Analysis of Moderation Strategies on a Reddit-like Network

Fabian, Yuri, Frederick, Max

Key Questions

- How useful is it to moderate and try to influence a social network like reddit?

More specifically:

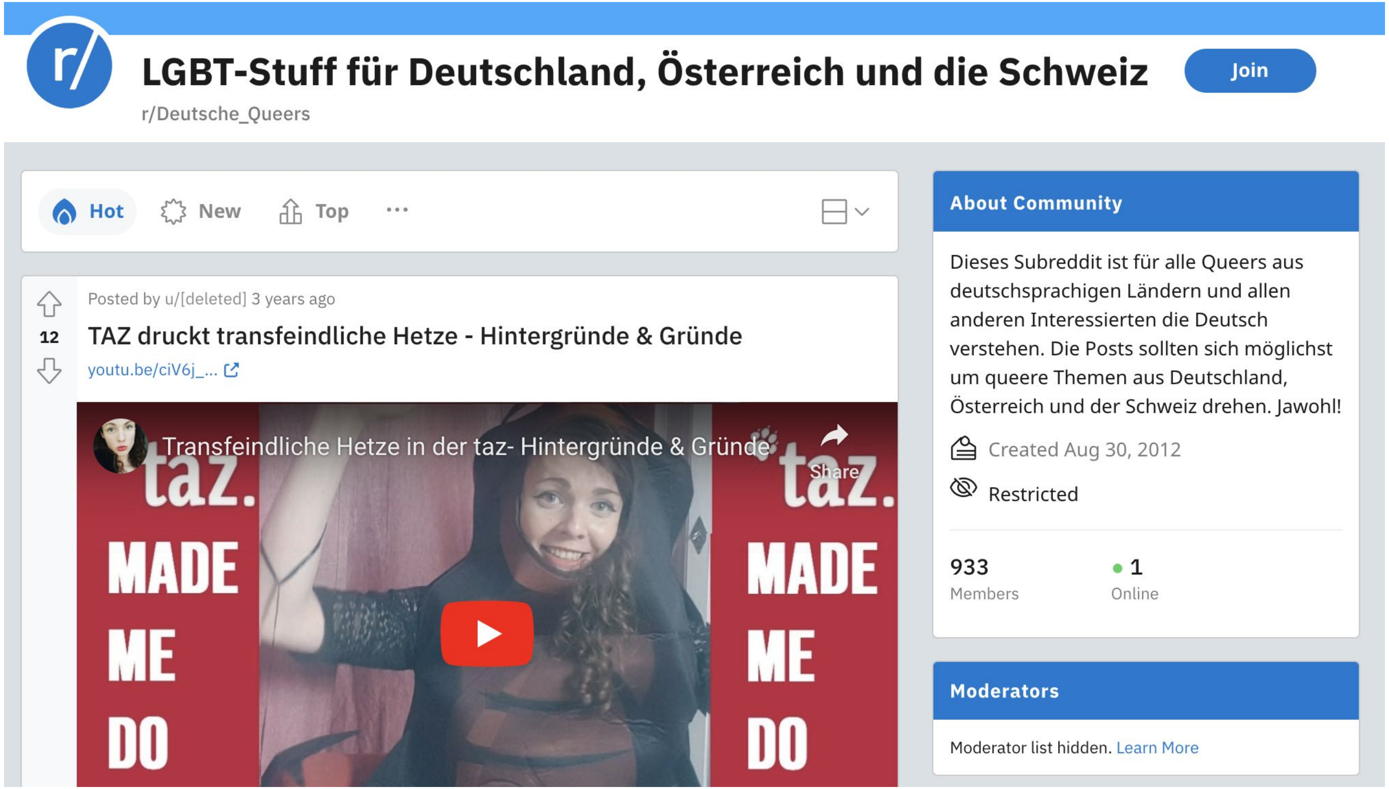
1. How can we best simulate Reddit on an agent-based model in Python?
2. What (realistic!) moderation techniques can we simulate and what are its effects on the opinion dynamics inside such a network?

Recap: How Reddit Works

- Users follow Subreddits
- Users create/read Posts
- Users rate Posts
- Two Post Sorting Methods (Hot and New)

redditinc.com

How Reddit Works: Example of a Subreddit



LGBT-Stuff für Deutschland, Österreich und die Schweiz

r/Deutsche_Queers

Posted by u/[deleted] 3 years ago

TAZ druckt transfeindliche Hetze - Hintergründe & Gründe

youtu.be/ciV6j_...

About Community

Dieses Subreddit ist für alle Queers aus deutschsprachigen Ländern und allen anderen Interessierten die Deutsch verstehen. Die Posts sollten sich möglichst um queere Themen aus Deutschland, Österreich und der Schweiz drehen. Jawohl!

Created Aug 30, 2012

Restricted

933 Members

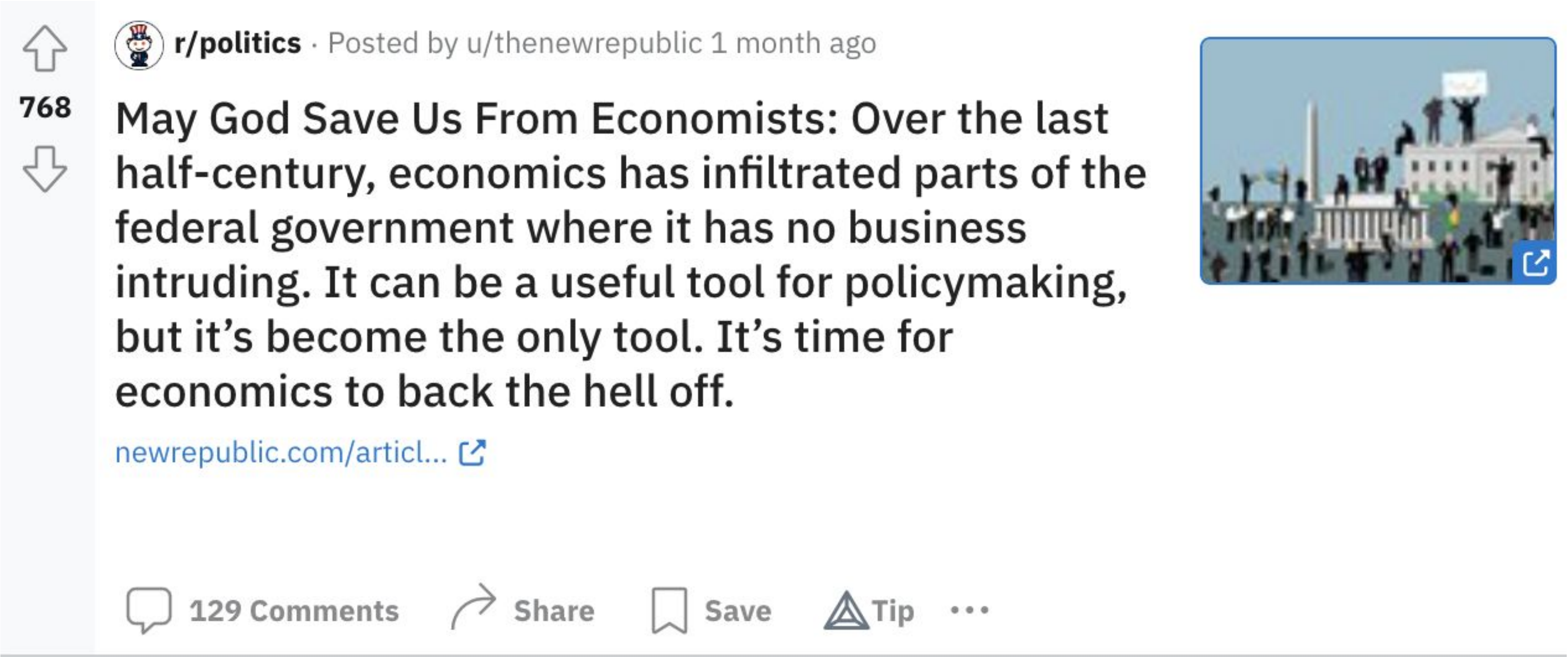
● 1 Online

Moderators

Moderator list hidden. Learn More

reddit.com/r/Deutsche_Queers

How Reddit Works: Example of a Post



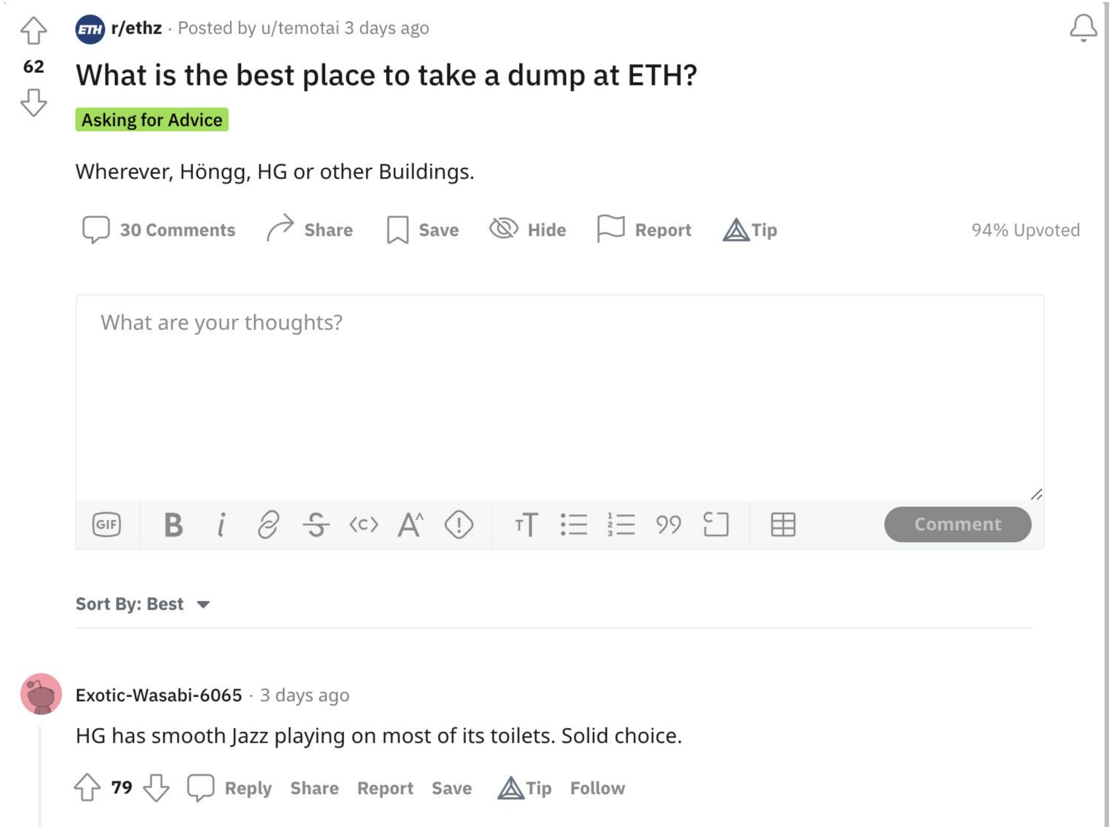
r/politics · Posted by u/thenewrepublic 1 month ago

May God Save Us From Economists: Over the last half-century, economics has infiltrated parts of the federal government where it has no business intruding. It can be a useful tool for policymaking, but it's become the only tool. It's time for economics to back the hell off.

newrepublic.com/articl...

reddit.com/r/politics

How Reddit Works: Example of a Post



r/ethz · Posted by u/temotai 3 days ago

62

What is the best place to take a dump at ETH?

