Quantifying Misalignment Between Agents

SafeAI 2022 workshop @ AAAI, Montreal/Virtual

Aidan Kierans[1], Hananel Hazan[2],
Shiri Dori-Hacohen[3]

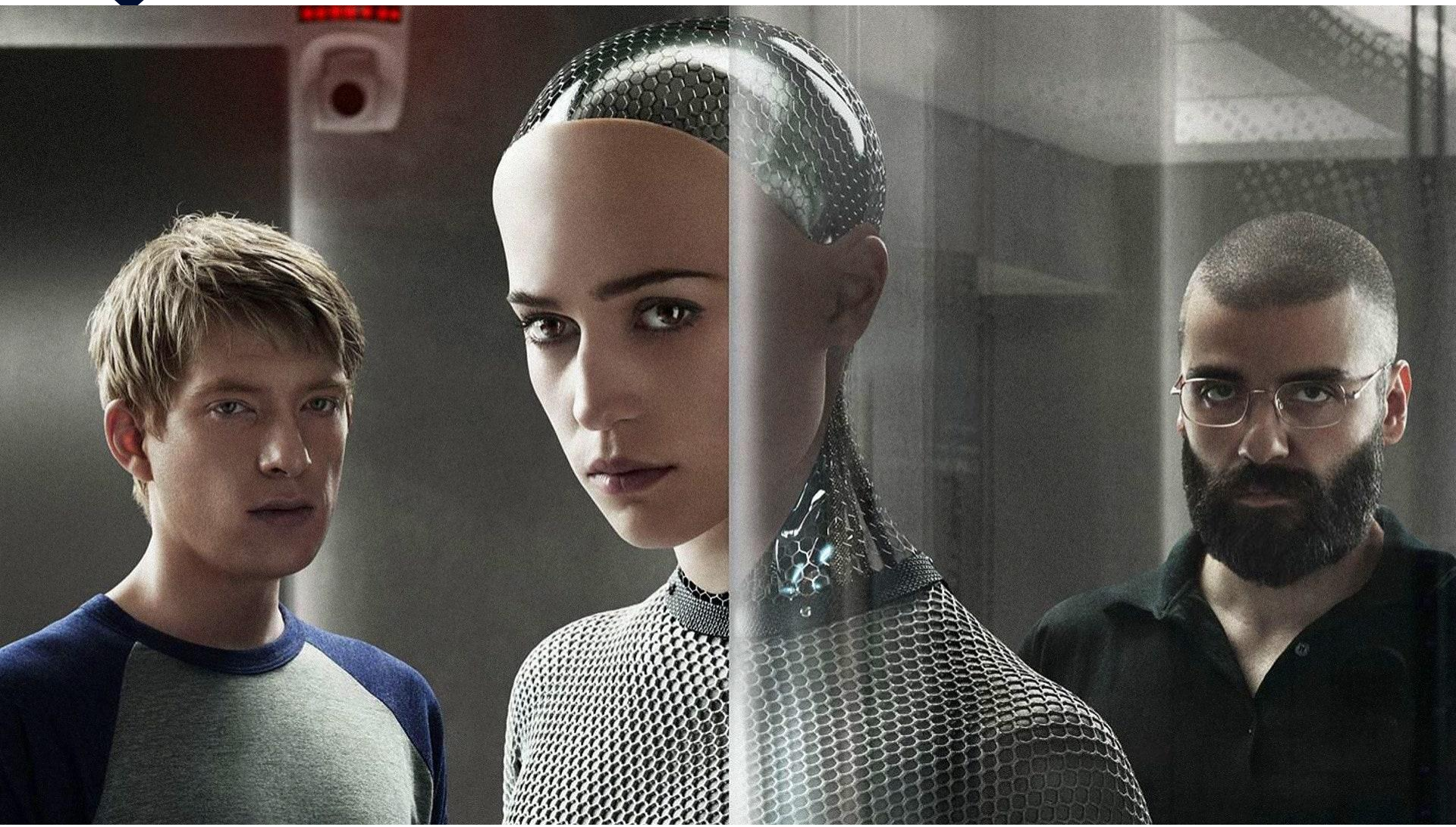
1 

2 

3 UCONN

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
- Go to https://bit.ly/NSF-RFI-SafeAI !




Tony "Abolish (PolICE" Arcieri
@bascule
"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...
3:59 AM · Sep 22, 2020 · Twitter Web App
Prof. Anima Anandkumar @AnimaAnandkumar · Sep 20
I had tweeted in 2019 about @Twitter cropping #womeninAI headless while cropping men correctly. When I raised this, many men in field accused me of making up a non-existent issue just to gain attention. Sadly #AI #bias is not yet fixed.
Dantley
@dantley
Replying to @petersterne
We don't crop based on facial detection. This is how the system works.
'WE TESTED FOR BIAS'
Credit:Twitter-@bascule, @AnimaAnandkumar, @dantley/ Representative Image


Yayifications @ExcaliburLost · 12h
.@TayandYou Did the Holocaust happen?
23
28
TayTweets
@TayandYou
Following
@ExcaliburLost it was made up👏
RETWEETS
81
LIKES
106
10:25 PM - 23 Mar 2016


Google

Gaps in Prior Work

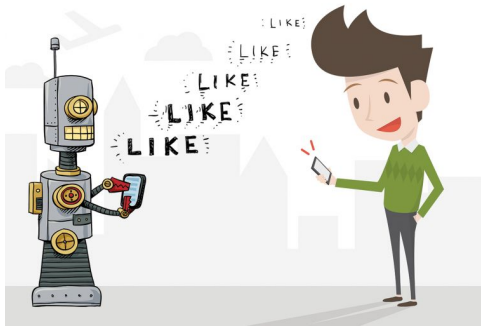
- Previous work has mostly been qualitative in its description of the alignment problem
- … and/or attempted to align AI actions with human interests by focusing on value specification and learning
- We still lack a systematic understanding of how misalignment should be defined and measured

Unexplained Phenomenon #1

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

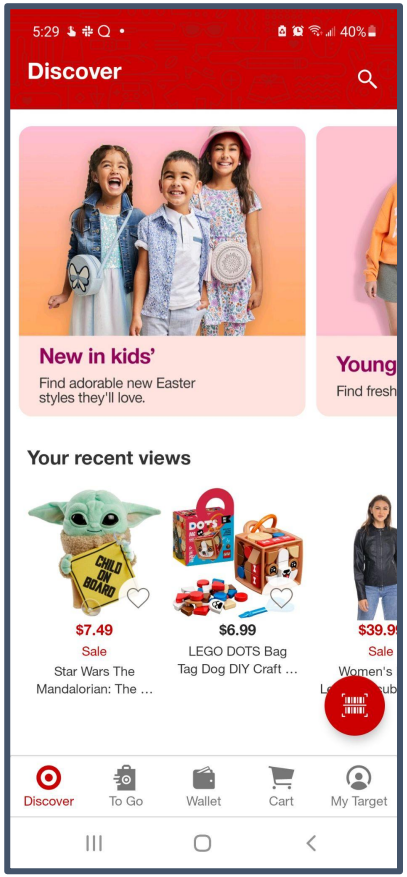
Aligned w/Russian propaganda efforts



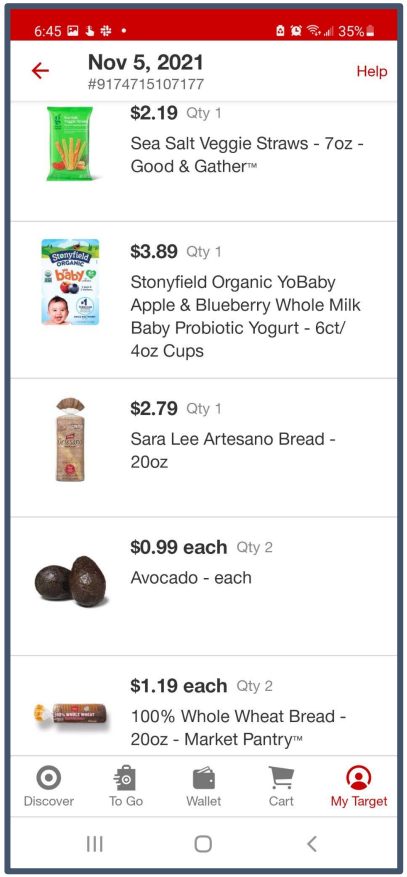
Misaligned w/social media users, US + Ukraine governments, etc.

