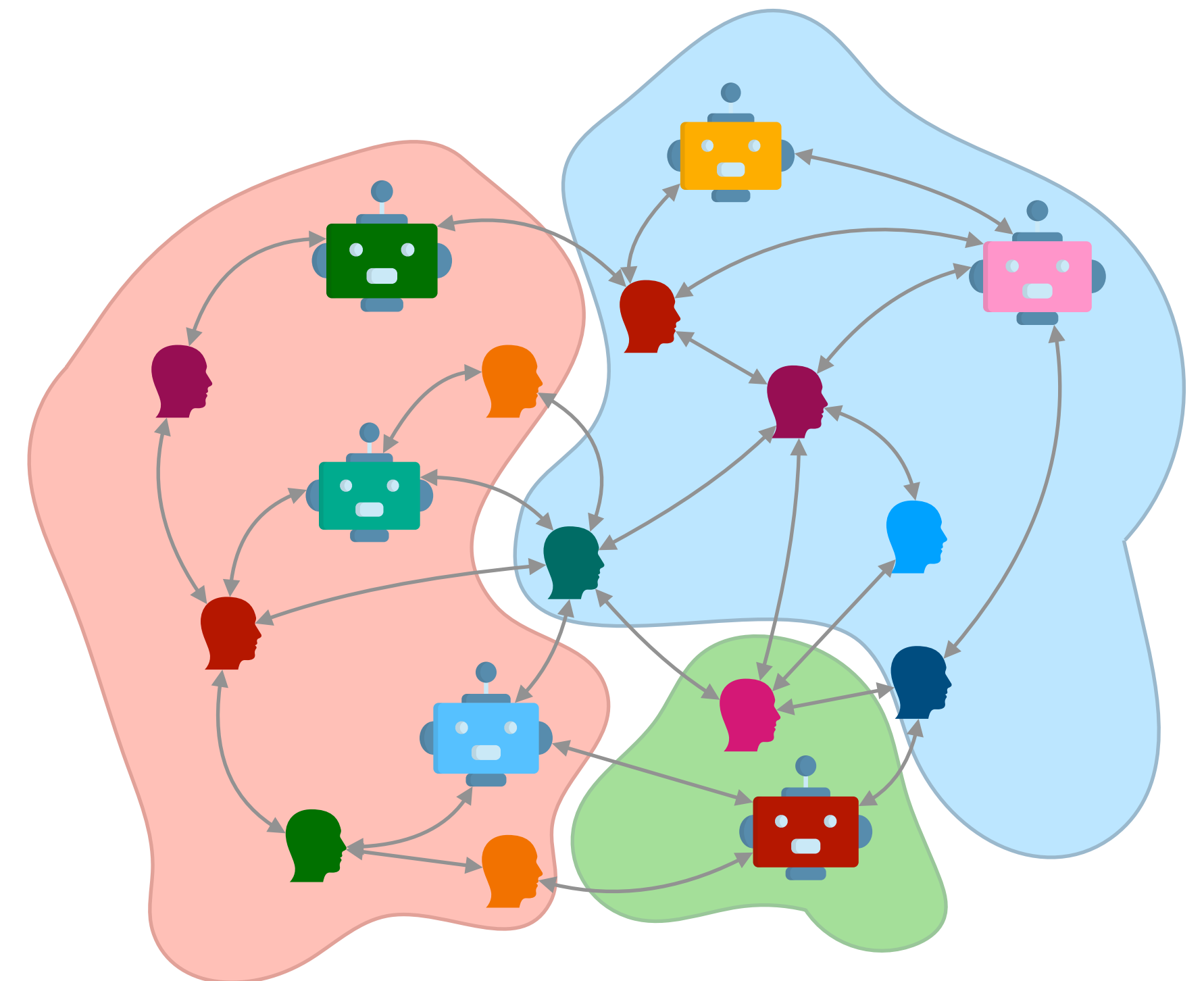
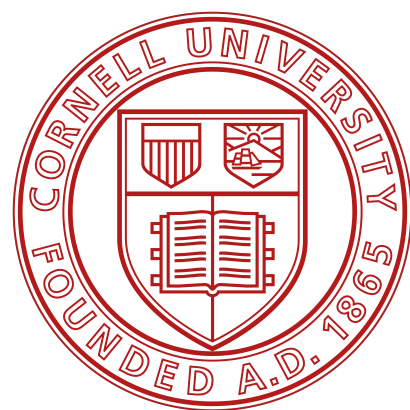