Asking for Advice

Wherever, Höngg, HG or other Buildings.

30 Comments Share Save Hide Report Tip 94% Upvoted

What are your thoughts?

Comment

Sort By: Best

Exotic-Wasabi-6065 · 3 days ago

HG has smooth Jazz playing on most of its toilets. Solid choice.

79 Reply Share Report Save Tip Follow

Source: reddit.com/r/ethz

Why Reddit in the first place?

- Anonymous,

  more liberal/extremist opinions

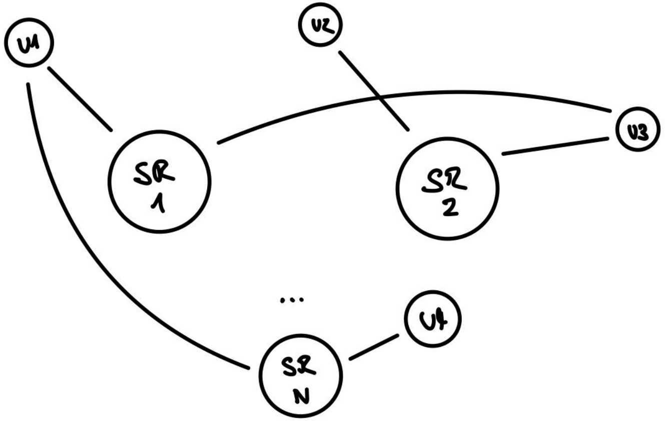
- Unique Structure
- Bottom-Up Moderation

- Contrast to restricted moderation

https://images.thequint.com/thequint%2F2022-01%2Fb6f0e23e-b291-481d-80b8-9955ea0c6406%2Freddit.jpg?rect=0%2C0%2C2000%2C1125&auto=format%2Ccompress&fmt=webp&width=720

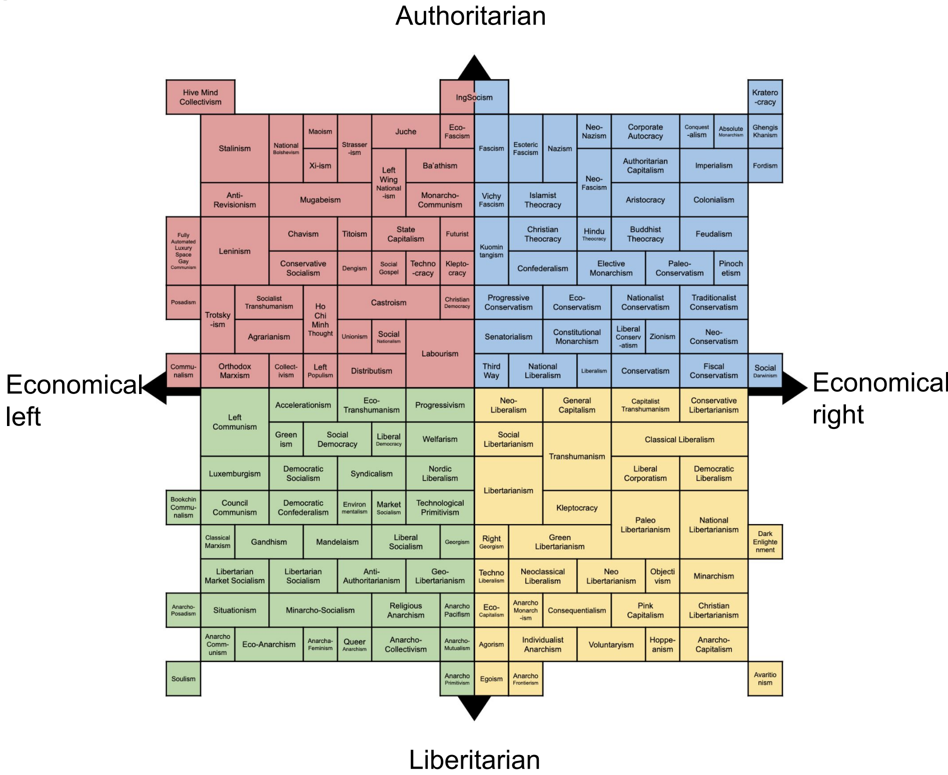
Our Model: Basic Graph

- Model as a Graph with **Users** and **Subreddits** as **Vertices**
- **Edge** exists if User follows Subreddit
- No edges between Users



Our Model: Opinion Measuring

- One dimensional opinions are oversimplifying
- Opinion as 2-D vector with real values between 0 and 1
- Two-Dimensional opinion fields a.k.a. political compasses are often used to model politicians positions
- For the rest of the presentation: **Opinion value =: Bias**



Our Model: Initialization

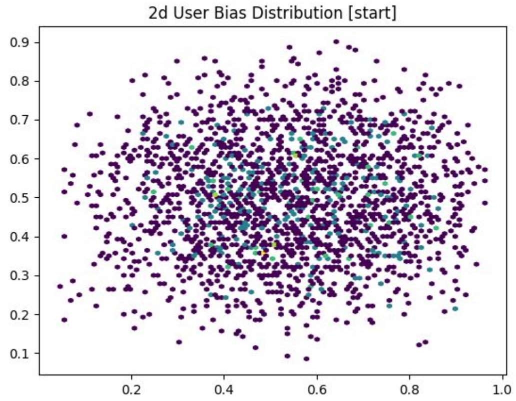
- Fixed amount of Users and Subreddits
- User bias generated with Beta-Distribution
    - Bias value 1: Alpha = 3.1, Beta = 2.7
    - Bias value 2: Alpha = 4.9, Beta = 5.2
    - Normal distribution too simplified

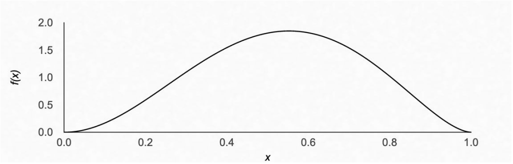
Our Model: Initialization

- Fixed amount of Users and Subreddits
- User bias generated with Beta-Distribution
    - Bias value 1: Alpha = 3.1, Beta = 2.7
    - Bias value 2: Alpha = 4.9, Beta = 5.2
    - Normal distribution too simplistic
- Users randomly allocated among Subreddits

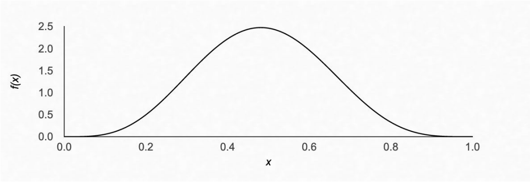
Beta Distribution Formula

$$f(x) = \frac{(x-a)^{p-1}(b-x)^{q-1}}{B(p,q)(b-a)^{p+q-1}} \qquad B(\alpha,\beta) = \int_1^0 t^{(\alpha-1)}(1-t)^{(\beta-1)}dt.$$