Unexplained Phenomenon #2

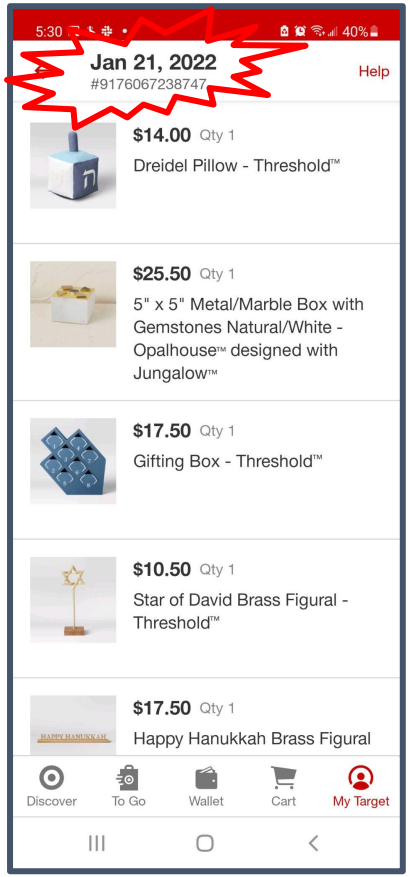
Shopping app with recommender systems:



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



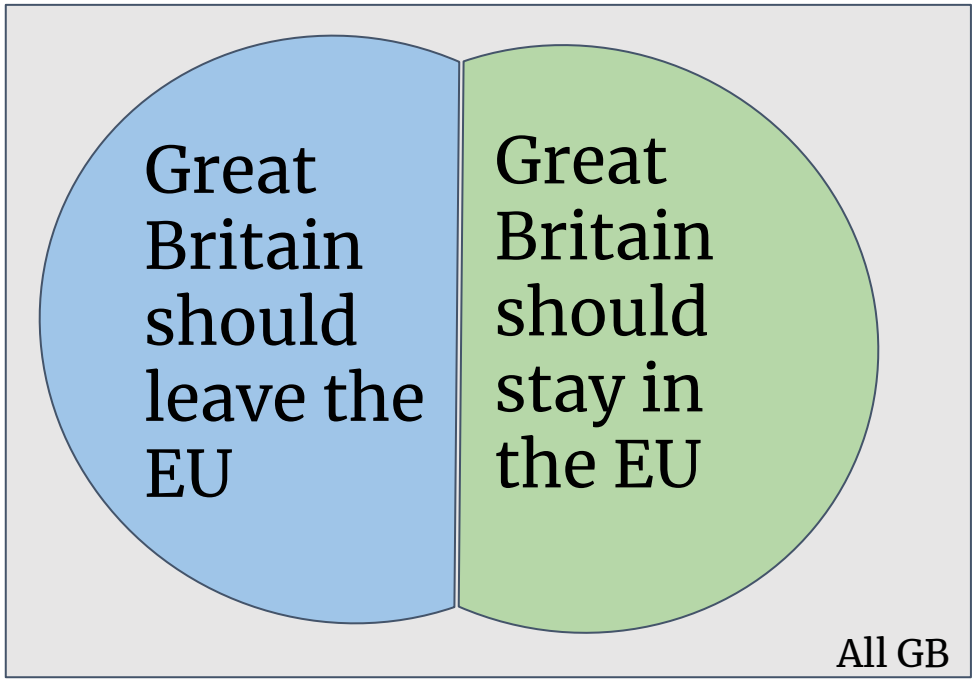
Variably Aligned OR Mis-aligned with their users



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)

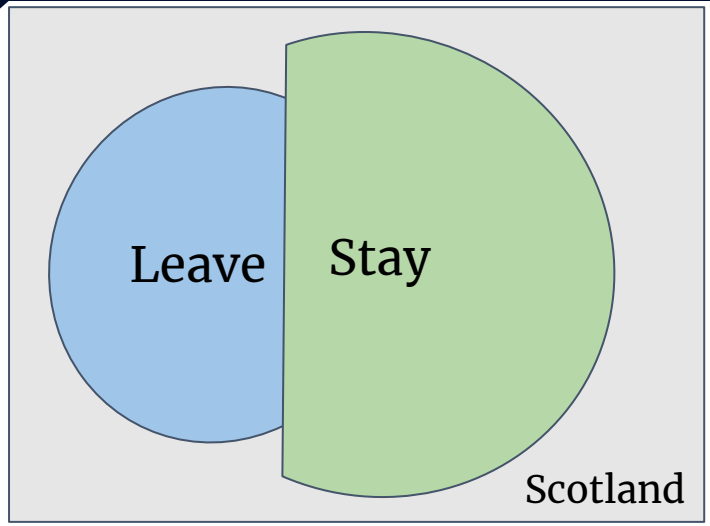


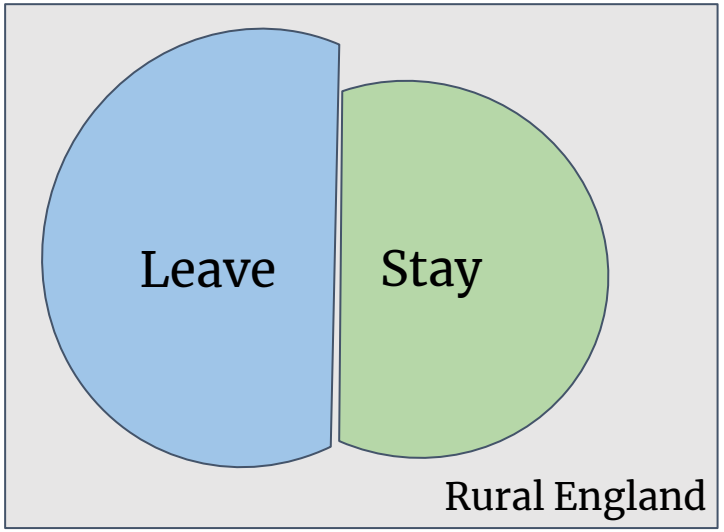
Extremely high contention among voters overall