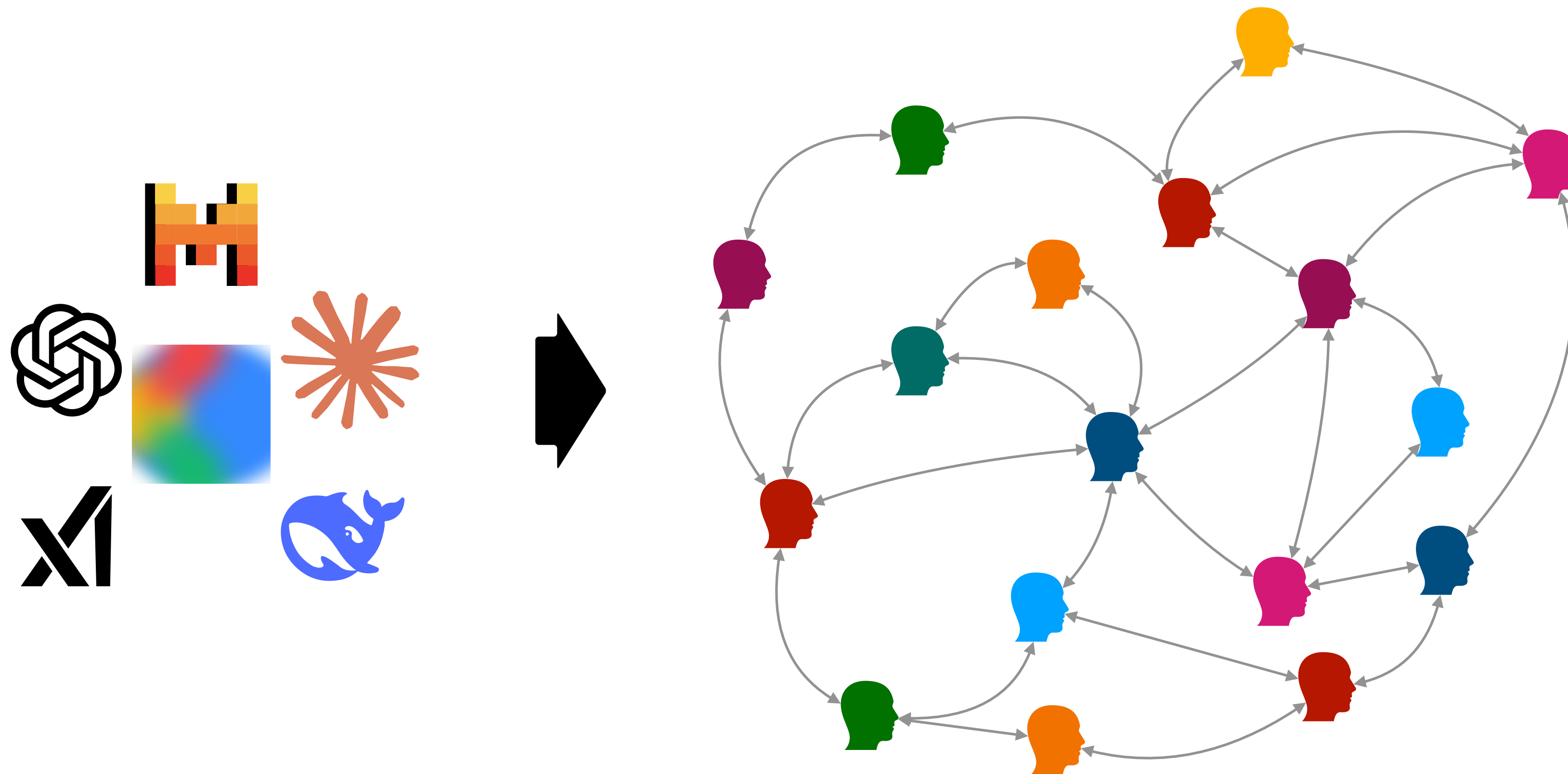


