Learning and Human Values
BRIAN CHRISTIAN
Best-Selling Author, Algorithms to Live By

Fourth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society
A virtual conference
May 19-21, 2021

Institute for Ethics in AI
Oxford leading the way in AI ethics

Stanford University
One Hundred Year Study on Artificial Intelligence (AI100)
Even almost perfectly aligned agents can perform horribly in equilibrium

- Two agents each provide part of a service, each choose quality $q_i$
- **Overall quality** determined by $\min_i q_i$
- Agents care primarily about overall quality, but also have a slight incentive to be the lower one

<table>
<thead>
<tr>
<th></th>
<th>100</th>
<th>90</th>
<th>80</th>
<th>70</th>
<th>60</th>
<th>50</th>
<th>40</th>
<th>30</th>
<th>20</th>
<th>10</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>111, 111</td>
<td>90, 112</td>
<td>80, 102</td>
<td>70, 92</td>
<td>60, 82</td>
<td>50, 72</td>
<td>40, 62</td>
<td>30, 52</td>
<td>20, 42</td>
<td>10, 32</td>
<td>0, 22</td>
</tr>
<tr>
<td>90</td>
<td>112, 90</td>
<td>101, 101</td>
<td>80, 102</td>
<td>70, 92</td>
<td>60, 82</td>
<td>50, 72</td>
<td>40, 62</td>
<td>30, 52</td>
<td>20, 42</td>
<td>10, 32</td>
<td>0, 22</td>
</tr>
<tr>
<td>80</td>
<td>102, 80</td>
<td>102, 80</td>
<td>91, 91</td>
<td>70, 92</td>
<td>60, 82</td>
<td>50, 72</td>
<td>40, 62</td>
<td>30, 52</td>
<td>20, 42</td>
<td>10, 32</td>
<td>0, 22</td>
</tr>
<tr>
<td>70</td>
<td>92, 70</td>
<td>92, 70</td>
<td>92, 70</td>
<td>81, 81</td>
<td>60, 82</td>
<td>50, 72</td>
<td>40, 62</td>
<td>30, 52</td>
<td>20, 42</td>
<td>10, 32</td>
<td>0, 22</td>
</tr>
<tr>
<td>60</td>
<td>82, 60</td>
<td>82, 60</td>
<td>82, 60</td>
<td>82, 60</td>
<td>71, 71</td>
<td>50, 72</td>
<td>40, 62</td>
<td>30, 52</td>
<td>20, 42</td>
<td>10, 32</td>
<td>0, 22</td>
</tr>
<tr>
<td>50</td>
<td>72, 50</td>
<td>72, 50</td>
<td>72, 50</td>
<td>72, 50</td>
<td>72, 50</td>
<td>61, 61</td>
<td>40, 62</td>
<td>30, 52</td>
<td>20, 42</td>
<td>10, 32</td>
<td>0, 22</td>
</tr>
<tr>
<td>40</td>
<td>62, 40</td>
<td>62, 40</td>
<td>62, 40</td>
<td>62, 40</td>
<td>62, 40</td>
<td>62, 40</td>
<td>51, 51</td>
<td>30, 52</td>
<td>20, 42</td>
<td>10, 32</td>
<td>0, 22</td>
</tr>
<tr>
<td>30</td>
<td>52, 30</td>
<td>52, 30</td>
<td>52, 30</td>
<td>52, 30</td>
<td>52, 30</td>
<td>52, 30</td>
<td>52, 30</td>
<td>41, 41</td>
<td>20, 42</td>
<td>10, 32</td>
<td>0, 22</td>
</tr>
<tr>
<td>10</td>
<td>32, 10</td>
<td>32, 10</td>
<td>32, 10</td>
<td>32, 10</td>
<td>32, 10</td>
<td>32, 10</td>
<td>32, 10</td>
<td>32, 10</td>
<td>32, 10</td>
<td>32, 10</td>
<td>21, 21</td>
</tr>
<tr>
<td>0</td>
<td>22, 0</td>
<td>22, 0</td>
<td>22, 0</td>
<td>22, 0</td>
<td>22, 0</td>
<td>22, 0</td>
<td>22, 0</td>
<td>22, 0</td>
<td>22, 0</td>
<td>0, 22</td>
<td>11, 11</td>
</tr>
</tbody>
</table>

*(Cf. Traveler’s Dilemma)*
From *The Atlantic*, “Want to See How Crazy a Bot-Run Market Can Be?”