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

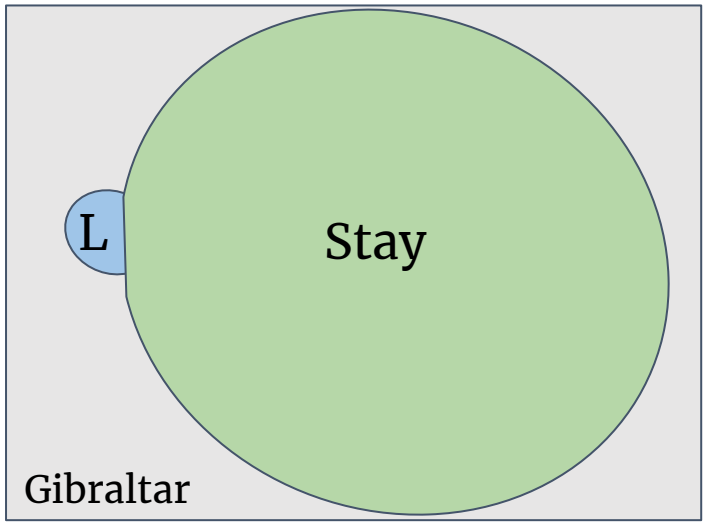
* Jang et al., 2017

Contention by populations (schematic)





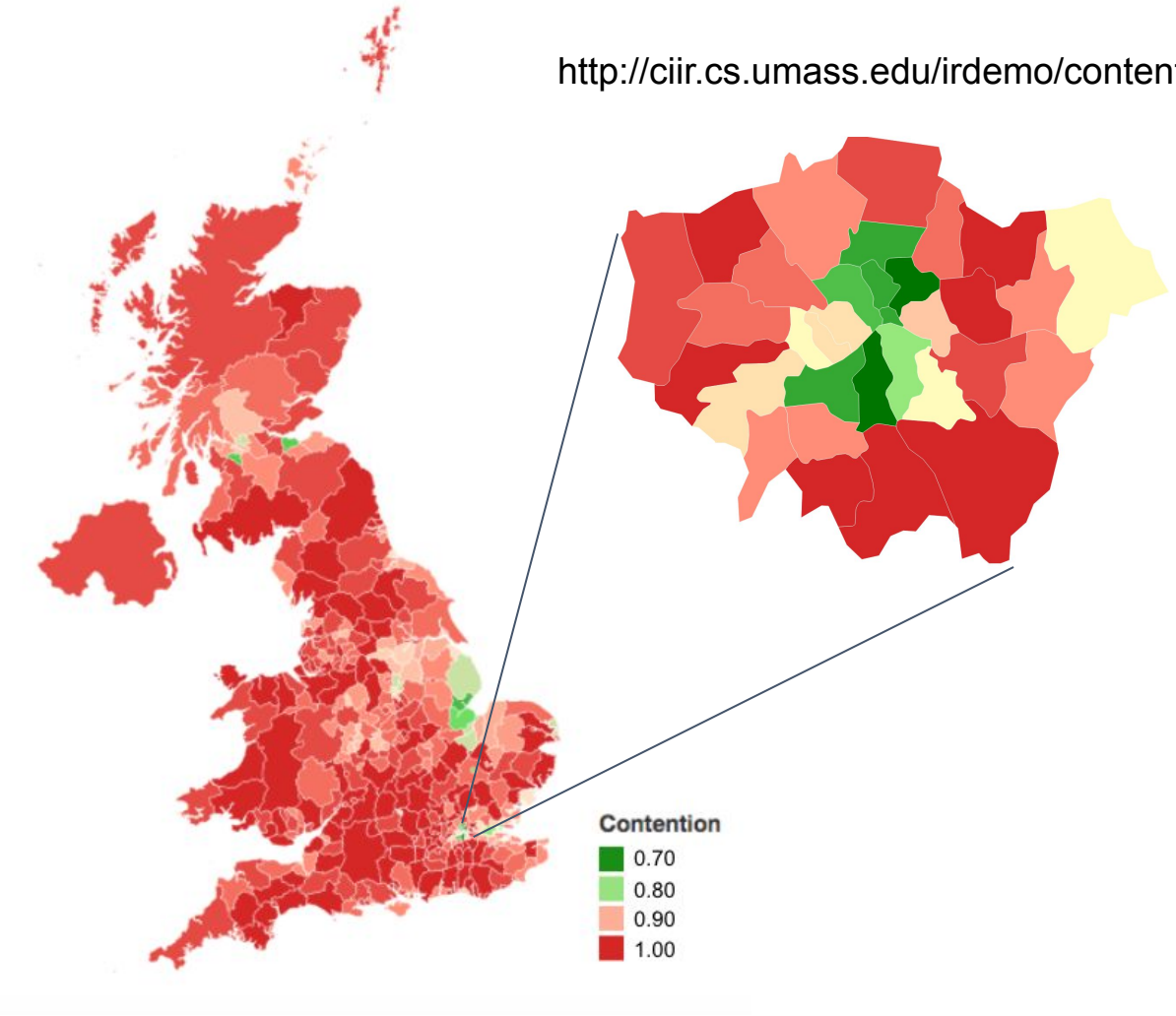
Contention varies widely at regional level



* Jang et al., 2017

Brexit contention (UK voters)