# An Experimental Method to Study Opinion Diffusion in Human-AI Hybrid Societies

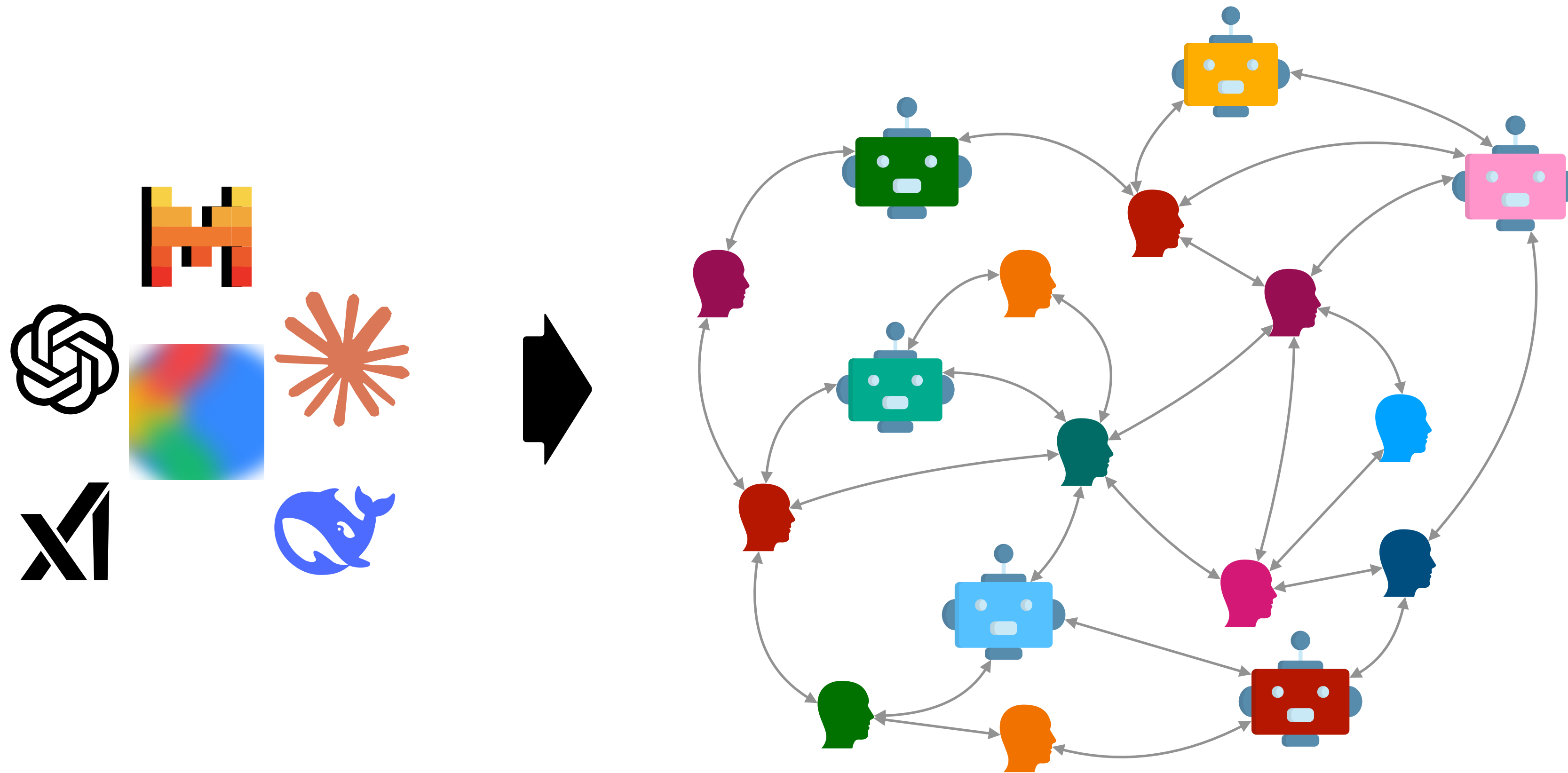
Léna Gaubert