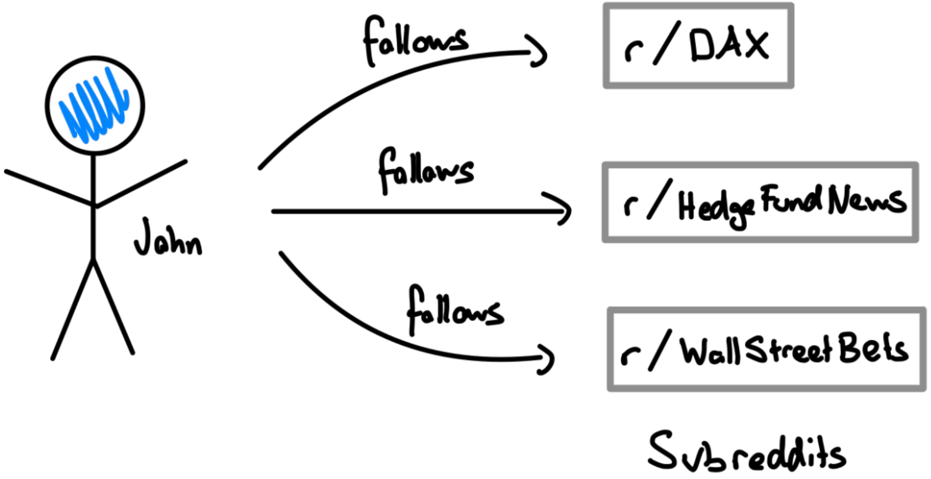
Beta-Distribution of Bias value 1

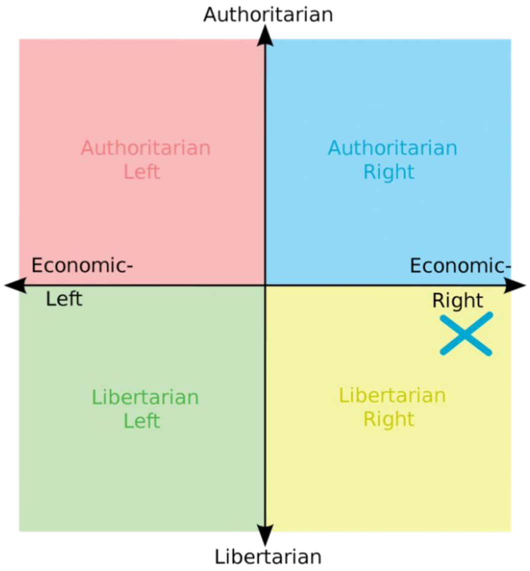


Beta-Distribution of Bias value 2

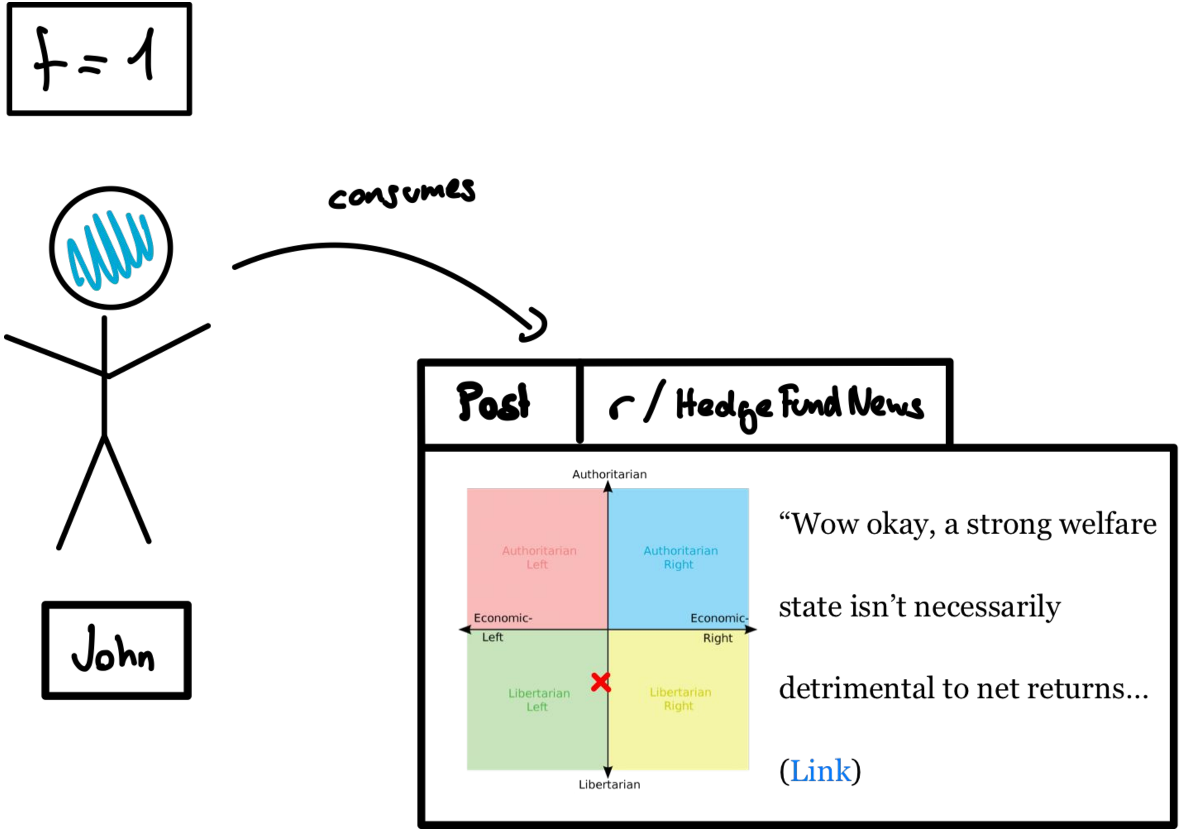
Our Model: Interactions - Example: John

- Banker & Staunch capitalist
- Pretty far right on the economic spectrum
- Very centric on the degree of authoritarianism he prefers



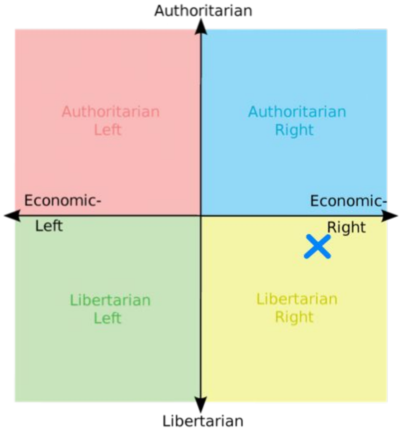


Our Model - Interactions

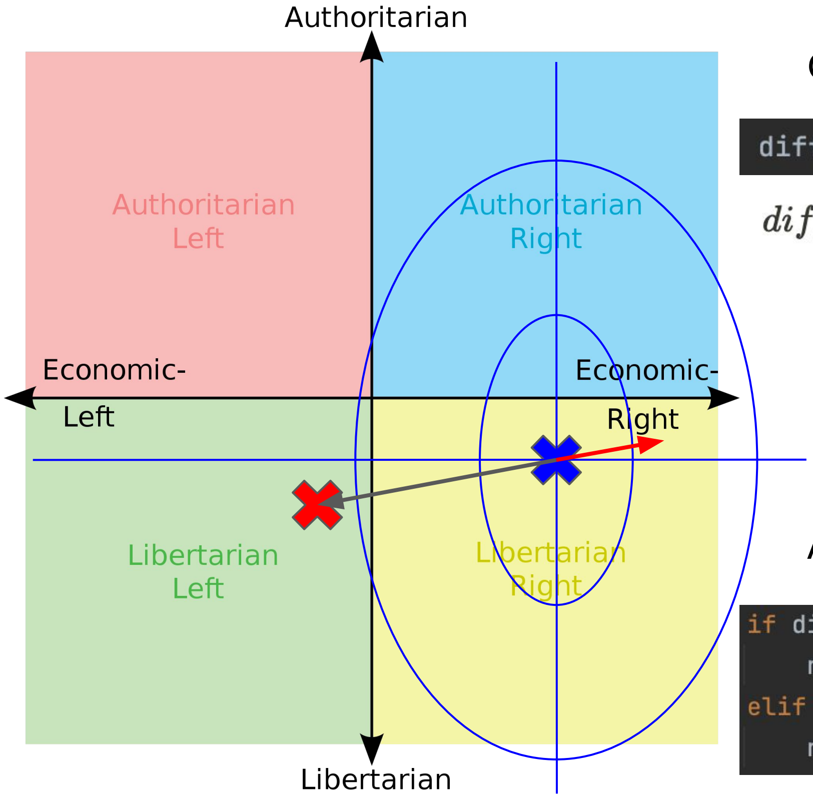


"Wow okay, a strong welfare state isn't necessarily detrimental to net returns...

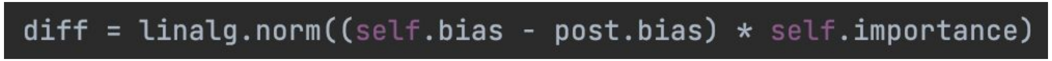
(Link)



Our Model - Interactions



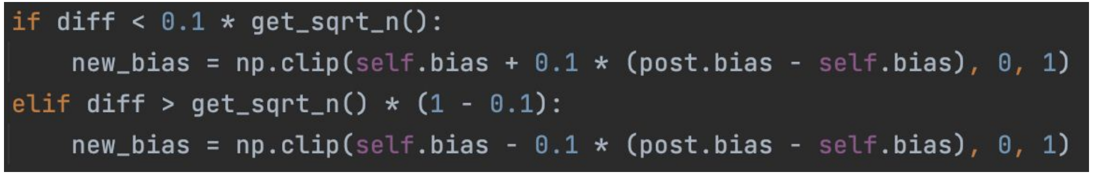
Opinion Distance



```
diff = linalg.norm((self.bias - post.bias) * self.importance)
```

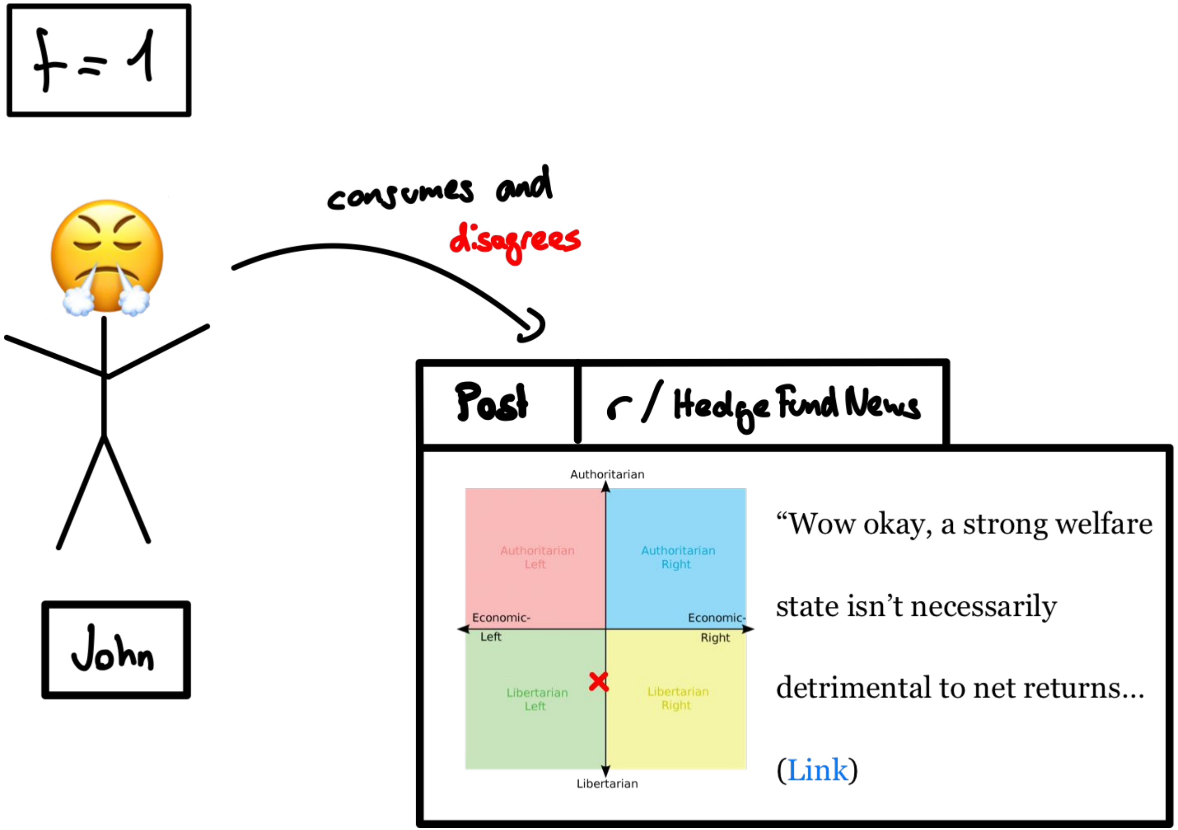
$diff = \|(\vec{v} - \vec{u}) * \vec{i}\|$, where * denotes component-wise multiplication.