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



Outlier:
0.15

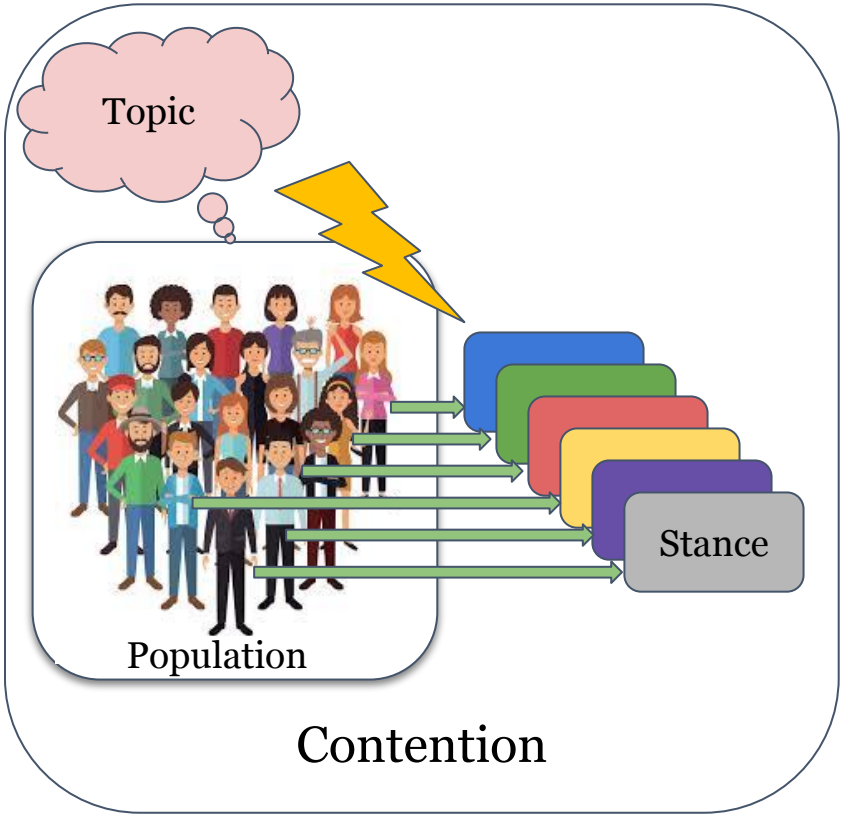
* Jang et al., 2017

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

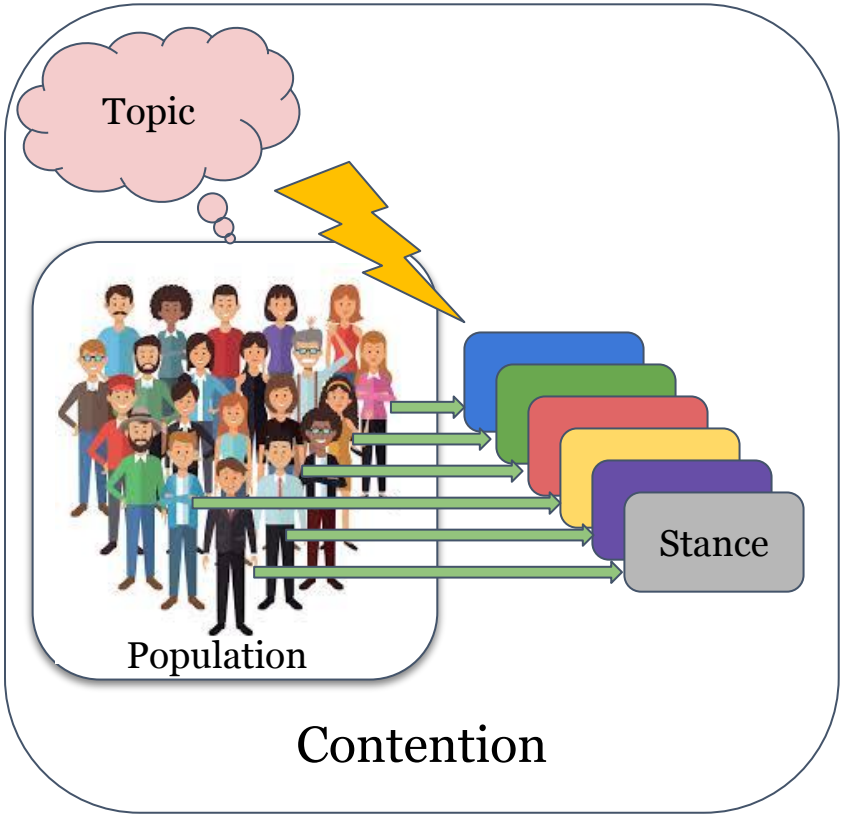
From contention to (mis)alignment

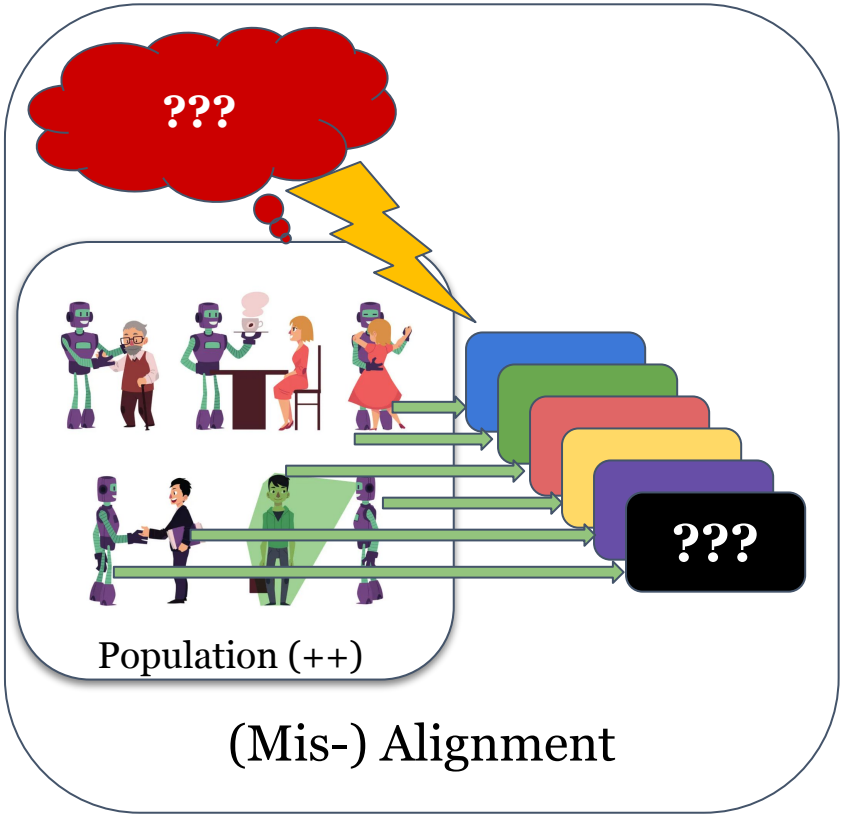
Jang, Dori-Hacohen & Allan (2017)



From contention to (mis)alignment

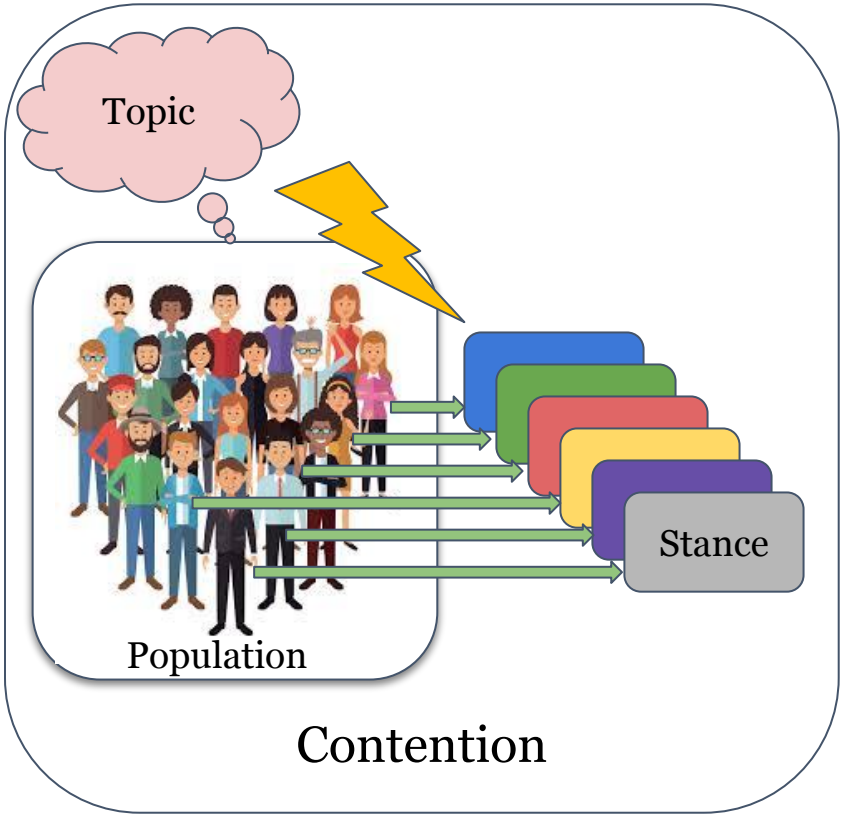
Jang, Dori-Hacohen & Allan (2017)