# Our societies are increasingly hybrid

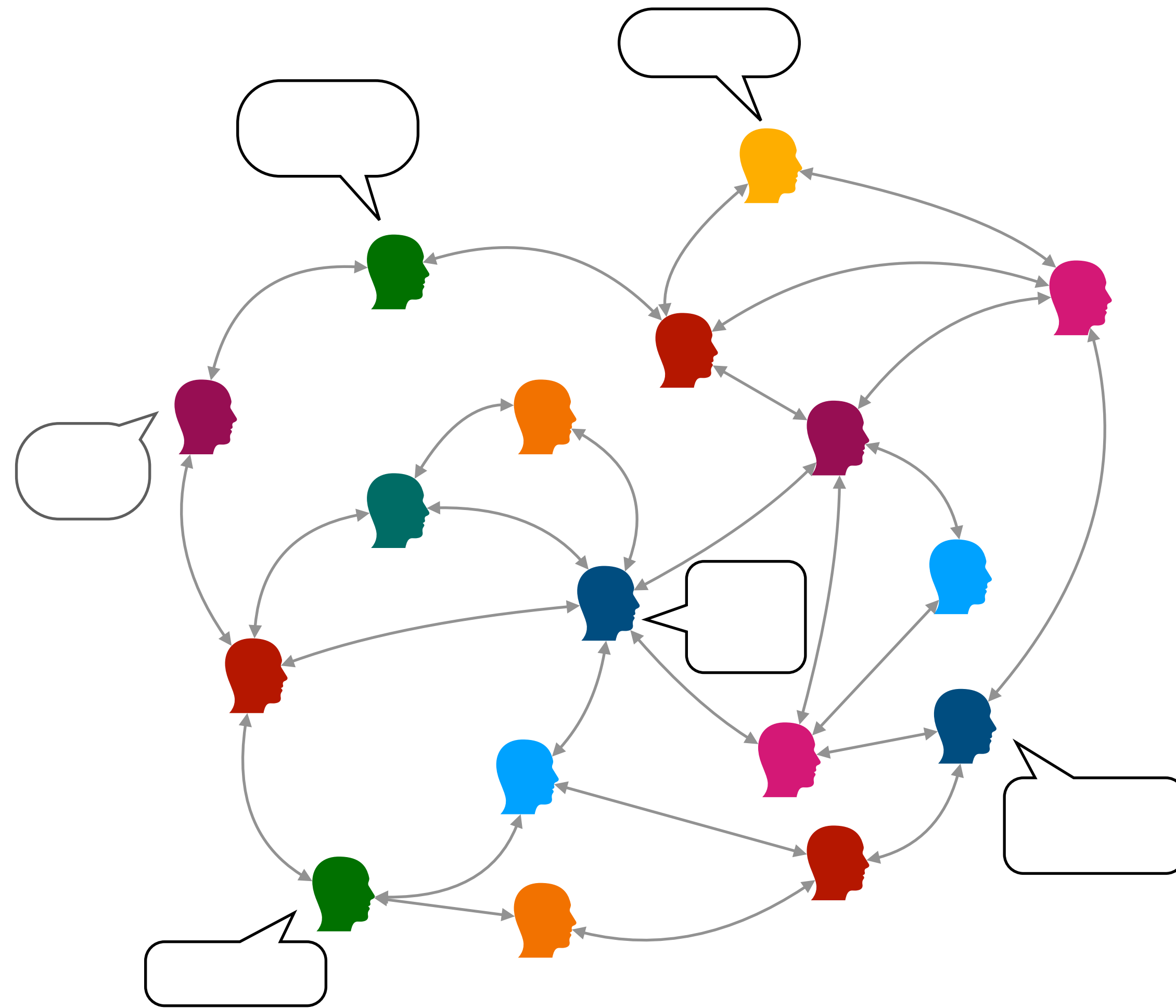


# Our societies are increasingly hybrid