Agreement



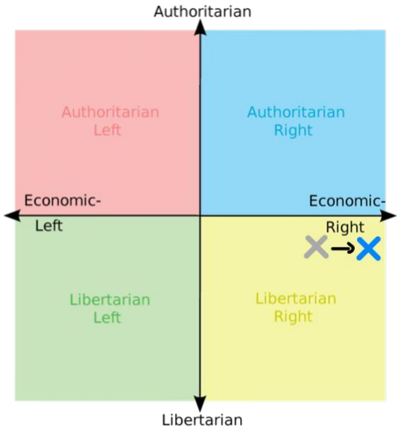
```
if diff < 0.1 * get_sqrt_n():
    new_bias = np.clip(self.bias + 0.1 * (post.bias - self.bias), 0, 1)
elif diff > get_sqrt_n() * (1 - 0.1):
    new_bias = np.clip(self.bias - 0.1 * (post.bias - self.bias), 0, 1)
```

Our Model - Interactions

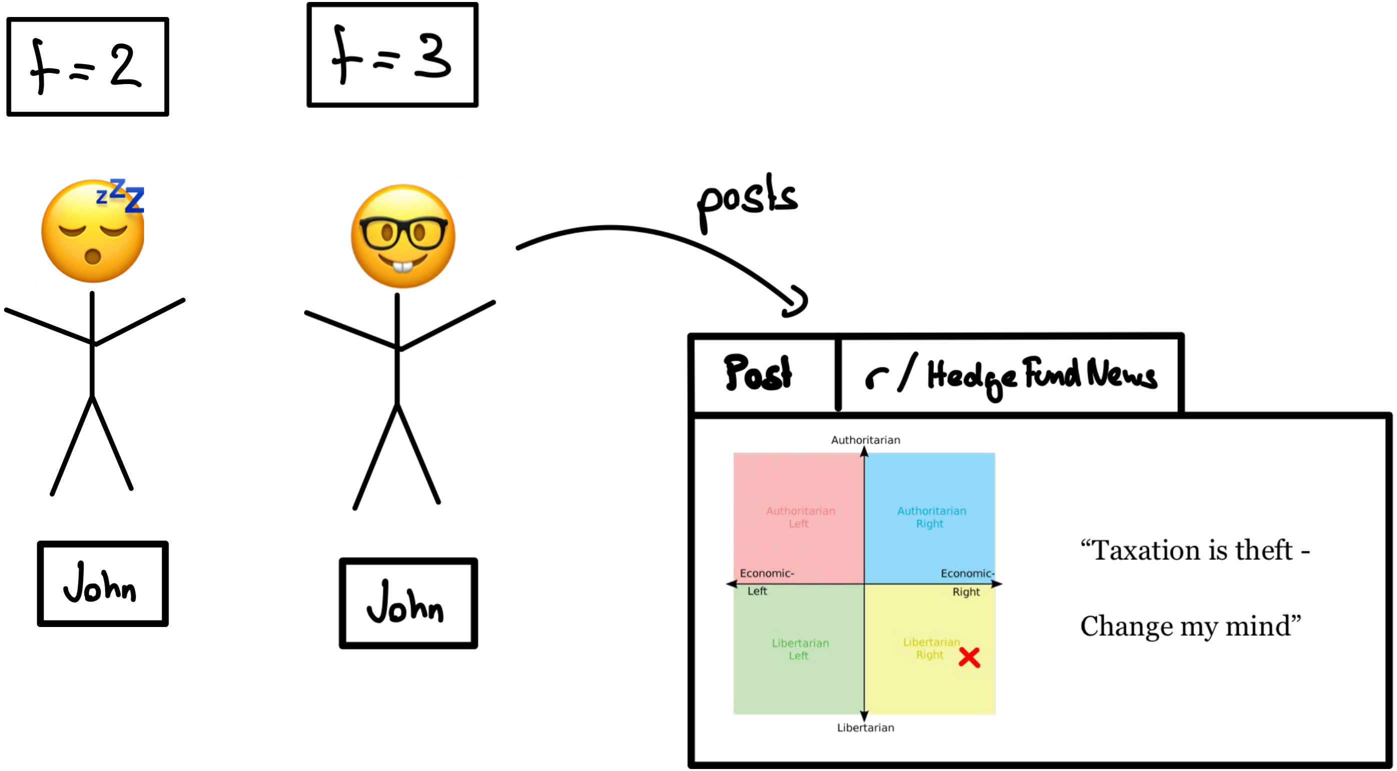


“Wow okay, a strong welfare state isn’t necessarily detrimental to net returns…

(Link)



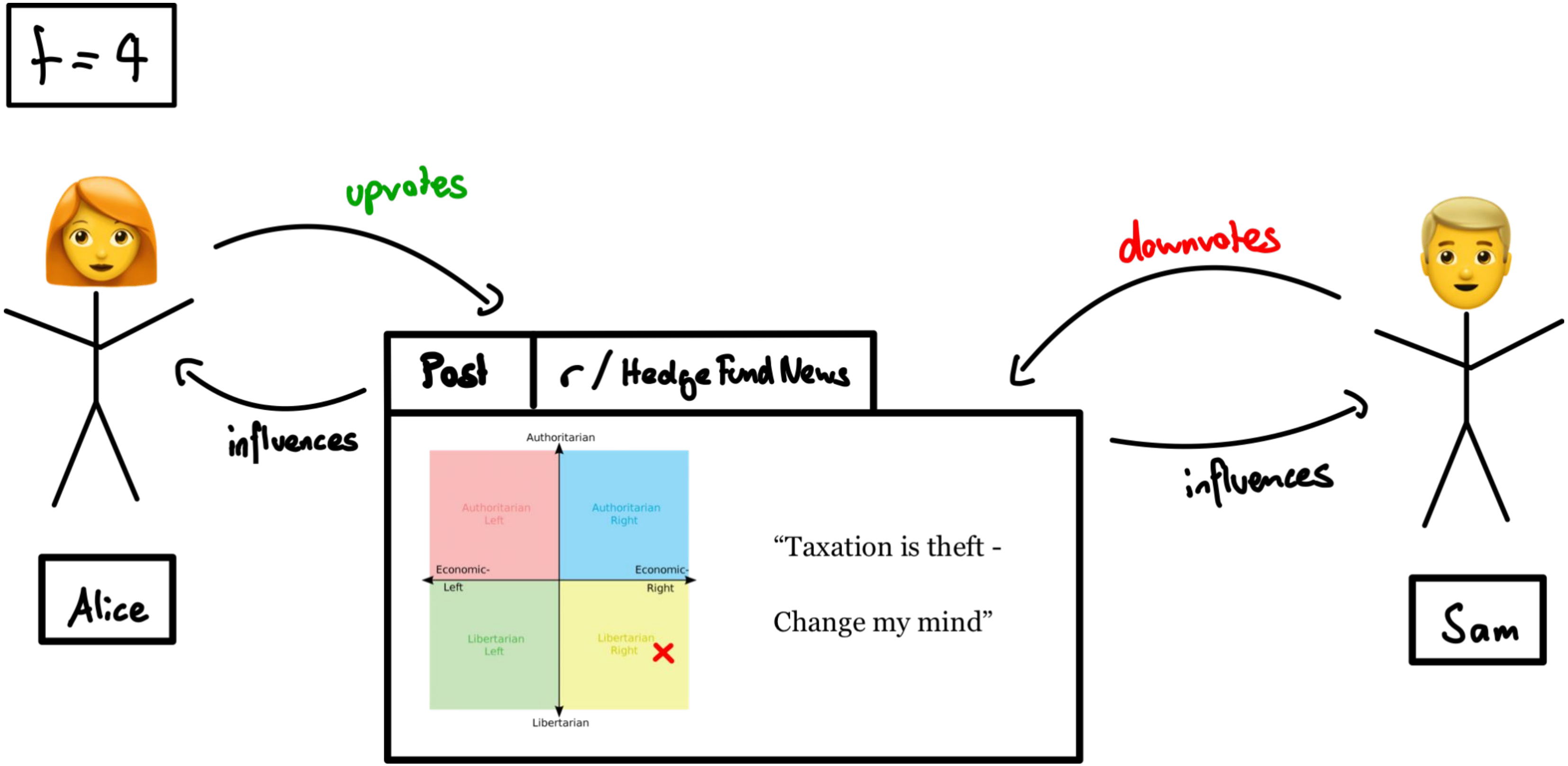
Our Model - Interactions



"Taxation is theft -

Change my mind"

Our Model - Interactions

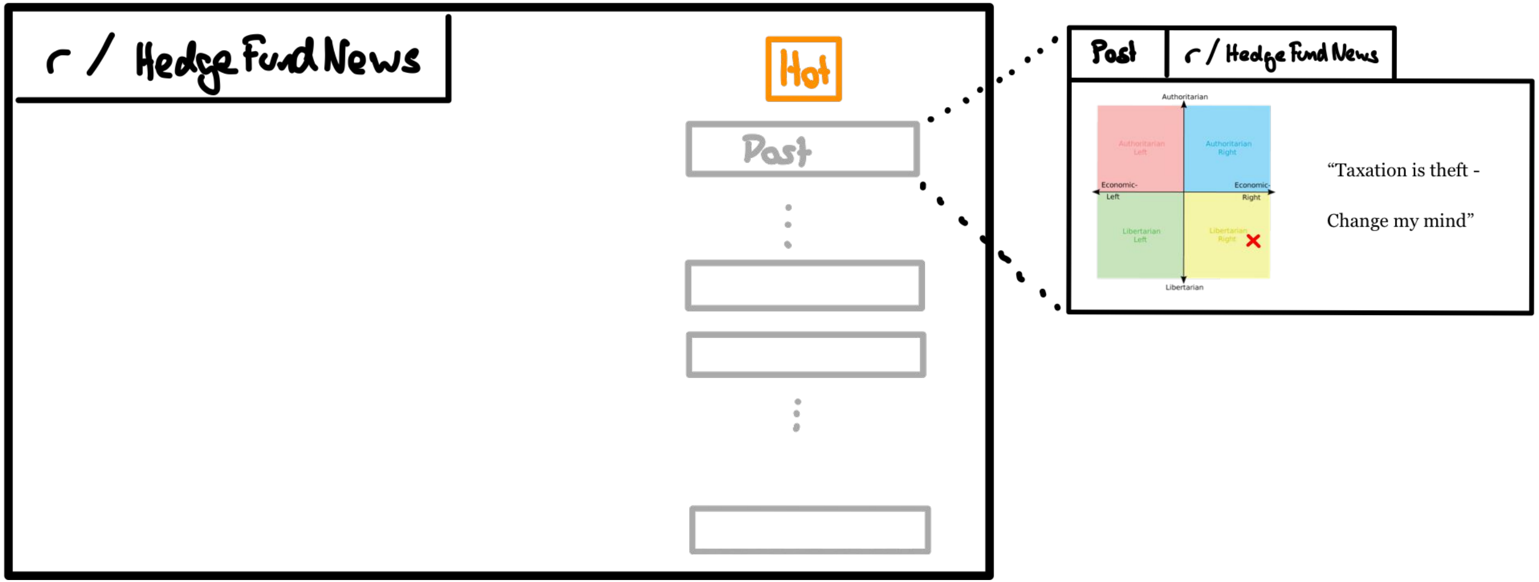


“Taxation is theft -

Change my mind”

Our Model - Interactions

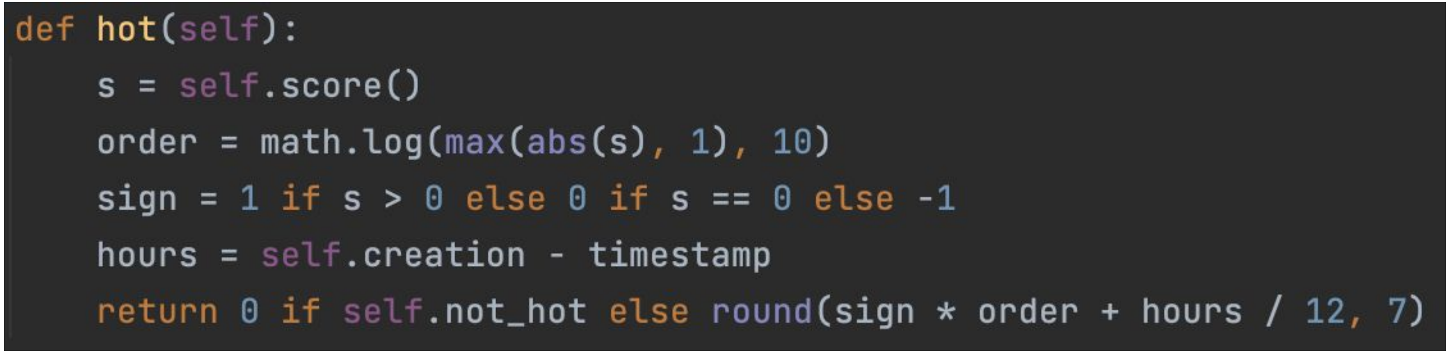
$f = 5$



"Taxation is theft -

Change my mind"