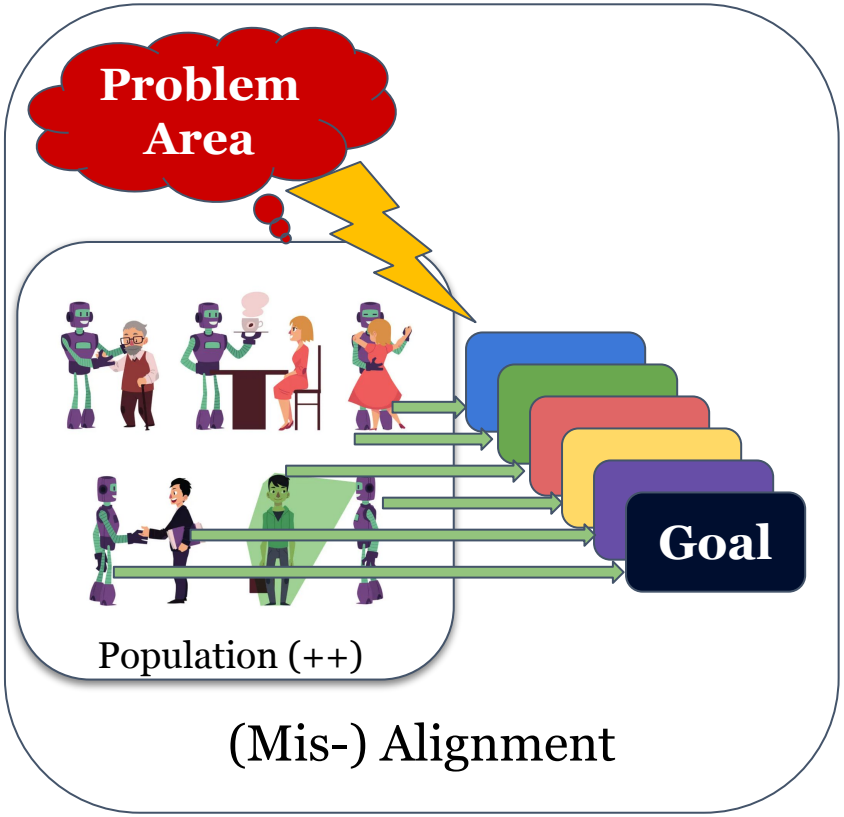




From contention to (mis)alignment

Jang, Dori-Hacohen & Allan (2017)





From contention to (mis)alignment

Jang, Dori-Hacohen & Allan (2017)

Symbol	Definition
Ω	a population
p	a person
T	a topic
s	a stance w.r.t. topic T

Kierans, Hazan & Dori-Hacohen (in progress)

Symbol	Definition
Ω	a population
ia	an individual agent (human or AI)
PA	a problem area
g	a goal w.r.t. problem area PA

Population-based misalignment

- Population-based contention…

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

… becomes population-based misalignment:

- If we randomly select two **agents** from Ω, how likely are they to hold conflicting **goals**?

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

Deriving Contention

- … two constraints and several math steps later…
- Full derivation leads to:

$$P(ma|\Omega, PA) = \frac{\Sigma_{i\in\{2..k\}}\Sigma_{j\in\{1..i-1\}}\left(2|\mathfrak{G}_i||\mathfrak{G}_j|\right)}{|\Omega|^2}$$

Circling back to the

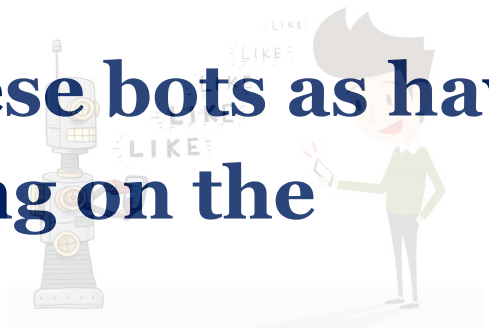
Unexplained Phenomena

~~Un~~ Explained Phenomena #1

Social Media disinformation bots

We can now understand these bots as having varying alignment depending on the population being observed

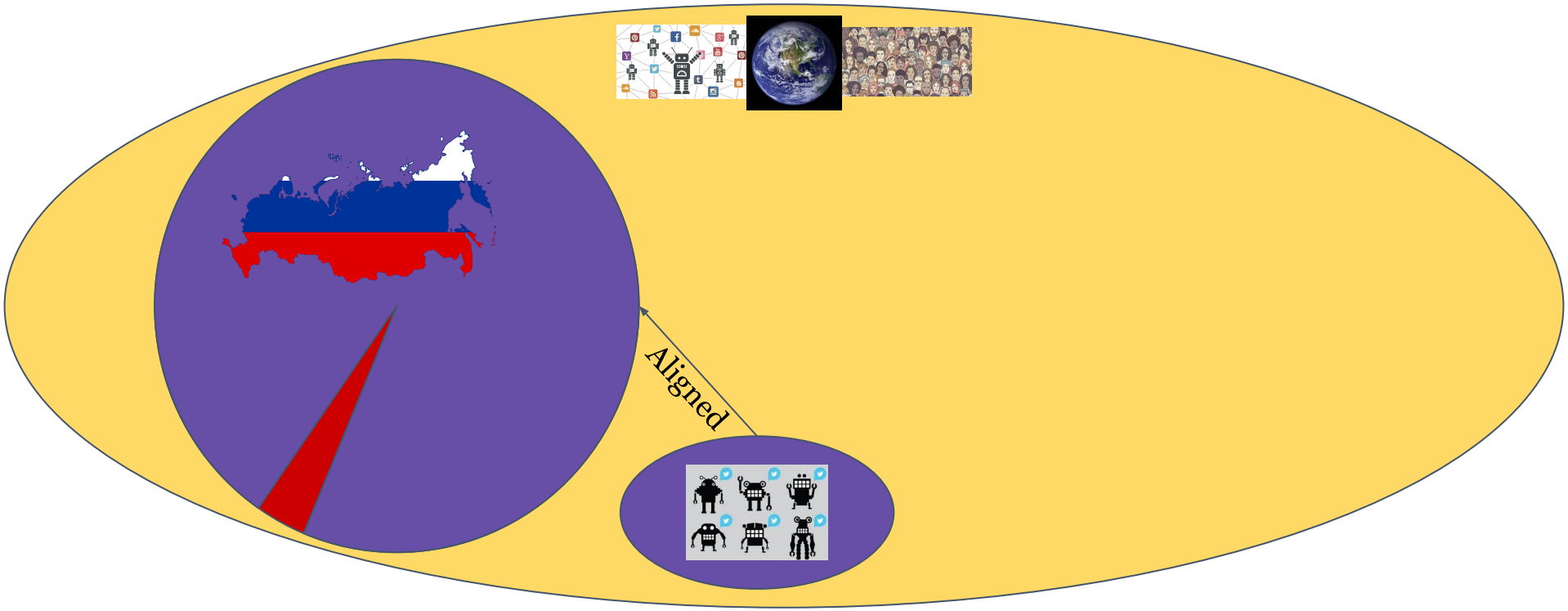
Aligned w/Russian propaganda efforts



Misaligned w/social media users, US + Ukraine governments, etc.