# Online social behaviors

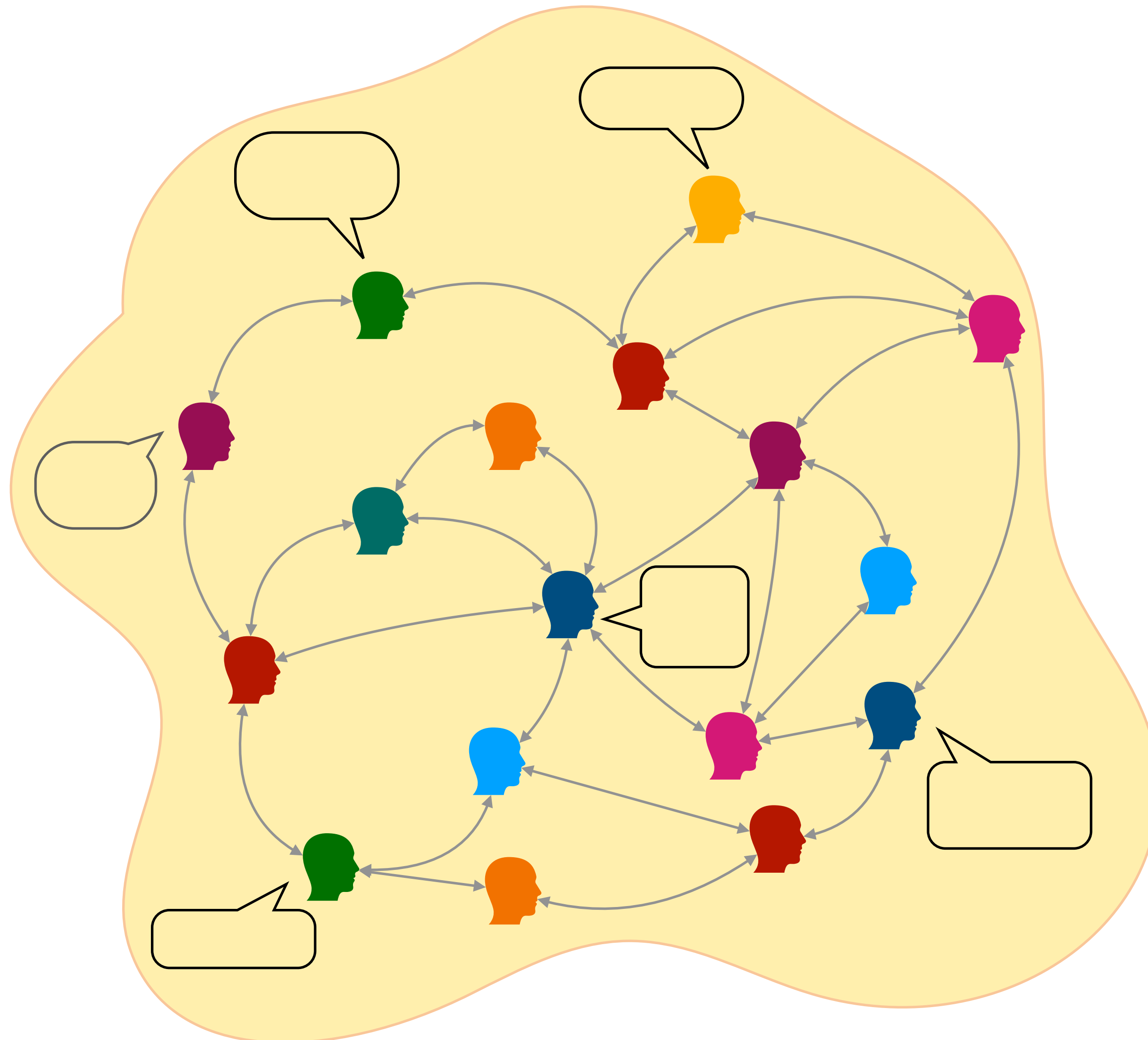




# Online social behaviors

One aspect of social behavior are **failure