*By James Fallows*

April 23, 2011
How A Book About Flies Came To Be Priced $24 Million On Amazon

Two booksellers using Amazon's algorithmic pricing to ensure they were generating marginally more revenue than their main competitor ended up pushing the price of a book on evolutionary biology — Peter Lawrence's The Making of a Fly — to $23,698,655.93. [partner id="wireduk"]The book, which was published in 1992, is out of print but is commonly […]

Two booksellers using Amazon's algorithmic pricing to ensure they were generating marginally more revenue than their main competitor ended up pushing the price of a book on evolutionary biology -- Peter Lawrence's *The Making of a Fly* -- to $23,698,655.93.

[partner id="wireduk"]The book, which was published in 1992, is out of print but is commonly used as a reference text by fly experts. A post doc student working in Michael Eisen's lab at UC Berkeley first discovered the pricing glitch when looking to buy a copy. As documented on Eisen's blog, it was discovered that Amazon had 17 copies for sale -- 15 used from $35.54 and two new from $1,730,045.91 (one from seller profnath at that price and a second from bordeebook at $2,198,177.95).

This was assumed to be a mistake, but when Eisen returned to the page the next day, he noticed the price had gone up, with both copies on offer for around $2.8 million. By the end of the day, profnath had raised its price again to $3,536,674.57. He worked out that once a day, profnath set its price to be 0.9983 times the price of the copy offered by bordeebook (keen to undercut its competitor), meanwhile the prices of bordeebook were rising at 1.270589 times the price offered by profnath.
The **May 6, 2010, flash crash**, [1][2][3] also known as the **crash of 2:45** or simply the **flash crash**, was a United States trillion-dollar[4] **stock market crash**, which started at 2:32 p.m. **EDT** and lasted for approximately 36 minutes.[5]:1

Between 2:45:13 and 2:45:27, HFTs traded over 27,000 contracts, which accounted for about 49 percent of the total trading volume, while buying only about 200 additional contracts net.
Prisoner’s Dilemma

<table>
<thead>
<tr>
<th></th>
<th>cooperate</th>
<th>defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>cooperate</td>
<td>2, 2</td>
<td>0, 3</td>
</tr>
<tr>
<td>defect</td>
<td>3, 0</td>
<td>1, 1</td>
</tr>
</tbody>
</table>
How is human cooperation different?

Alicia P. Melis and Dirk Sammann

ABSTRACT

Although cooperation is a widespread phenomenon in nature, human cooperation exceeds that of all other species with regard to the scale and range of cooperative activities. Here we review and

Why We're So Nice: We're Wired to Cooperate

By Natalie Angier

July 23, 2002
When the System Fails
COVID-19 and the Costs of Global Dysfunction
By Stewart Patrick July/August 2020

The chaotic global response to the coronavirus pandemic has tested the faith of even the most ardent internationalists. Most nations, including the world’s most powerful, have turned inward, adopting travel bans, implementing export controls, hoarding or obscuring...