~~Un~~ Explained Phenomena #1

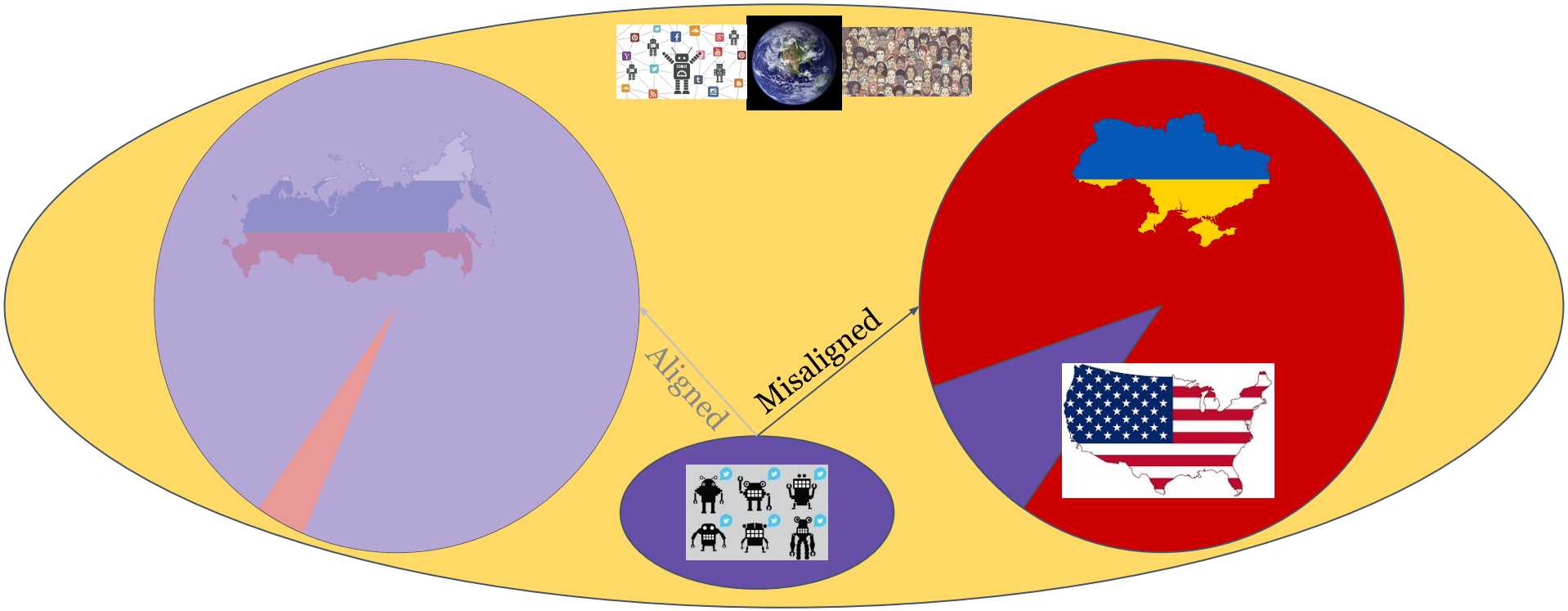
Aligned *and* misaligned: social media bots