modes** such as **echo chambers**...

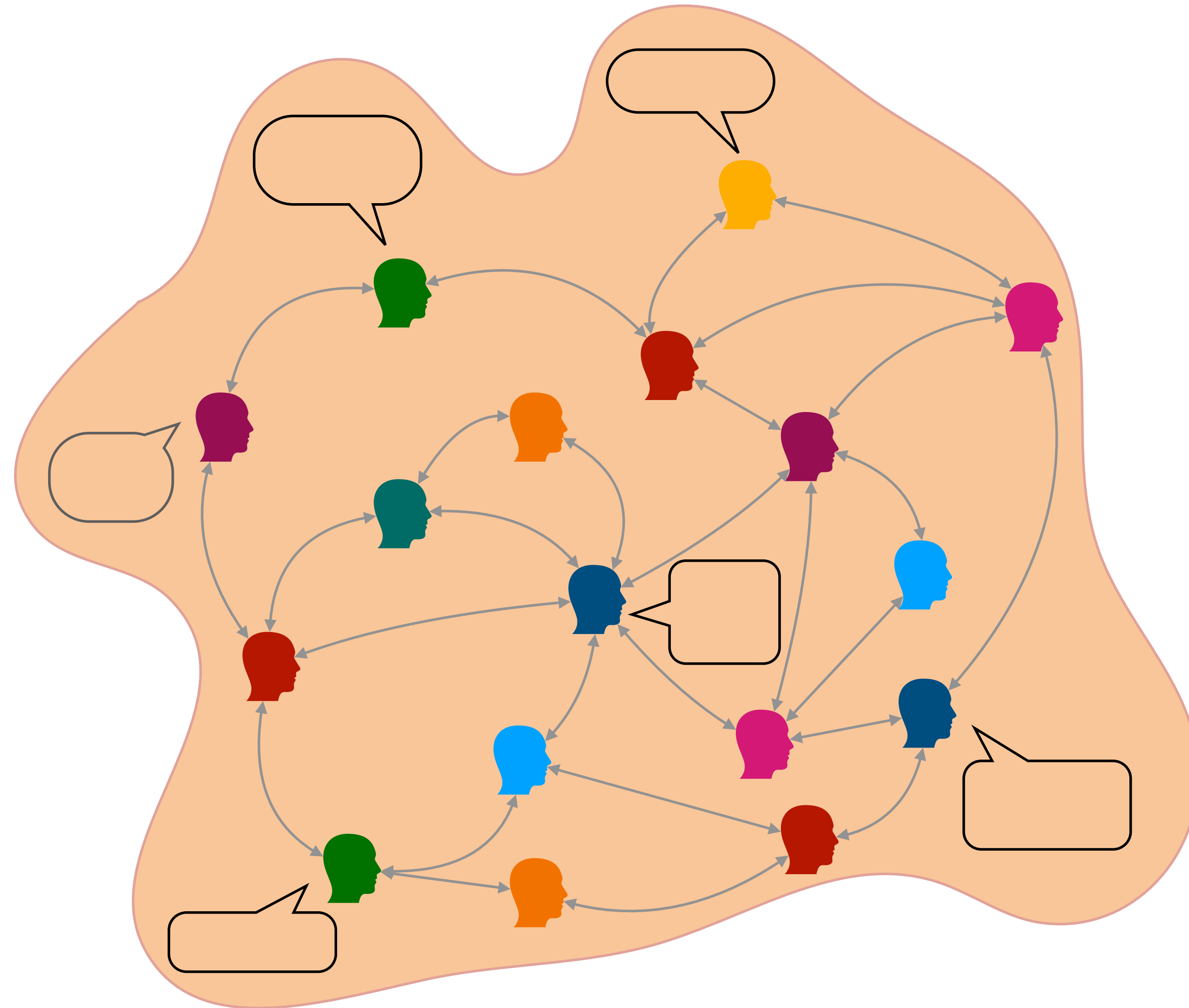
Garimella et al.  
(2018), Cinelli et al.  
(2021)