WHY COOPERATION FAILED IN 1914
By STEPHEN VAN EVERA

The essays in this volume explore how three sets of factors affect the degree of cooperation or non-cooperation between states. The first set comprises the "structures of payoffs" that states receive in return for adopting cooperative or noncooperative policies; payoff structures are signified by the rewards and penalties accruing to each state from mutual cooperation (CC); cooperation by one state and "defection" by another (CD and DC); and mutual defection (DD). The second set comprises the "strategic setting" of the international "game"—that is, the rules and conditions under which international relations are conducted. Two aspects of the strategic setting are considered: the size of the "shadow of the future," and the ability of the players to "recognize" past cooperators and defectors, and to distinguish between them. The third set is the number of players in the game, and the influence these...

The Global Climate Talks Ended In Disappointment
One activist group pronounced the conclusions a "pile of shite" and dumped manure outside the meeting hall.

Zahra Hizjii
BuzzFeed News Reporter
J. Lester Feder
BuzzFeed News Reporter

Posted on December 13, 2019, at 10:29 a.m. ET
**Infinitely Repeated Prisoner’s Dilemma**

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

- **Grim trigger** strategy: cooperate as long as everyone cooperates; after that, defect forever. (Equilibrium, if players are somewhat patient.)
- **Folk theorem**: with sufficiently patient players, can always sustain cooperation this way, in any game.
- Folk theorem can be used to efficiently compute equilibria (in infinitely repeated games with sufficiently patient players) [Littman & Stone DSS 2005, Andersen & C., AAAI’13]
Repetitive games on social networks
[Moon & C., IJCAI’15]

- **Common assumption**: an agent’s behavior is instantly observable to all other agents (instant punishment)
- What if there is a delay in knowledge propagation due to network structure?

- **Algorithm** for finding (unique) maximal set of cooperating agents

Catherine Moon
Experiments on random graphs: **Phase transition** between complete cooperation and complete defection

Random graph models:
Erdős–Rényi (ER)
Barabási–Albert preferential-attachment (PA)

Beta = cooperation benefit, delta = discount factor
# Disarmament Game

[Deng & C., AAAI’17, ’18]

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th>KR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>10, -5</td>
<td>0, -4</td>
<td>0, -4</td>
</tr>
<tr>
<td>PF</td>
<td>4, 1</td>
<td>-10, 3</td>
<td>-0.5, -0.5</td>
</tr>
<tr>
<td>PD</td>
<td>-6, 8</td>
<td>-10, 3</td>
<td>-0.5, -0.5</td>
</tr>
</tbody>
</table>
### Disarmament Game

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th>KR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R</strong></td>
<td>10, -5</td>
<td>0, -4</td>
<td>0, -4</td>
</tr>
<tr>
<td><strong>PF</strong></td>
<td>4, 1</td>
<td>-10, 3</td>
<td>-0.5, -0.5</td>
</tr>
<tr>
<td><strong>PD</strong></td>
<td>-6, 8</td>
<td>-9, 3</td>
<td>-0.5, -0.5</td>
</tr>
</tbody>
</table>

Pure Nash equilibria
Pure Stackelberg equilibria (no matter who takes the lead)
Disarmament Game

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th>KR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>10, -5</td>
<td>0, -4</td>
<td>0, -4</td>
</tr>
<tr>
<td>PF</td>
<td>4, 1</td>
<td>-10, 3</td>
<td>-0.5, -0.5</td>
</tr>
<tr>
<td>PD</td>
<td>-6, 8</td>
<td>-10, 3</td>
<td>-0.5, -0.5</td>
</tr>
</tbody>
</table>

**Desired Outcome**
Pareto better than the Nash equilibrium outcome
Multiple-round (pure) commitments

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th>KR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>10, -5</td>
<td>0, -4</td>
<td>0, -4</td>
</tr>
<tr>
<td>PF</td>
<td>4, 1</td>
<td>-10, 3</td>
<td>-0.5, -0.5</td>
</tr>
<tr>
<td>PD</td>
<td>-6, 8</td>
<td>-10, 3</td>
<td>-0.5, -0.5</td>
</tr>
</tbody>
</table>
## Multiple-round (pure) commitments