Reddit Hot Function

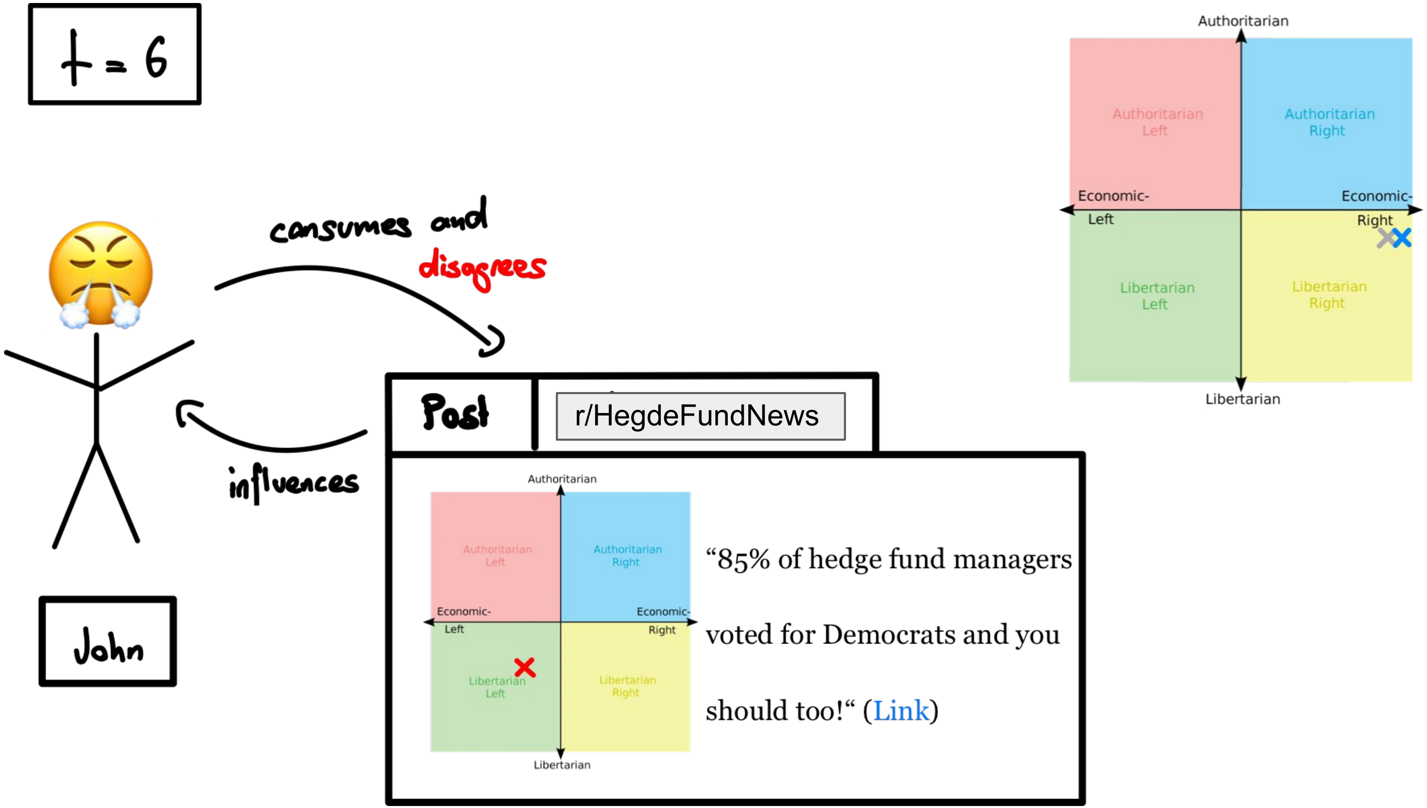


```
def hot(self):
    s = self.score()
    order = math.log(max(abs(s), 1), 10)
    sign = 1 if s > 0 else 0 if s == 0 else -1
    hours = self.creation - timestamp
    return 0 if self.not_hot else round(sign * order + hours / 12, 7)
```

$$hot = \log_a(s) - h/12$$

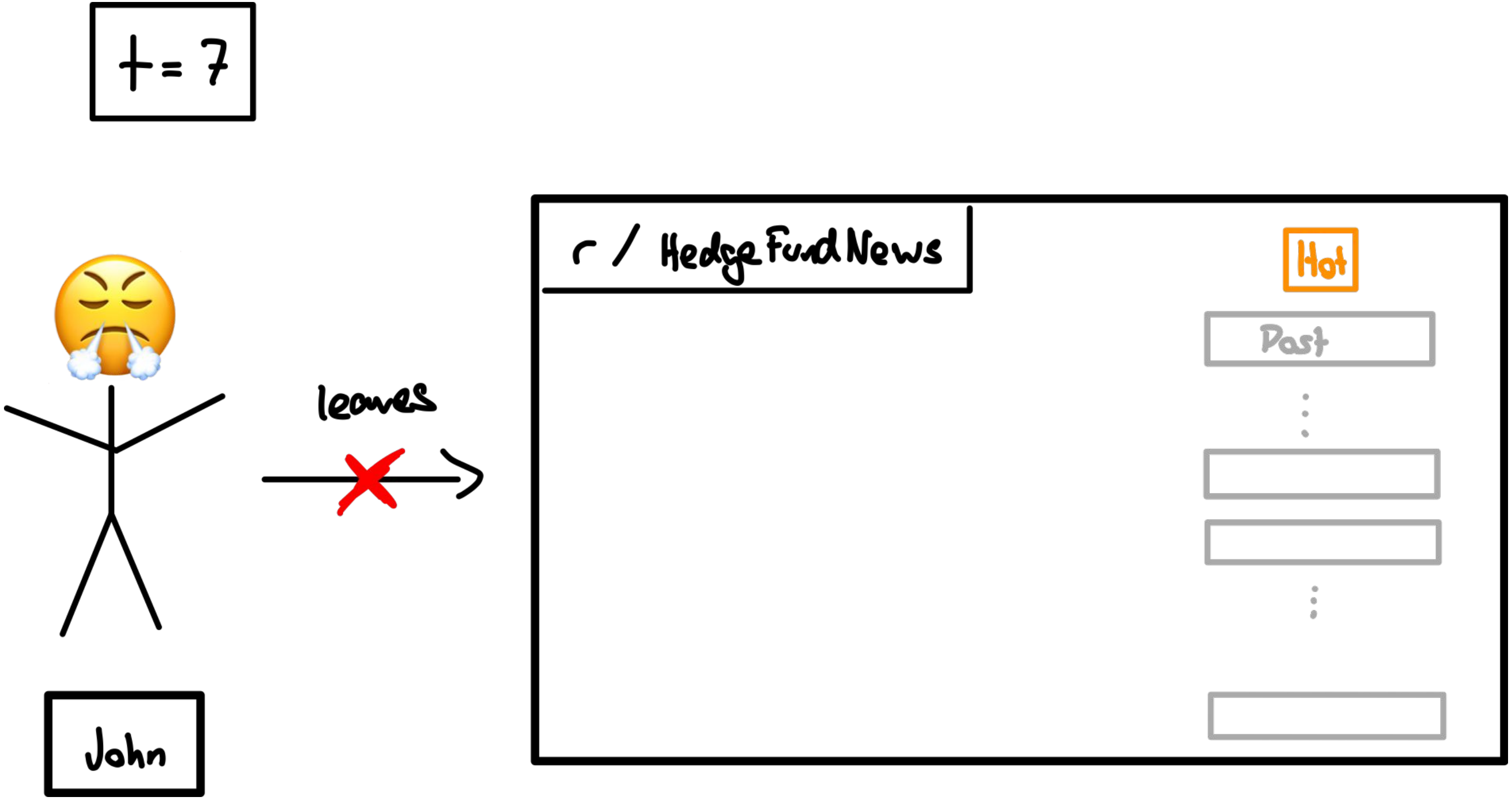
https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9

Our Model - Interactions



“85% of hedge fund managers voted for Democrats and you should too!“ (Link)

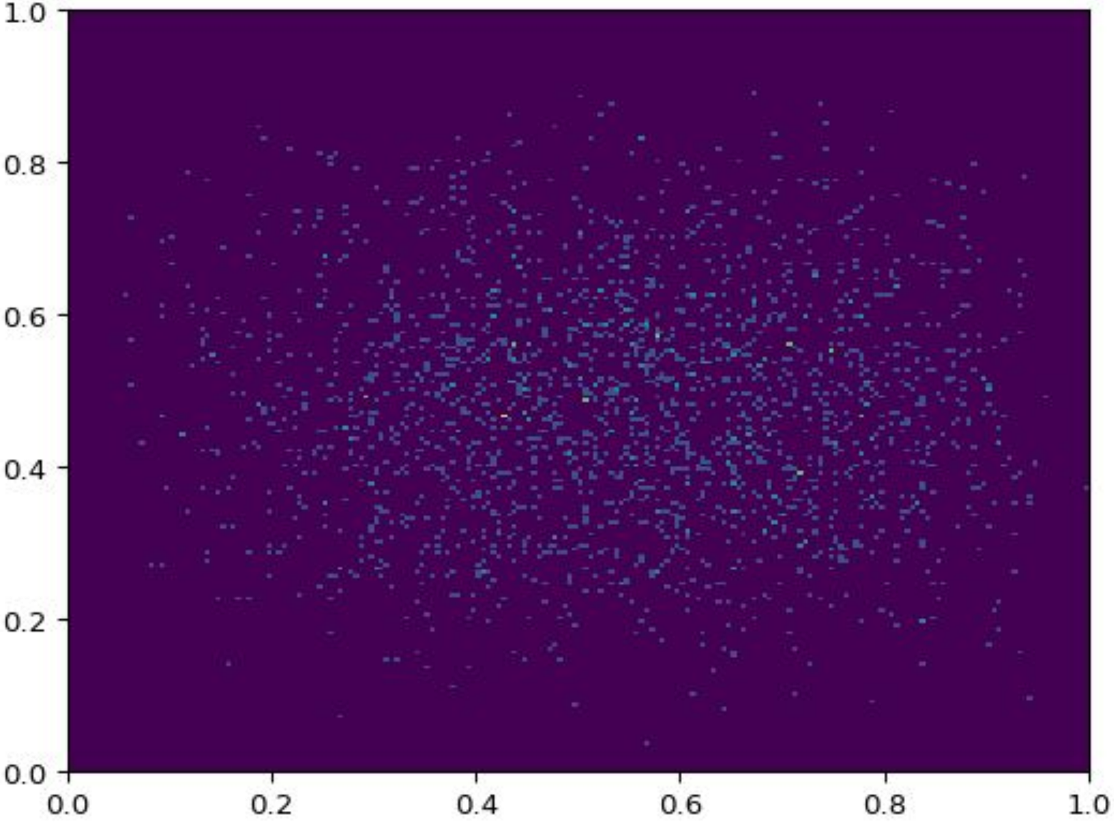
Our Model - Interactions



The outcome of our model (without moderation)