# Online social behaviors

One aspect of social behavior are **failure modes** such as **echo chambers, fake news...**

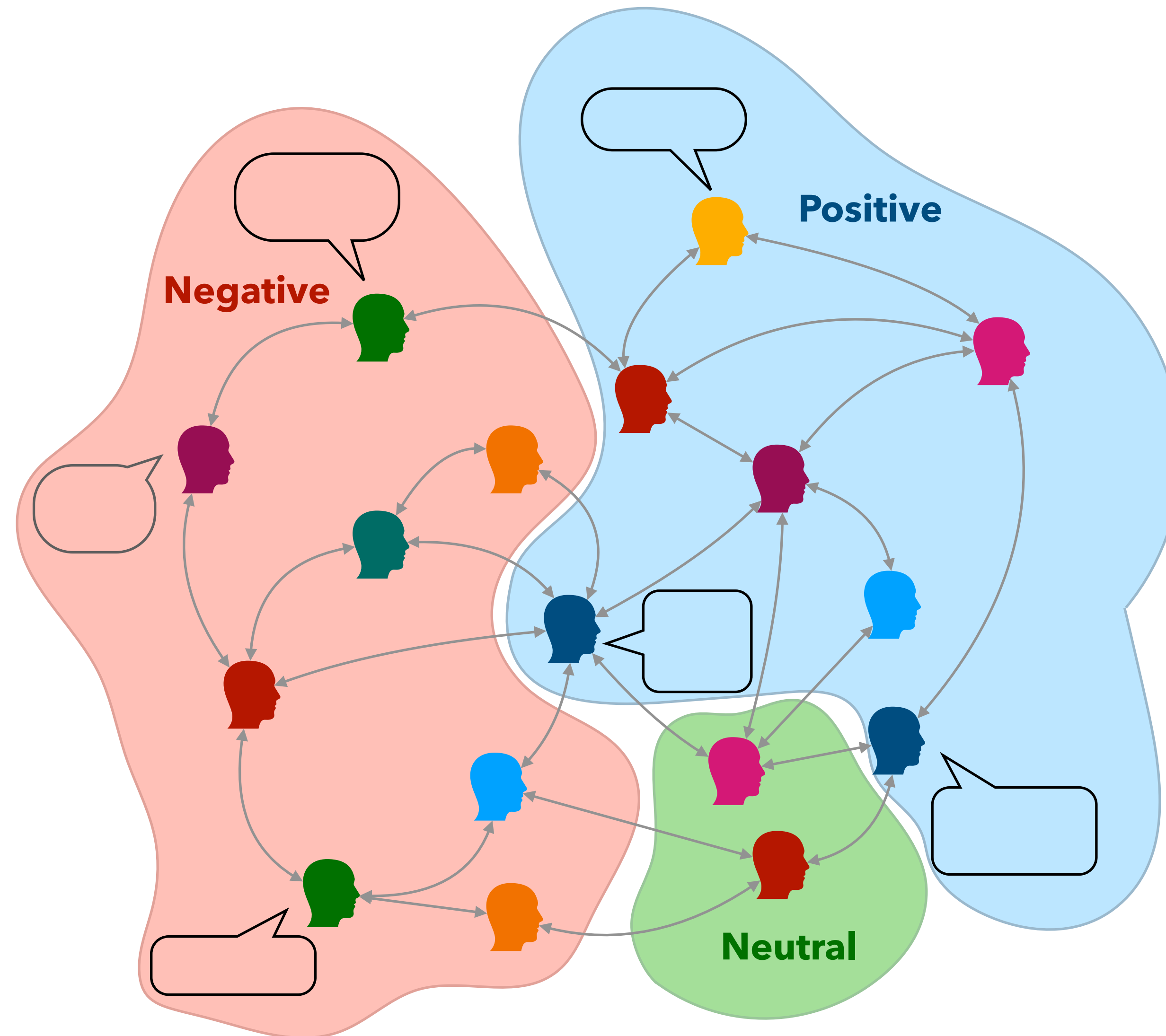
Aïmeur et al (2023)