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>10, -5</td>
<td>0, -4</td>
</tr>
<tr>
<td>PF</td>
<td>4, 1</td>
<td>-0.5, -0.5</td>
</tr>
<tr>
<td>PD</td>
<td>-6, 8</td>
<td>-0.5, -0.5</td>
</tr>
</tbody>
</table>
### Multiple-round (pure) commitments

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R</strong></td>
<td>10, -5</td>
<td>0, -4</td>
</tr>
<tr>
<td><strong>PF</strong></td>
<td>4, 1</td>
<td>-0.5, -0.5</td>
</tr>
<tr>
<td><strong>PD</strong></td>
<td>-6, 8</td>
<td>-0.5, -0.5</td>
</tr>
</tbody>
</table>

*Incentivize Row to commit in the next round*
Multiple-round (pure) commitments

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>10, -5</td>
<td>0, -4</td>
</tr>
<tr>
<td>PF</td>
<td>4, 1</td>
<td>-0.5, -0.5</td>
</tr>
<tr>
<td>PD</td>
<td>-6, 8</td>
<td>-0.5, -0.5</td>
</tr>
</tbody>
</table>
Multiple-round (pure) commitments

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th></th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PF</td>
<td>4, 1</td>
<td></td>
<td>-0.5, -0.5</td>
</tr>
<tr>
<td>PD</td>
<td>-6, 8</td>
<td></td>
<td>-0.5, -0.5</td>
</tr>
</tbody>
</table>
## Multiple-round (pure) commitments

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF</td>
<td>4, 1</td>
<td>-0.5, -0.5</td>
</tr>
<tr>
<td>PD</td>
<td>-6, 8</td>
<td>-0.5, -0.5</td>
</tr>
</tbody>
</table>
Multiple-round (pure) commitments

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th>KR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>10, -5</td>
<td>0, -4</td>
<td>0, -4</td>
</tr>
<tr>
<td>PF</td>
<td>4, 1</td>
<td>-10, 3</td>
<td>-0.5, -0.5</td>
</tr>
<tr>
<td>PD</td>
<td>-6, 8</td>
<td>-10, 3</td>
<td>-0.5, -0.5</td>
</tr>
</tbody>
</table>
Multiple-round (pure) commitments

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th></th>
<th>KR</th>
<th></th>
<th>ST</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>10, -5</td>
<td></td>
<td>0, -4</td>
<td></td>
<td>0, -4</td>
<td></td>
</tr>
<tr>
<td>PF</td>
<td>4, 1</td>
<td></td>
<td>-10, 3</td>
<td></td>
<td>-0.5, -0.5</td>
<td></td>
</tr>
<tr>
<td>PD</td>
<td>-6, 8</td>
<td></td>
<td>-10, 3</td>
<td></td>
<td>-0.5, -0.5</td>
<td></td>
</tr>
</tbody>
</table>

**Fact:** The desired outcome cannot be achieved if Row commits first.

In general, it is an **NP-hard problem** to determine whether an outcome can be reached without creating incentive to deviate from disarmament.
THE PARKING GAME
(cf. the trust game [Berg et al. 1995])

Letchford, C., Jain [2008] define a solution concept capturing this.
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; ...)?
Role assignment [Moon & C., IJCAI’16]

- Two individuals for roles in two committees
- Committee 1

<table>
<thead>
<tr>
<th>Chair</th>
<th>Member sabotaged</th>
<th>Cooperate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selfish</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Cooperate</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

- Committee 2

<table>
<thead>
<tr>
<th>Chair</th>
<th>Member sabotaged</th>
<th>Cooperate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selfish</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Cooperate</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Computation for optimal role assignment

- Problem is NP-hard

- Dynamic programming approach:

- Integer programming approach:
Agents through time

**information** (data, sensor input, inbound communication, ...)

**decisions** (actions, effector use, outbound communication, ...)

---

AI / software (e.g., personal assistant)

---

an idealized human being

---

**instruction**$_1$

**instruction**$_2$

...
What should you do if...