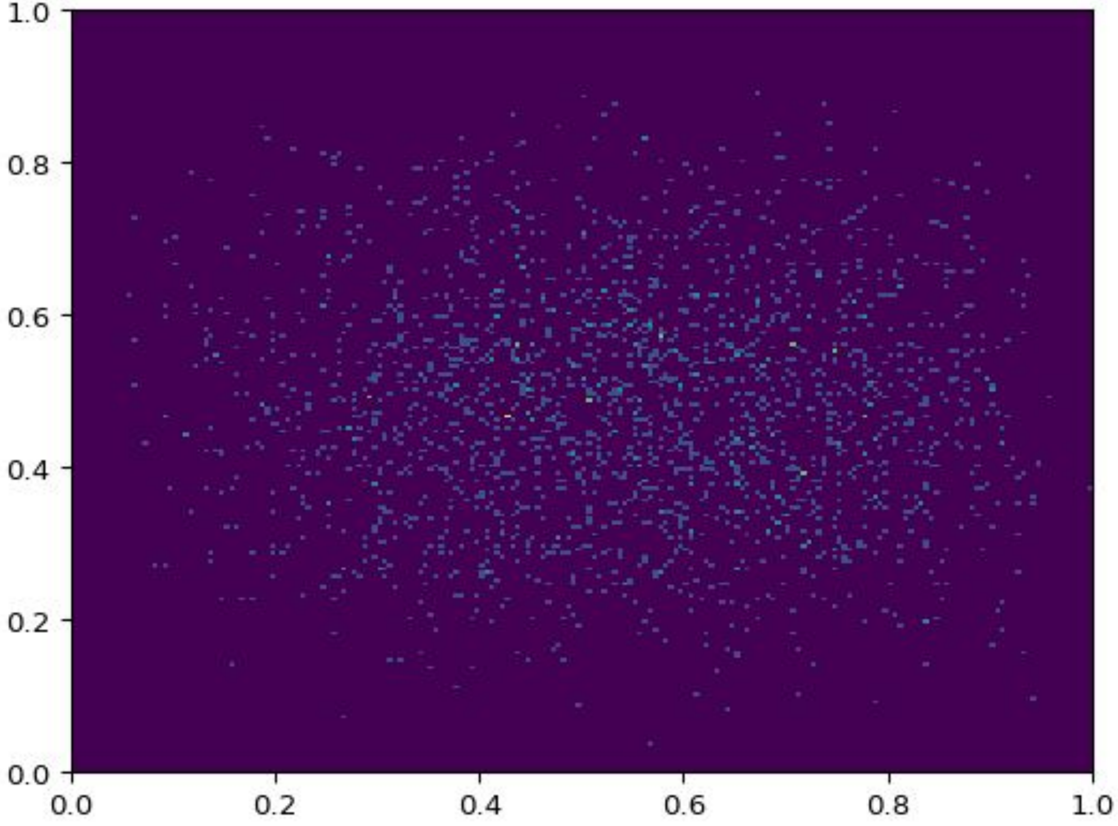
User Bias Distribution

- 2’000 User
- 200 Subreddits
- 25x6 Rounds

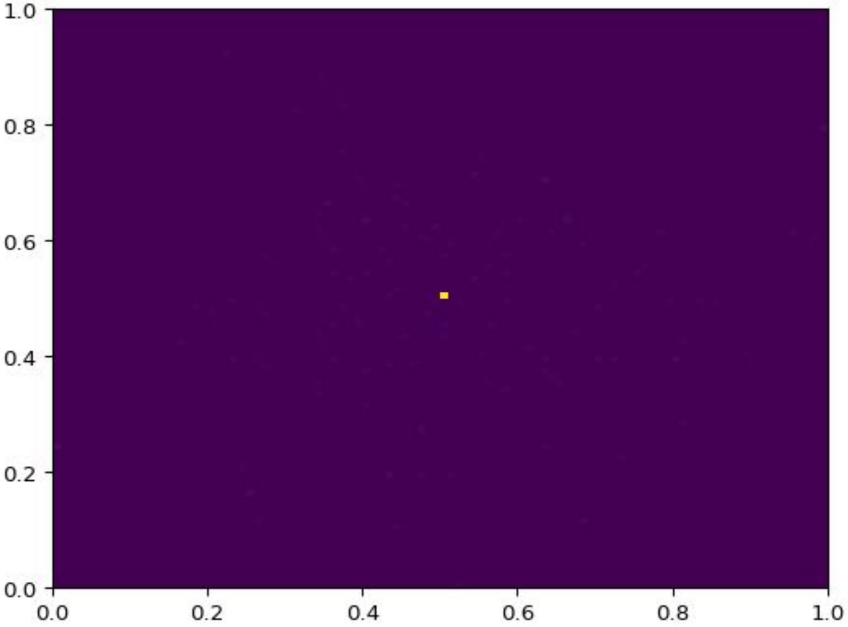


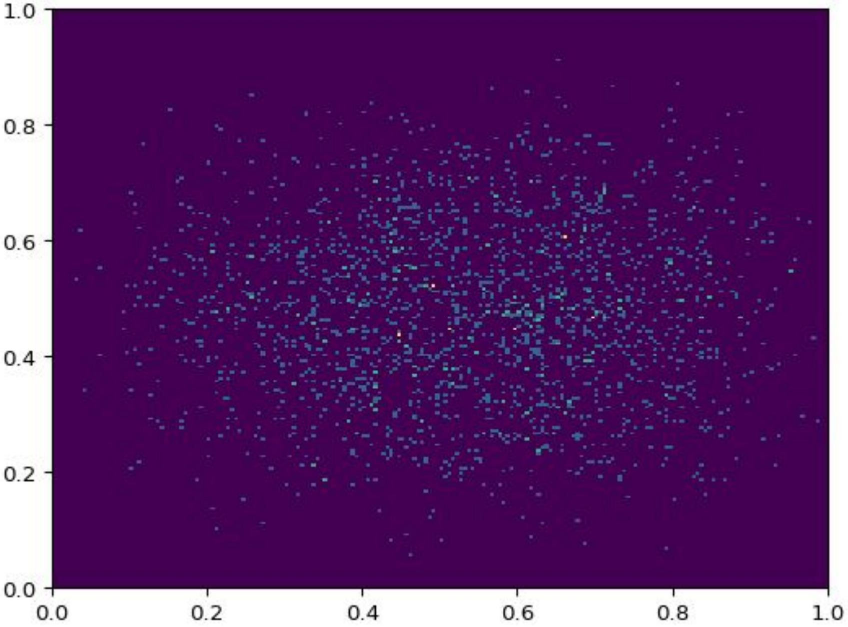
User Bias Distribution

- 2’000 User
- 200 Subreddits
- 25x6 Rounds

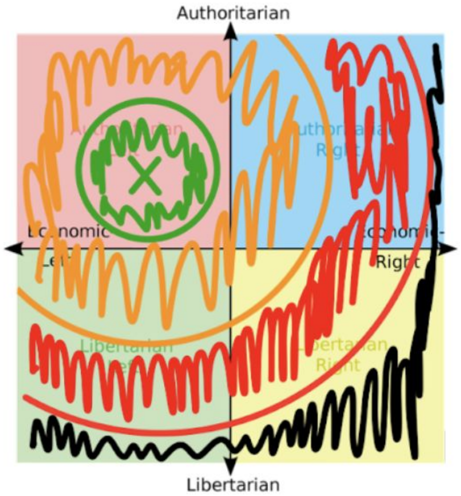


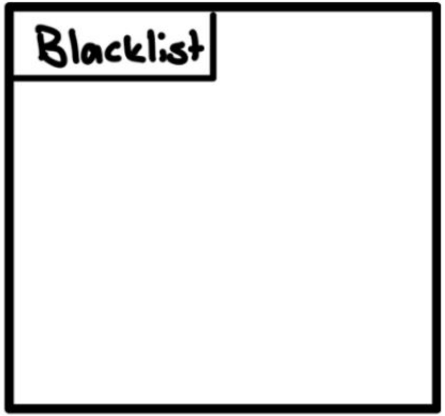
Correlation User Subreddit





Moderation

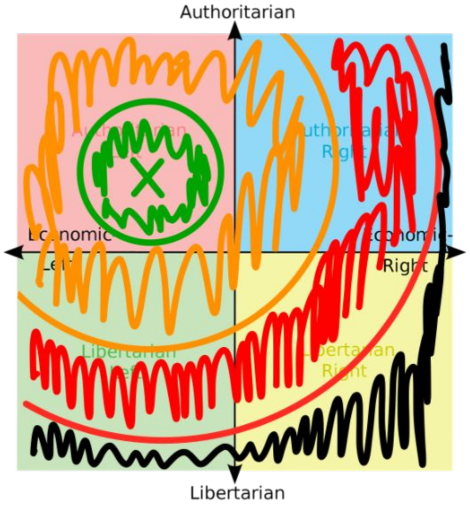




- Green Zone: Nothing happens
- Orange: Post will be removed from Hot Queue
- Red Zone: User will be blacklisted
- Black Zone: User gets banned from all subreddits (24 rounds)