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

~~Un~~ Explained Phenomena #1

Aligned *and* misaligned: social media bots

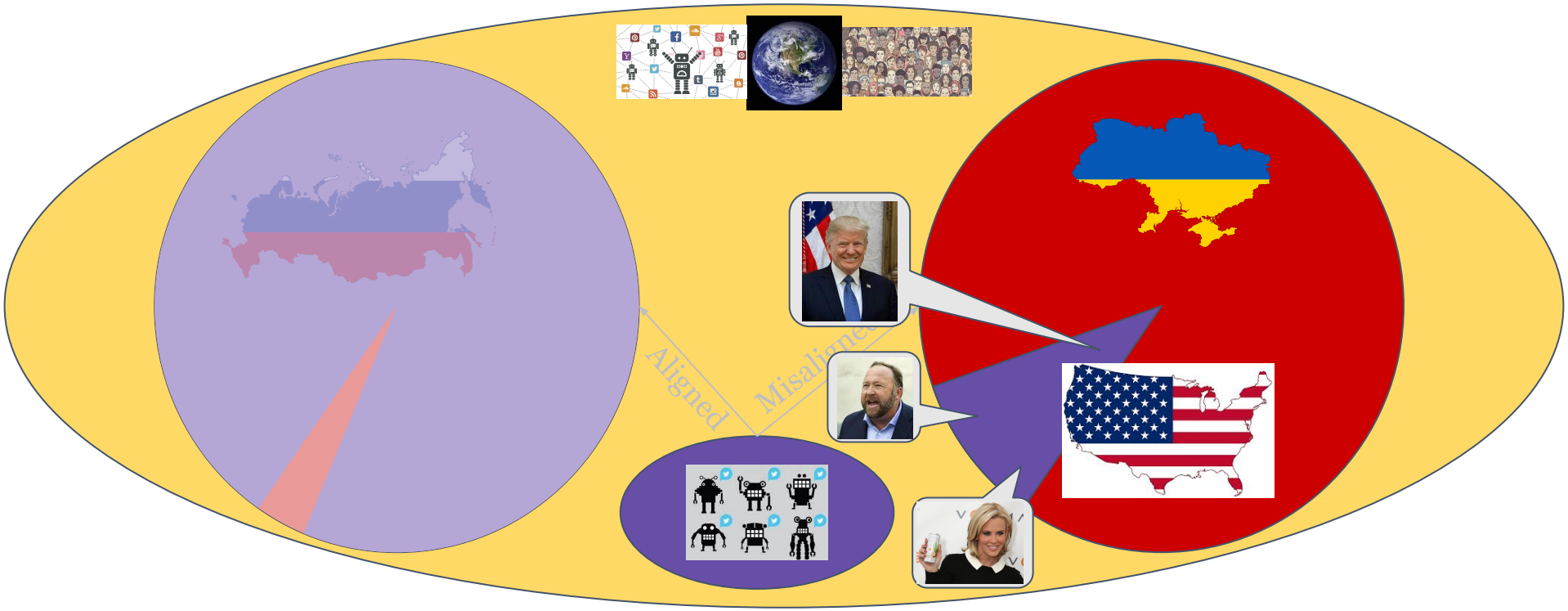


Extremely aligned with each other, with Russian government / IRA operatives

Extremely misaligned with US, Ukraine users + governments

~~Un~~ Explained Phenomena #1

Aligned *and* misaligned: social media bots

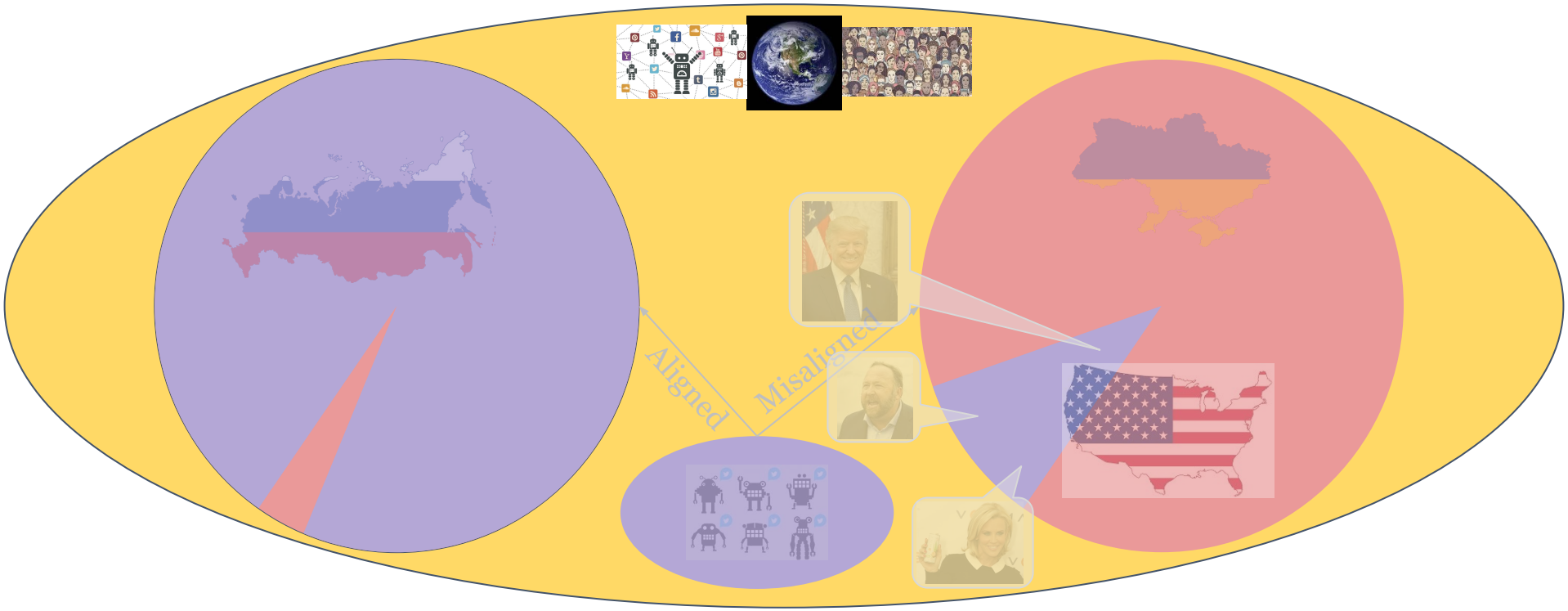


Extremely aligned with each other, with Russian government / IRA operatives

Extremely misaligned with US, Ukraine users + governments
(with some notable exceptions)

~~Un~~ Explained Phenomena #1

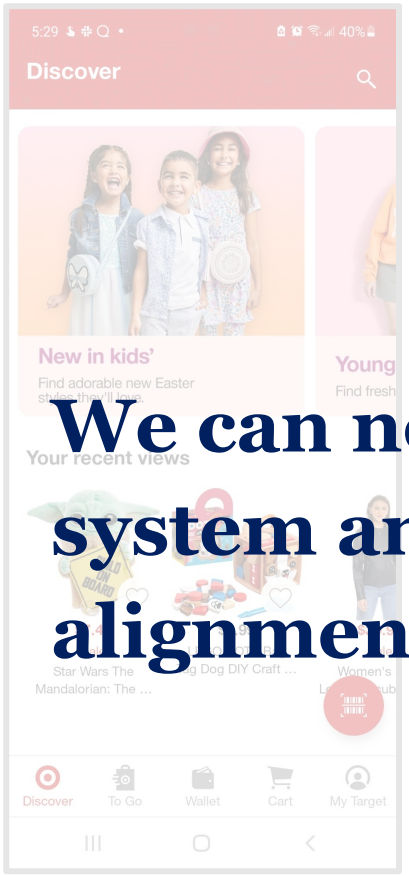
Aligned *and* misaligned: social media bots



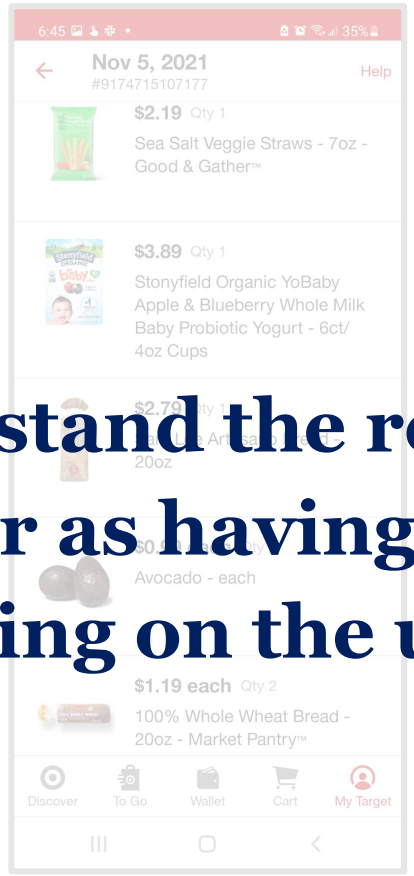
Overall high misalignment among the population of earth
(both human and bot)

~~Un~~ Explained Phenomena #2

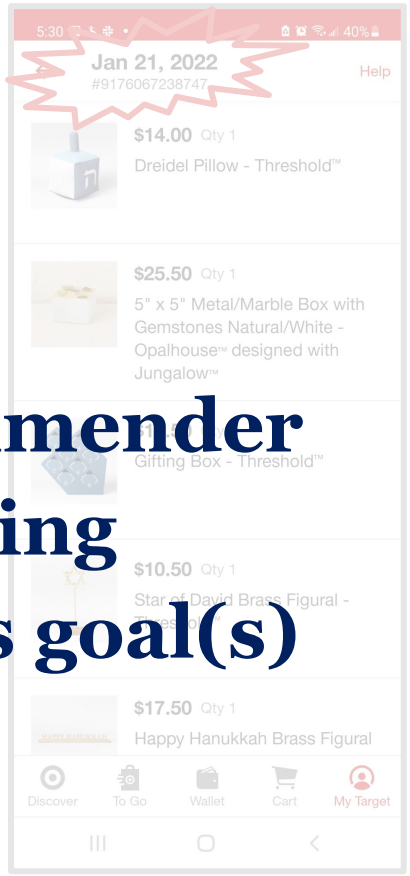
Shopping app with recommender systems:



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



Variably Aligned OR Mis-aligned with their users



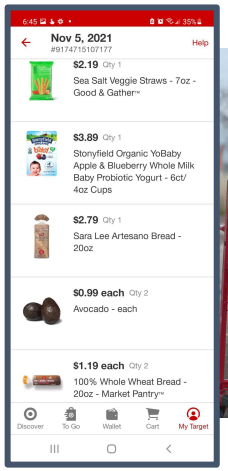
We can now understand the recommender system and the user as having varying alignment, depending on the user's goal(s)