• ... you knew *others could read your code*?

• ... you knew you *were facing someone running the same code*?

• ... you knew you *had been in the same situation before but can’t possibly remember what you did*?
Program equilibrium [Tennenholz 2004]

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

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

• See also: [Fortnow 2009, Kalai et al. 2010, Barasz et al. 2014, Critch 2016, Oesterheld 2018, ...]
Robust program equilibrium [Oesterheld 2018]

• Can we make the equilibrium less fragile?

With probability $\varepsilon$
  Cooperate
Else
  Do what the other program does against this program

<table>
<thead>
<tr>
<th></th>
<th>cooperate</th>
<th>defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>cooperate</td>
<td>2, 2</td>
<td>0, 3</td>
</tr>
<tr>
<td>defect</td>
<td>3, 0</td>
<td>1, 1</td>
</tr>
</tbody>
</table>
Safe Pareto improvements for delegated game playing [AAMAS’21], with Caspar Oesterheld

Delegated game playing

- Representatives are competent at playing games and the original players trust the representatives.
  => Default: aligned delegation
- DL, RL are strictly dominated and therefore never played
- Equilibrium selection problem
  => Pareto-suboptimal outcome (DM, DM) might occur

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>RM</th>
<th>DL</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>-5,-5</td>
<td>2,0</td>
<td>5,-5</td>
<td>5,-5</td>
</tr>
<tr>
<td>RM</td>
<td>0,2</td>
<td>1,1</td>
<td>5,-5</td>
<td>5,-5</td>
</tr>
<tr>
<td>DL</td>
<td>-5,5</td>
<td>-5,5</td>
<td>1,1</td>
<td>2,0</td>
</tr>
<tr>
<td>RL</td>
<td>-5,5</td>
<td>-5,5</td>
<td>0,2</td>
<td>1,1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DL</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>DL</td>
<td>-5,-5</td>
<td>2,0</td>
</tr>
<tr>
<td></td>
<td>(1,1)</td>
<td>(2,0)</td>
</tr>
<tr>
<td>RL</td>
<td>0,2</td>
<td>1,1</td>
</tr>
<tr>
<td></td>
<td>(0,2)</td>
<td>(1,1)</td>
</tr>
</tbody>
</table>

- Each player’s contract says: Play this alternative game if the other player adopts an analogous contract.
- The games are essentially isomorphic.
  - DM ~ DL
  - RM ~ RL
- Safe Pareto improvement on the original game: outcome of new game is better for both players with certainty.
Disarmament revisited: Committing to your first few lines of code

1. With probability 40%, cooperate
2. With probability 40%, cooperate
3. With probability 40%, cooperate
4. With probability 40%, cooperate

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

... 

- E.g., if Blue refuses to add line 2, then Red defects with probability .6, resulting in at most \(.4 \times 3 + .6 \times 1 = 1.8\) for Blue
- “Folk theorem” [Deng & C., AAAI’17, ‘18] that cooperation can always be achieved this way!
Prisoner’s Dilemma against (possibly) a copy

- What if you play against your twin that you always agree with?
- What if you play against your twin that you almost always agree with?

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

related to working paper
[Oesterheld, Demski, C.]

Caspar Oesterheld  Abram Demski
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 would you do?
The lockdown dilemma

• Lockdown is **monotonous**: you forget what happened before, you forget what day it is
• Suppose you know lockdown lasts two days (unrealistic)
• Every morning, you can decide to eat an unhealthy cookie! (or not)
• Eating a cookie will give you +1 utility immediately, but then -3 later the next day
• **But, carpe diem**: you only care about today
• Should you eat the cookie right now?

related to working paper [C.]
Your own choice is evidence...