Our Model - Moderation

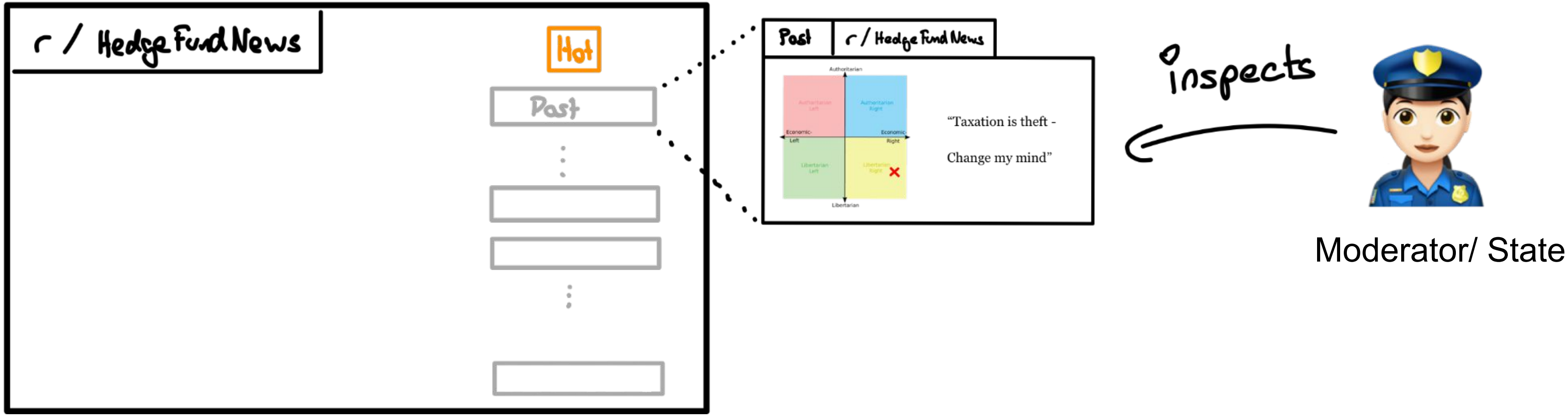
t = 8



Blacklist

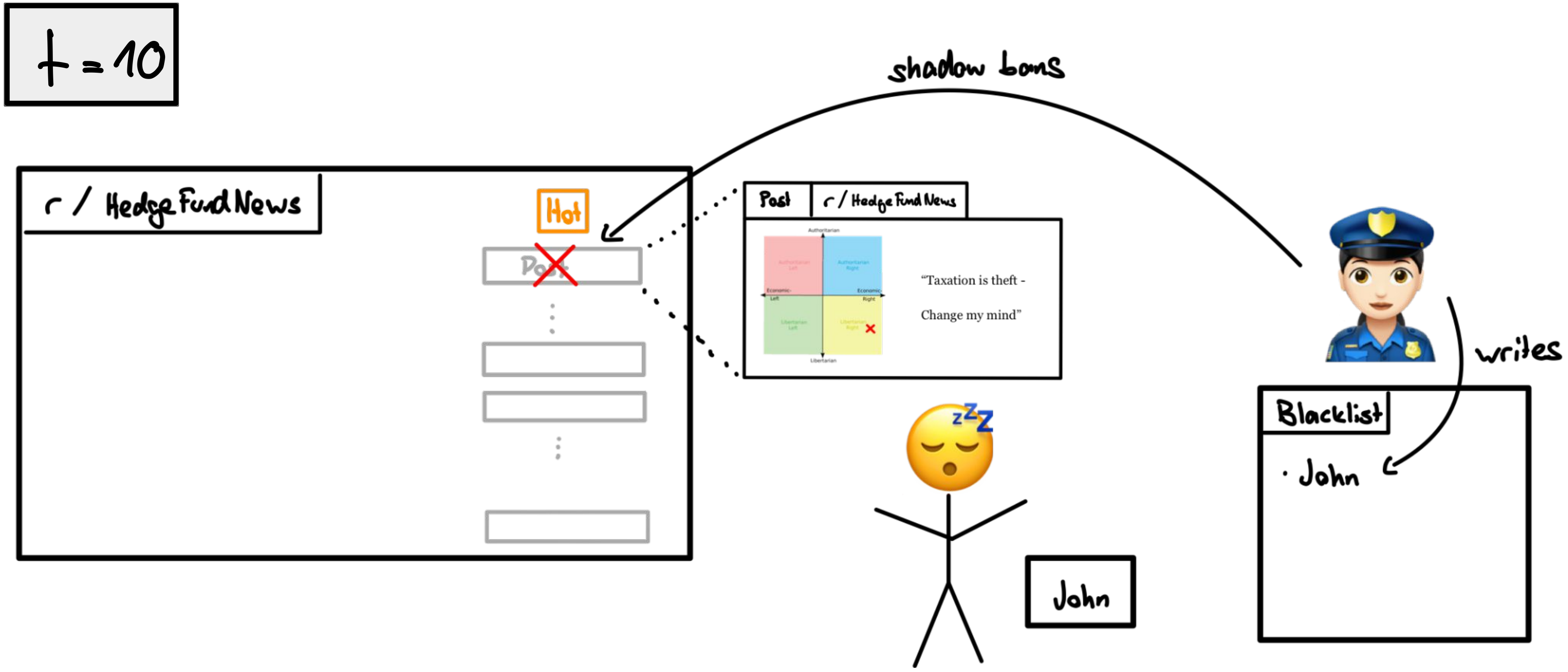
Our Model - Moderation

$$\boxed{f = g}$$



Moderator/ State

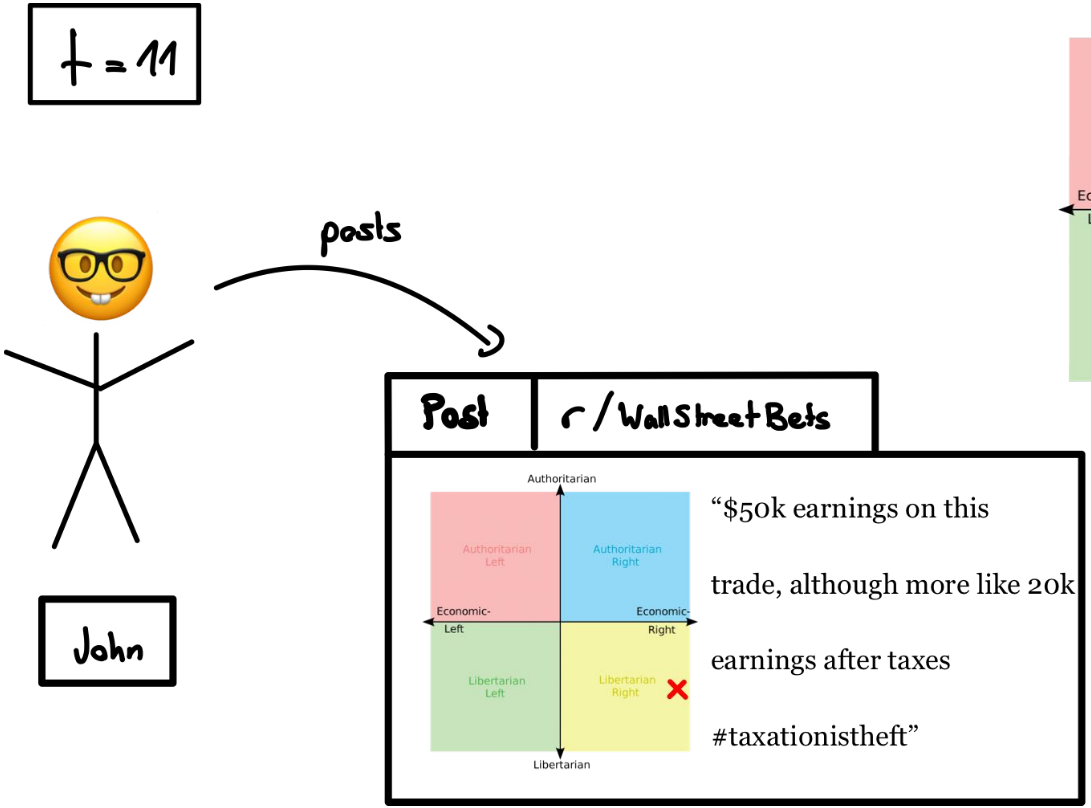
Our Model - Interactions



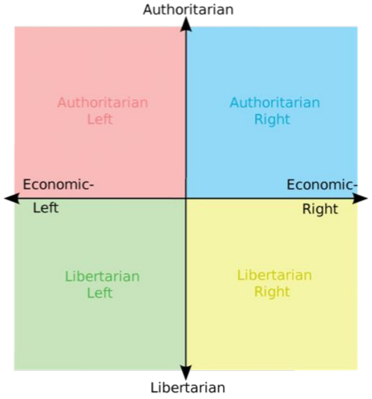
"Taxation is theft -

Change my mind"

Our Model - Moderation

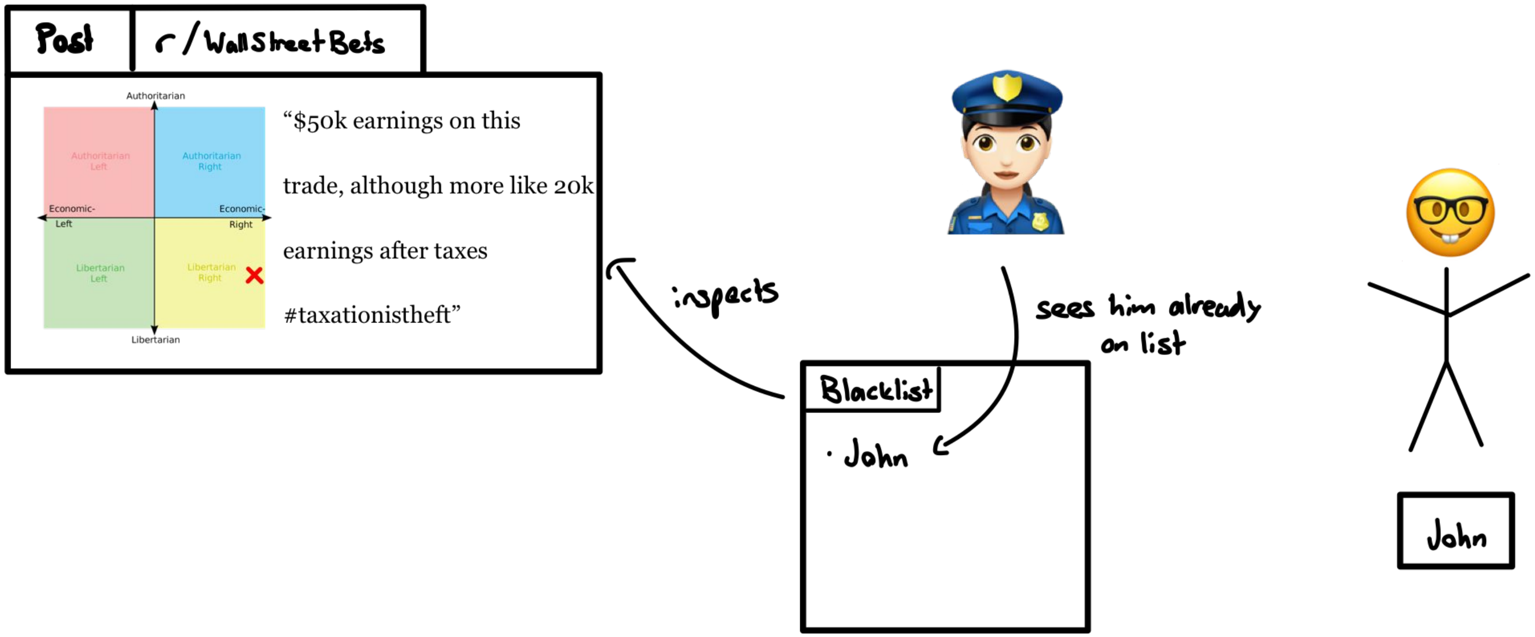


“$50k earnings on this trade, although more like 20k earnings after taxes #taxationistheft”



