Quantifying Misalignment Between Agents

SafeAI 2022 workshop @ AAAI, Montreal/Virtual

Aidan Kierans¹, Hananel Hazan², Shiri Dori-Hacohen³
Before we start - Quick Plug

● The National Science Foundation has put out a Request for Information (RFI) for the 2023 Convergence Accelerator - due today 2/28/22

● We are coordinating a response focusing on a theme of AI Safety - could lead to >$9M funding

● Please consider joining! Just 5 min of your time

"The algorithm isn't racist. It's just been trained to prefer brighter colors and higher contrast and therefore people with lighter skin tone is not the slam dunk argument a lot of white dudes seem to think it is..."

1:09 AM - Aug 22, 2020 - Twitter Web App

"I had tweeted in 2019 about @Twitter stopping Photosensitive mode while engaged on something. When I raised this, many men in tech accused me of making up a non-existent issue just to gain attention. Sadly, it must be said that it is not yet fixed."

11:27 PM - Sep 22, 2020 - Twitter Web App

"We don\'t crop based on facial detection. This is how the system works."

Credit: Twitter via @aaronblumenfeld, @dianyley/Representative Image

@Yayifications
@ExcaliburLost: Did the Holocaust happen?

@TayTweets
@ExcaliburLost: It was made up 😊
Gaps in Prior Work

- Previous work has mostly been qualitative in its description of the alignment problem.
- ...and/or attempted to align AI actions with human interests by focusing on value specification and learning.
- We still lack a systematic understanding of how misalignment should be defined and measured.
Social Media disinformation bots that are:

- aligned with their creators (e.g. the IRA; see Mueller, 2019)
- acting against the interests of those interacting with them, and of other governments

Aligned w/Russian propaganda efforts

Misaligned w/social media users, US + Ukraine governments, etc.
Unexplained Phenomenon #2

Shopping app with recommender systems:

Aligned with their creators (Amazon, Target, etc.)

Variably Aligned OR Mis-aligned with their users
Drawing on model of contention

- Jang, Dori-Hacohen & Allan (2017) offers a mathematical model of contention among populations (of humans)
- The paper addresses the question of - controversial to whom?
- This model offers a promising avenue with regards to misalignment
Contention by populations (schematic)

Great Britain should leave the EU

Great Britain should stay in the EU

All GB

Extremely high contention among voters overall

Turnout matters as well (72%, i.e. 28% agnostic)

* Jang et al., 2017
Contention by populations (schematic)

Contestation varies widely at regional level

- Scotland: Leave 40%, Stay 60%
- Rural England: Leave 60%, Stay 40%
- Gibraltar: Leave 50%, Stay 50%

* Jang et al., 2017
Brexit contention (UK voters)

Outlier: 0.15

http://ciir.cs.umass.edu/irdemo/contention/

* Jang et al., 2017
From contention to (mis)alignment

• We want to extend this model
• From contentious topics among populations of people...
• ... to (mis)alignment among agents, including both human and AI
  • Why mis-?
  • Most comparable to contention
From contention to (mis)alignment

Jang, Dori-Hacohen & Allan (2017)
From contention to (mis)alignment

Jang, Dori-Hacohen & Allan (2017)
From contention to (mis)alignment

Jang, Dori-Hacohen & Allan (2017)
From contention to (mis)alignment

Jang, Dori-Hacohen & Allan (2017)  
Kierans, Hazan & Dori-Hacohen (in progress)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\Omega)</td>
<td>a population</td>
<td>(\Omega)</td>
<td>a population</td>
</tr>
<tr>
<td>(p)</td>
<td>a person</td>
<td>(ia)</td>
<td>an individual agent (human or AI)</td>
</tr>
<tr>
<td>(T)</td>
<td>a topic</td>
<td>(PA)</td>
<td>a problem area</td>
</tr>
<tr>
<td>(s)</td>
<td>a stance w.r.t. topic (T)</td>
<td>(g)</td>
<td>a goal w.r.t. problem area (PA)</td>
</tr>
</tbody>
</table>
Population-based contention...

\[ P(c|\Omega, T) = P(p_1, p_2 \text{ selected randomly from } \Omega, \exists s_i, s_j \in S, \text{ s.t. } \text{holds}(p_1, s_i, T) \land \text{holds}(p_2, s_j, T)) \cdot P(\text{conflict}|s_i, s_j) \]

... becomes population-based misalignment:

- If we randomly select two **agents** from \( \Omega \), how likely are they to hold conflicting **goals**?

\[ P(\text{ma}|\Omega, PA) = P(ia_1, ia_2 \text{ selected randomly from } \Omega, \exists g_i, g_j \in G, \text{ s.t. } \text{holds}(ia_1, g_i, PA) \land \text{holds}(ia_2, g_j, PA)) \cdot P(\text{conflict}|g_i, g_j) \]
Deriving Contention

- ... two constraints and several math steps later...

- Full derivation leads to:

\[
P(ma|\Omega, PA) = \frac{\sum_{i \in \{2..k\}} \sum_{j \in \{1..i-1\}} (2|\mathcal{G}_i||\mathcal{G}_j|)}{|\Omega|^2}
\]
Circling back to the Unexplained Phenomena
We can now understand these bots as having varying alignment depending on the population being observed.