• ... for what the demon put in the boxes
• ... for whether your twin defects
• ... for whether you eat the cookie on the other day

• *Evidential Decision Theory (EDT)*: When considering how to make a decision, consider how happy you expect to be conditional on taking each option and choose an option that maximizes that

• *Causal Decision Theory (CDT)*: Your decision should focus on what you causally affect
Turning causal decision theorists into money pumps

[Oesterheld and C., Phil. Quarterly]

- **Adversarial Offer:**
  - Demon (really, any good predictor) put $3 into each box it predicted you would not choose
  - Each box costs $1 to open; can open at most one
  - Demon 75% accurate (you have no access to randomization)
  - CDT will choose one box, *knowing that it will regret doing so*
  - Can add earlier **opt-out** step where the demon promises not to make the adversarial offer later, if you pay the demon $0.20 now
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’00]

• 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?
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?

Sunday Monday Tuesday
• 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

Nature

Heads

Monday

player 1

Tails

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

11 0 -20 -10 -10 0

Left=accept, Right= decline
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
  - [C., Synthese’15]
  - ... though we’ve argued against CDT in other work [Oesterheld & C, Phil. Quarterly’21]
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?

See also: [Hare 2007-2010, Valberg 2007, Hellie 2013, Merlo 2016, ...]
Absentminded Driver Problem

Piccione and Rubinstein, 1997

- Driver on monotonous highway wants to take second exit, but exits are indistinguishable and driver is forgetful
- Deterministic (behavioral) strategies are not stable
- Optimal randomized strategy: exit with probability $p$ where $p$ maximizes $4p(1-p) + (1-p)^2 = -3p^2 + 2p + 1$, so $p^* = 1/3$
- What about “from the inside”? P&R analysis: Let $b$ be the belief/credence that we’re at X, and $p$ the probability that we exit. Maximize with respect to $p$: $(1-b)(4p+1(1-p)) + b(4p(1-p) + 1(1-p)^2) = -3bp^2 + (3-b)p + 1$, so $p^* = (3-b) / (6b) = 1/(2b) - 1/6$
- But if $p = 1/3$, then $b = 3/5$, which would give $p^* = 5/6 - 1/6 = 2/3$? So also not stable?
- Resembles EDT reasoning... But not really halving... Shouldn’t $b$ depend on $p$...

Fig. 1. The absent-minded driver problem.

Image from Aumann, Hart, Perry 1997
A different analysis  
[Aumann, Hart, Perry, 1997]

• AHP reason more along thirders / CDT lines:
  • Imagine we normally expect to play $p = 1/3$. Should we deviate **this time only**?
  • If we exit now, get $(3/5)*0 + (2/5)*4 = 8/5$
  • If we continue now, get $(3/5)*((1/3)*4+(2/3)*1) + (2/5)*1 = 8/5$
• So indifferent and willing to randomize (equilibrium)
• **Questions**
• **Joint work with:**

  Scott Emmons  Caspar Oesterheld  Andrew Critch  Stuart Russell

• Does this always work? Yes! (See also Taylor [2016])
• Does some version of EDT work with some version of belief formation?

**Fig. 1.** The absent-minded driver problem.
A challenging example for the evidential decision theorist

• Optimal strategy to commit to is to just go left: \((p_l, p_s, p_r) = (1, 0, 0)\)
• If you’re at an intersection, what does EDT say you should do?
• When considering \((p_l, p_s, p_r) = (1, 0, 0)\), you presumably expect to be at X and get 1 (really just need: no more than 1)
• When considering \((p_l, p_s, p_r) = (0, \frac{1}{2}, \frac{1}{2})\), then say \(b\) is your subjective probability of being at Y
  • Assume: \(b > 0\)
  • Assume: \(b\) is not a function of \(\epsilon\)
• So, expected utility: \(b \times \frac{1}{2} \times (4-\epsilon) + (1-b) \times \frac{1}{4} \times (4-\epsilon) = 1+b-\frac{1}{4}\epsilon-\frac{1}{4}b\epsilon\)
• For sufficiently small \(\epsilon\) this is greater than 1
• Hence EDT suggests \((0, \frac{1}{2}, \frac{1}{2})\) over \((1, 0, 0)\)!
• ... right? ... right?
A way for EDT to get the right answer (+SSA)