Our Model - Interactions

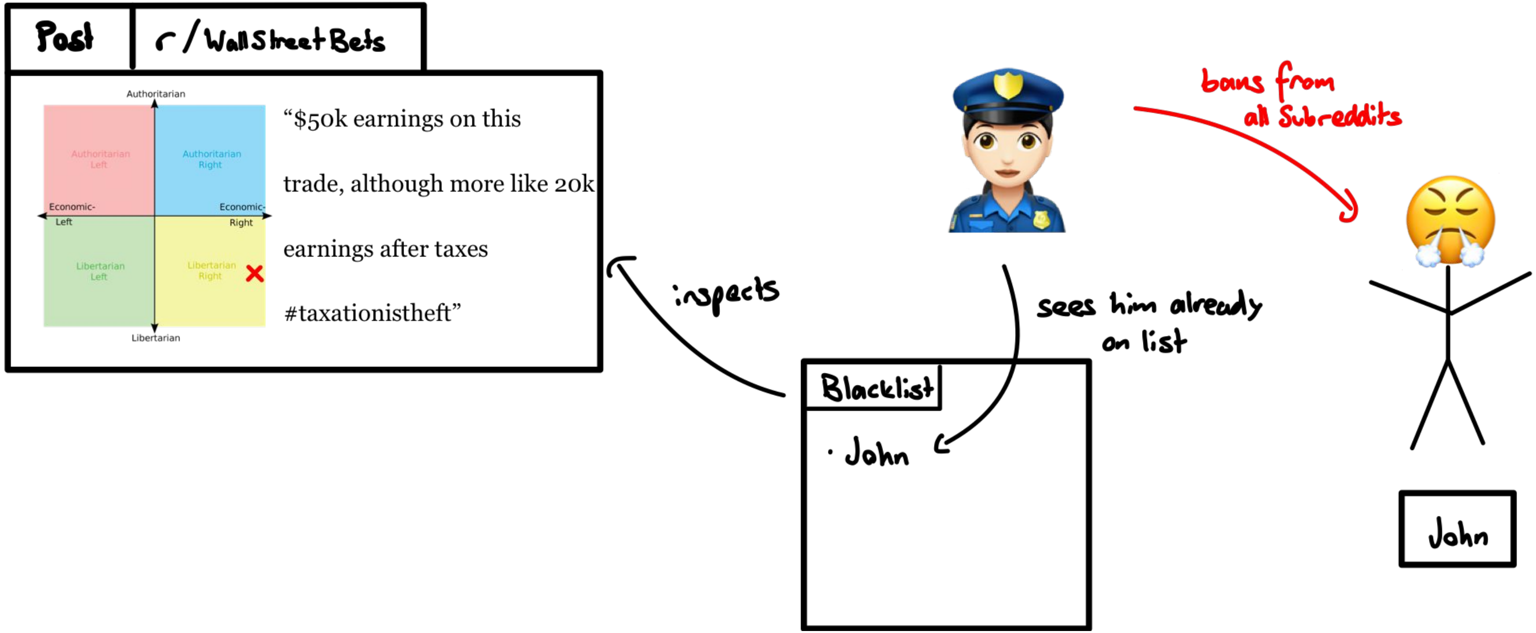
t = 12



“$50k earnings on this trade, although more like 20k earnings after taxes #taxationistheft”

Our Model - Interactions

t = 12



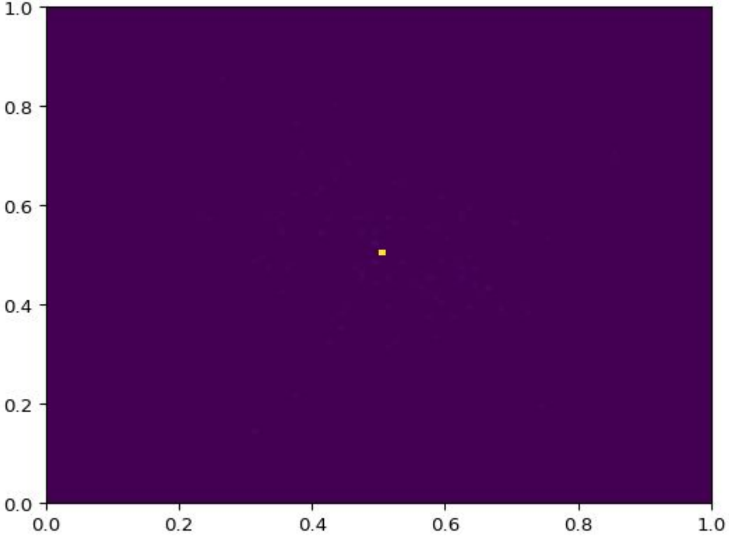
“$50k earnings on this trade, although more like 20k earnings after taxes #taxationistheft”

The Outcome of the Model with Moderation

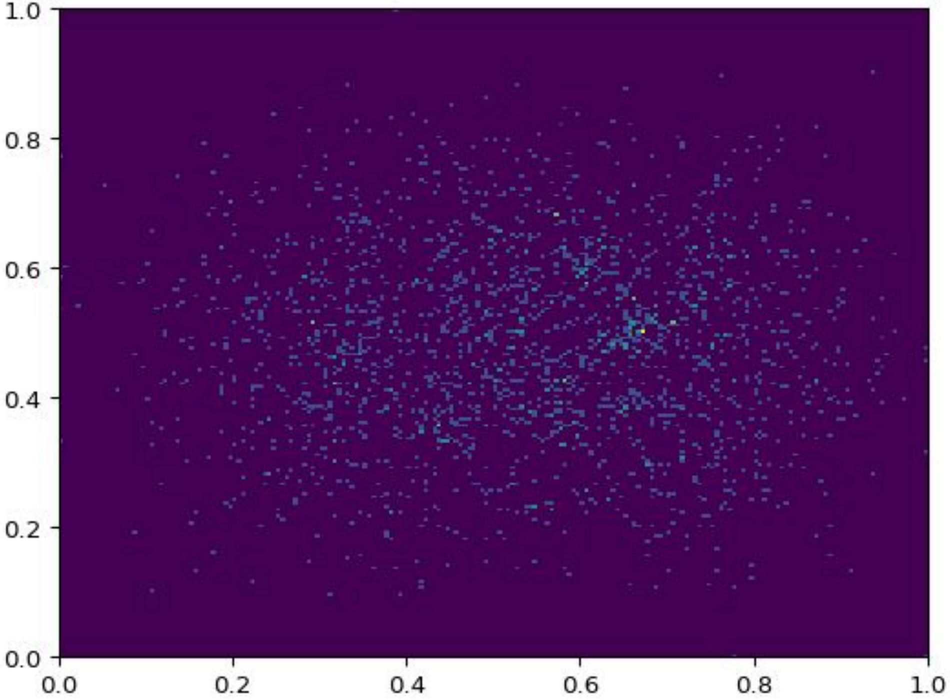
With Moderation

- 2’000 User
- 200 Subreddits
- 25x12 Rounds

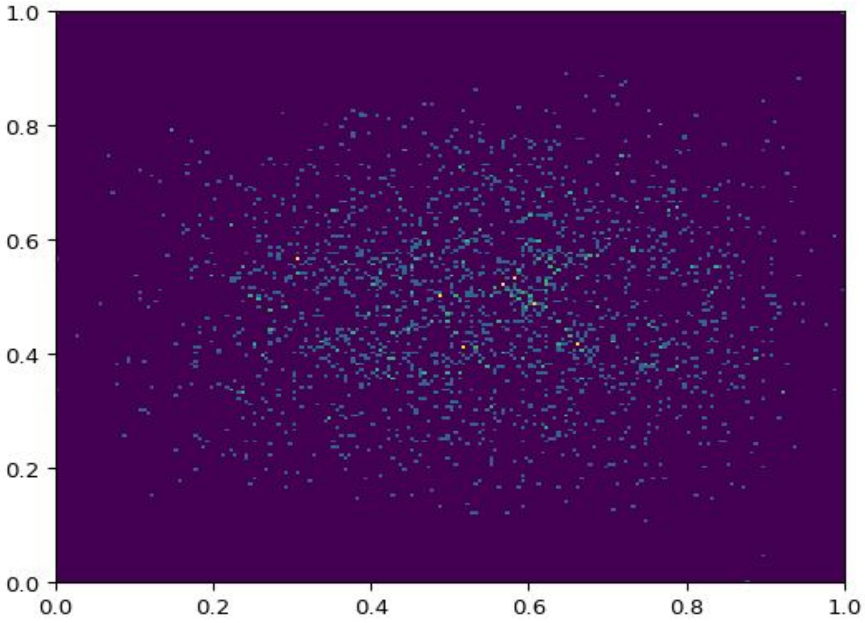
Subreddits

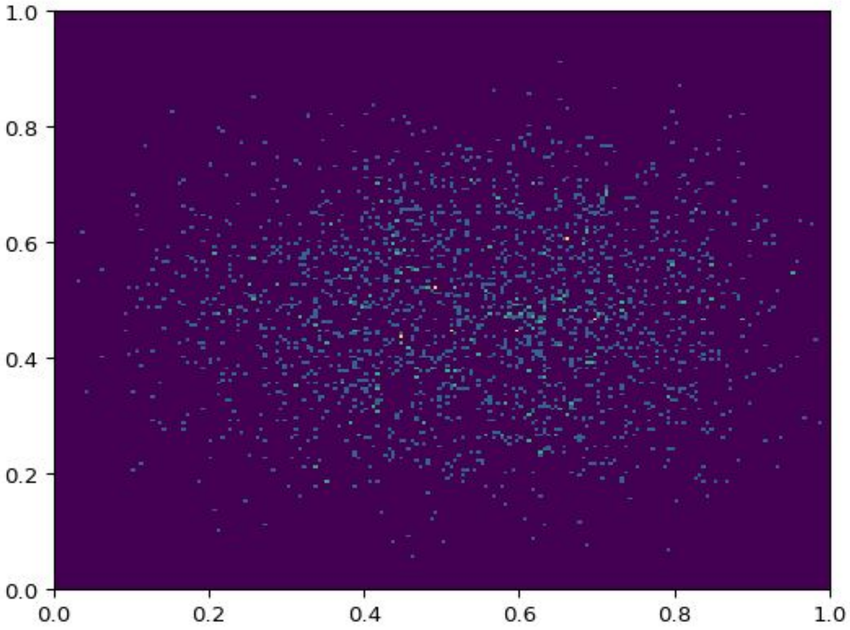


Agent opinions



Comparison Moderation / Without Moderation





Our Model: Assumptions

Inaccuracies:

- Post treated as one entity (no differentiation between text, photos and videos)
- Opinion reduced to quantitative 2D-Vector
- Agents can't develop tactics to spread their opinions (could be done with Machine Learning)

Inspiration:

- Many models in this field used similar approaches:
  → Opinion model:
    - real Values between 0,1 (Deffuant et al., 2000)
    - multi dimensional bias values (Stauffer, 2005)
- Introducing noise through probabilistic outcomes in many methods (Pineda et al., 2009)

Thank you for your attention!!!

Any Questions left?