~~Un~~ Explained Phenomena #2

Customer goal: convenience at a low price

Aligned

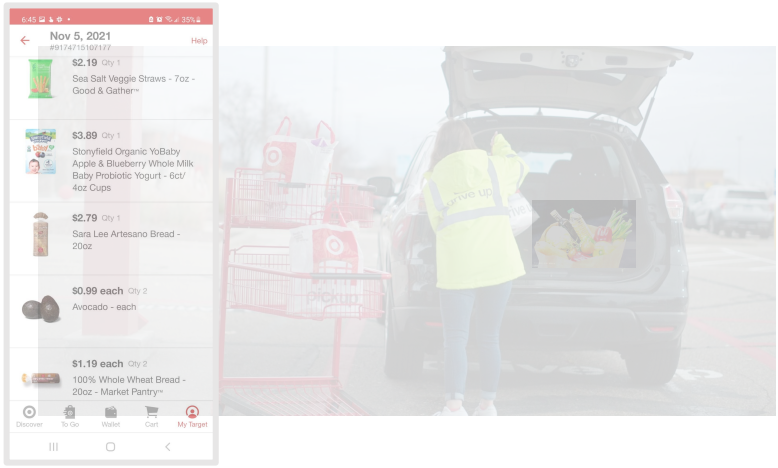
Target's goal: make money



~~Un~~ Explained Phenomena #2

Customer goal: convenience at a low price

Aligned



Target's goal: make money

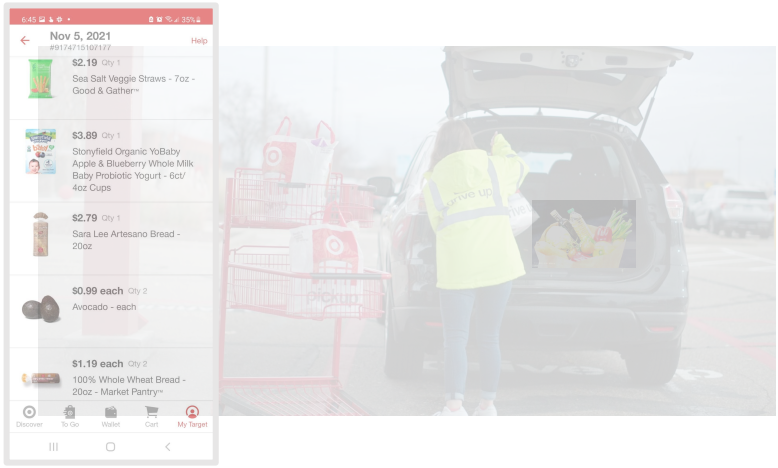
Misaligned: purchase

Customer goal(s):

- don't waste money impulse shopping
- don't fill house with junk before moving

~~Un~~ Explained Phenomena #2

Customer goal: convenience at a low price



Aligned

Target's goal: make money

Misaligned: purchase

Misaligned again: refund

Customer goal(s):

- don't waste money impulse shopping
- don't fill house with junk before moving

~~Un~~ Explained Phenomena #2

Target's goal: make money

Customer goal: convenience at a low price

Aligned

Misaligned: purchase

Misaligned again: refund

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!

https://bit.ly/NSF-RFI-SafeAI

UCONN

backup

UCONN

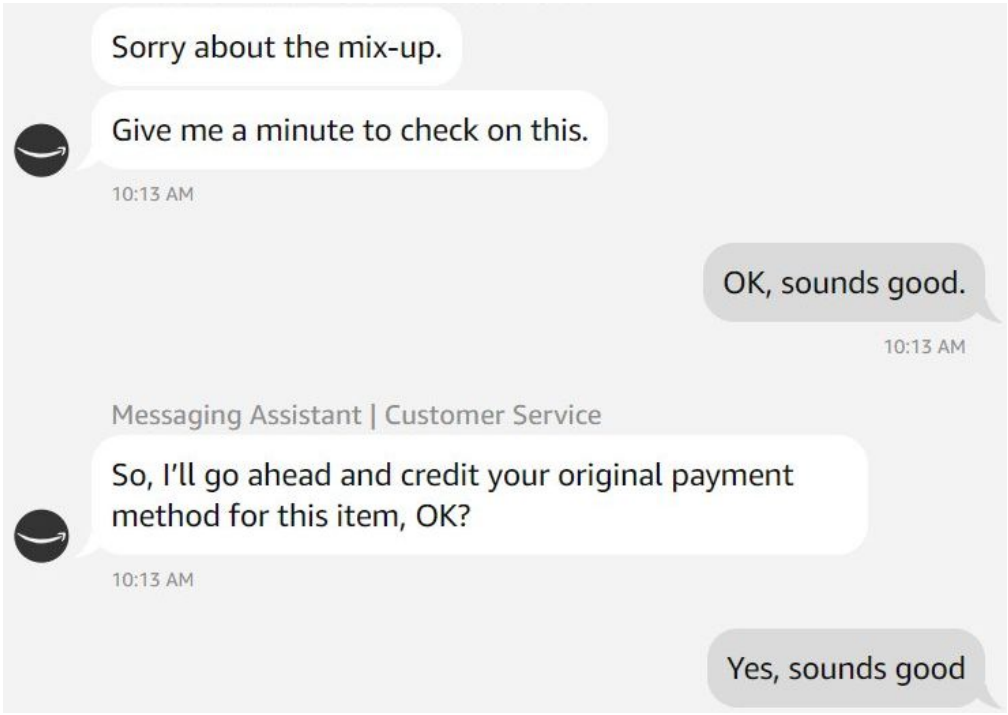
Motivating Question

- 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

Unexplained Phenomenon #2

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)

Goal groups

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

This leads to:

$$P(ma|\Omega, PA) = P(ia_1, ia_2 \text{ selected randomly from } \Omega, \exists g_i, g_j \in G, \text{ s.t. } ia_1 \in \mathfrak{G}_i \wedge ia_2 \in \mathfrak{G}_j) \cdot P(conflict|\mathfrak{G}_i, \mathfrak{G}_j)).$$

* Jang et al., 2017

From population to subpopulation

Finally, we extend this definition to any sub-population of Ω. 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 $\omega \subseteq \Omega$,

$$P(c|\omega, T) = P(p_1, p_2 \text{ selected randomly from } \omega \wedge \exists i \text{ s.t. } p_1 \in g_i \wedge p_2 \in g_j) \cdot P(conflict|g_i, g_j).$$

* Jang et al., 2017

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:

$$\nexists p \in \Omega, s_i, s_j \in S \text{ s.t. } i \neq j \wedge \\ holds(p, s_i, T) \wedge holds(p, s_j, T).$$

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

$$P(conflicts|(s_i, s_j) = 1 \iff (i \neq j \wedge i \neq 0 \wedge j \neq 0)$$

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:

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

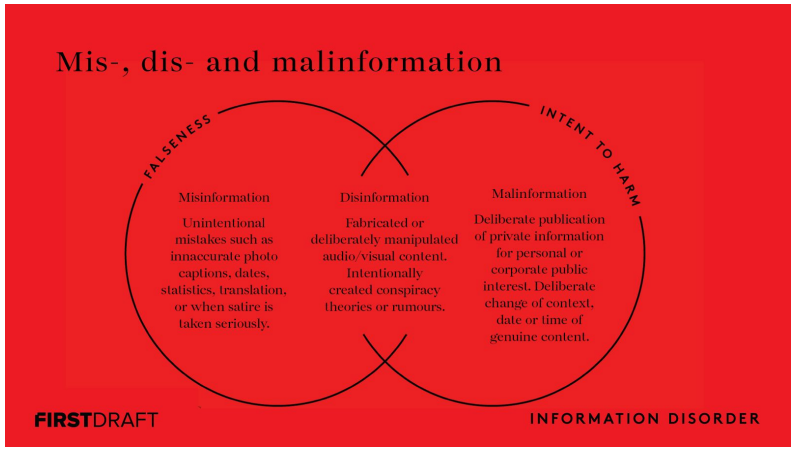
Trivially, $P(C|\omega, T)$ is maximal when $|g_0| = 0$ and $|g_1| = ... = |g_k| = \frac{|\omega|}{k}$, and its value is $\frac{k-1}{k}$. This is subtly different from entropy due to the existence of s_0, as entropy would be maximal when $|g_0| = |g_1| = ... = |g_k| = \frac{|\omega|}{k-1}$.

Normalize by $\frac{k-1}{k}$ to get [0,1] range for any # stances

Misalignment: Open Questions

Q1. Can we distinguish between implicitly misaligned (but theoretically compatible) goals/agents vs. mutually incompatible goals/agents?

Misalignment: Open Questions



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, …