- Consider probabilities of **whole trajectories, plus where you are**, under strategy \((0, \frac{1}{2}, \frac{1}{2})\), in a *halving sort of way*
- \(P(XY(4-\varepsilon), @X) = P(XY(4-\varepsilon)) \times P(@X | XY(4-\varepsilon)) = \frac{1}{4} * \frac{1}{2}\)
- \(P(XY(4-\varepsilon), @Y) = P(XY(4-\varepsilon)) \times P(@Y | XY(4-\varepsilon)) = \frac{1}{4} * \frac{1}{2}\)
- Any other trajectory with positive probability gives payoff 0
- So expected utility is \(2 * \frac{1}{4} * \frac{1}{2} * (4-\varepsilon) = 1 - \varepsilon/4\), which is worse than 1, so EDT gets the right answer

*What just happened?*

- Under this way of reasoning, if you tell me that I’m at X, it’s **more likely** that I’m on trajectory X(0) than on one of the XY ones
- \(P(XY(4-\varepsilon), @X) = \frac{1}{4} * \frac{1}{2} ; P(XY(0), @X) = \frac{1}{4} * \frac{1}{2} ; P(X(0), @X) = \frac{1}{2} * 1\)
- So \(P(X(0) | @X) = \frac{1}{2} / (\frac{1}{2} + \frac{1}{4}) = 2/3\) (**not 1/2**)
- Previous slide had **hidden assumption**: *where I am carries no information about my future coin tosses*
Making decisions with imperfect recall
[cf. absentminded driver problem: PR97, AHP97]

• Optimal strategy without recall: go Right with probability \( \frac{5}{8} \). (Outside view.) Follow that.

• You arrive at decision point. What is the probability that you’re there for the first time? (Inside view.)

  \[
  \text{Thirder: in expectation } 1 \text{ first awakening, and } \\
  \frac{(1/2)(5/8)(16/25)}{1-(5/8)(16/25)} = \frac{1}{3} \text{ later awakenings, so probability of first time } = \frac{1}{(4/3)} = \frac{3}{4}
  \]


• **Theorem.** This is always true!

• … but can have other equilibria
Fraction of time replicator dynamic finds **best** solution

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.93</td>
<td>0.81</td>
<td>0.68</td>
<td>0.65</td>
</tr>
<tr>
<td>3</td>
<td>0.81</td>
<td>0.70</td>
<td>0.58</td>
<td>0.46</td>
</tr>
<tr>
<td>4</td>
<td>0.76</td>
<td>0.58</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td>5</td>
<td>0.69</td>
<td>0.43</td>
<td>0.36</td>
<td>0.30</td>
</tr>
</tbody>
</table>

(a) RandomGame

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.58</td>
<td>0.45</td>
<td>0.40</td>
<td>0.33</td>
</tr>
<tr>
<td>3</td>
<td>0.57</td>
<td>0.35</td>
<td>0.29</td>
<td>0.27</td>
</tr>
<tr>
<td>4</td>
<td>0.53</td>
<td>0.37</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td>5</td>
<td>0.51</td>
<td>0.33</td>
<td>0.33</td>
<td>0.24</td>
</tr>
</tbody>
</table>

(b) CoordinationGame

\[ N = \#\text{players (or \#nodes)} \]

\[ A = \#\text{actions per player (or per node)} \]
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?)
Summary

• Game-theoretic failures to cooperate can happen even with almost perfectly aligned agents

• Some ways of getting to cooperation make sense for humans as well...

• ... but there are others that seem more natural for (advanced) AI agents

• Let’s not unnecessarily limit our toolkit!

THANK YOU FOR YOUR ATTENTION!