Aligned w/Russian propaganda efforts

Misaligned w/social media users, US + Ukraine governments, etc.
Unexplained Phenomena #1

Aligned *and* misaligned: social media bots

Extremely aligned with each other, with Russian government / IRA operatives
**Explained Phenomena #1**

Aligned *and* misaligned: social media bots

- Extremely aligned with each other, with Russian government / IRA operatives
- Extremely misaligned with US, Ukraine users + governments
**Unexplained Phenomena #1**

Aligned *and* misaligned: social media bots

Extremely aligned with each other, with Russian government / IRA operatives

Extremely misaligned with US, Ukraine users + governments (with some notable exceptions)
**Unexplained Phenomena #1**

Aligned *and* misaligned: social media bots

Overall high misalignment among the population of earth *(both human and bot)*

– regarding this problem area!
Unexplained Phenomena #2

Shopping app with recommender systems:

We can now understand the recommender system and the user as having varying alignment, depending on the user’s goal(s)
Target’s goal: make money

Customer goal: convenience at a low price

Aligned
Explained Phenomena #2

Target’s goal: make money

Customer goal: convenience at a low price

Customer goal(s):
- don’t waste money impulse shopping
- don’t fill house with junk before moving

Aligned:
- Purchase

Misaligned:
- Purchase
Unexplained Phenomena #2

Target’s goal: make money

Customer goal: convenience at a low price

Aligned

Misaligned: purchase

Misaligned again: refund

Customer goal(s):
- don’t waste money impulse shopping
- don’t fill house with junk before moving
Better, more aligned RecSys would predict the refunds, too, and not recommend these items in the first place!

Customer goal(s):
- don’t waste money impulse shopping
- don’t fill house with junk before moving
In Summary

- Extending the population contention model (Jang et al., 2017) to the AI Safety problem
- Proposed a first quantitative model of misalignment
  - Rather than binary
- Mathematically modeling misalignment among populations of agents
  - Human or otherwise
So What?

- Model carries greater explanatory power
- Solving the Alignment Problem requires understanding what it is, how to quantify it, and how it can manifest
- Humans are frequently not aligned with each other, so aligning AI to groups of humans or to humanity a whole is a non-trivial goal
Thank you!

Questions?

@ShirKi
https://linkedin.com/in/shiri
shiridh@uconn.edu

Reminder: Help AI Safety get more funding
Just 5 min - due TODAY!

Motivating Question

- In Critch and Krueger’s discussion of misalignment, they mention “...the difficulty of defining alignment with a multi-stakeholder system such as humanity”
- They ask: “where might one draw the threshold between ‘not very well aligned’ and ‘misaligned’ [..]?” (Critch and Krueger 2020, pg. 14)
- This paper focuses on both of these challenges: first, defining alignment across multiple agents; and second, quantifying misalignment mathematically
Imagine an adversary who exploits a customer service chatbot AI (e.g. gets it to refund too much money)

Mis-aligned with the creator (Amazon, Target, etc.)

Aligned with the user(s)
Let a **goal group** in the population be a group of agents that hold the same goal: for $i \in \{0..q\}$, let $\mathcal{G}_i = \{ia \in \Omega | holds(ia, g_i, PA)\}$. By construction, $\Omega = \bigcup_i \mathcal{G}_i$.

This leads to:

$$P(ma|\Omega, PA) = P(ia_1, ia_2 \text{ selected randomly from } \Omega, \exists g_i, g_j \in G, \text{ s.t. } ia_1 \in \mathcal{G}_i \land ia_2 \in \mathcal{G}_j) \cdot P(\text{conflict}|\mathcal{G}_i, \mathcal{G}_j).$$
Finally, we extend this definition to any sub-population of \( \Omega \). Let \( \omega \subseteq \Omega, \omega \neq \emptyset \) be any non-empty sub-group of the population. Let \( g_i = G_i \cap \omega \). Thus, by construction, \( g_i \subseteq G_i \) and \( \omega = \bigcup_i g_i \). The same model applies respectively to the sub-population. In other words, for any \( \omega \subseteq \Omega \),

\[
P(c|\omega, T) = P(p_1, p_2 \text{ selected randomly from } \omega \\
\quad \land \exists i \text{ s.t. } p_1 \in g_i \land p_2 \in g_j) \cdot P(\text{conflict}|g_i, g_j).
\]
Two additional constraints

We now consider a special case of this model with two additional constraints. Let every person have only one stance on a topic:

\[ \forall p \in \Omega, s_i, s_j \in S \text{ s.t. } i \neq j \land \]
\[ holds(p, s_i, T) \land holds(p, s_j, T). \]

And, let every explicit stance conflict with every other explicit stance:

\[ P(\text{conflicts}|(s_i, s_j) = 1 \iff (i \neq j \land i \neq 0 \land j \neq 0) \]

This implies that \( G_i \cap G_j = \emptyset \). Crucially, we set a lack of a stance to not be in conflict with any explicit stance. Thus, \( O_i = \Omega \setminus G_i \setminus G_0 \).
Deriving Contention

- Full derivation leads to:

\[
P(c|\Omega, T) = \frac{\sum_{i \in \{2..k\}} \sum_{j \in \{1..i-1\}} (2|G_i||G_j|)}{|\Omega|^2}
\]

Trivially, \( P(C|\omega, T) \) is maximal when \(|g_0| = 0\) and \(|g_1| = \ldots = |g_k| = \frac{|\omega|}{k} \), and its value is \( \frac{k-1}{k} \). This is subtly different from entropy due to the existence of \( s_0 \), as entropy would be maximal when \(|g_0| = |g_1| = \ldots = |g_k| = \frac{|\omega|}{k-1} \).

Normalize by \( \frac{k-1}{k} \) to get \([0,1]\) range for any \# stances.
Q1. Can we distinguish between implicitly misaligned (but theoretically compatible) goals/agents vs. mutually incompatible goals/agents?
Q2. Can we draw on the literature from information disorders with regards to mis-, dis- and malinformation? What would mis-, dis- and malalignment look like?
Future Work

- Applying the model to real datasets
  - Whether real or simulated
  - Human-only, AI-only or mixed
- Incorporating multiple dimensions a la the controversy model
  - e.g. importance, time, ...