# Online social behaviors

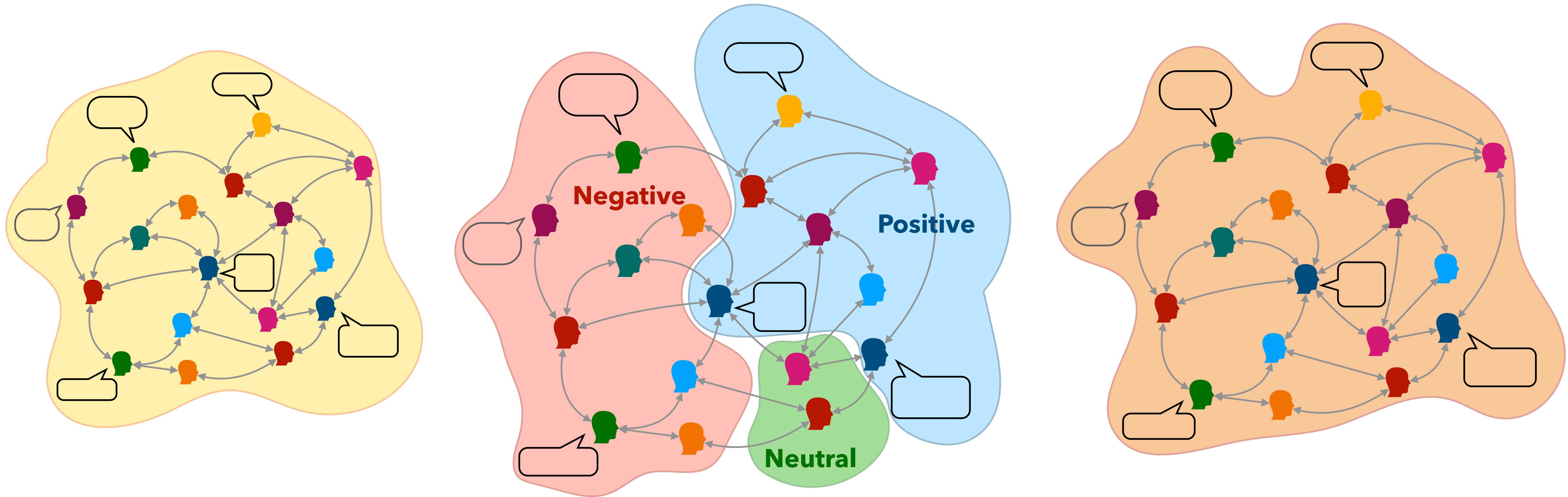
One aspect of social behavior are **failure modes** such as **echo chambers**, **fake news** or **polarization**.

Bramson et al. (2016),  
Pournaki (2025)





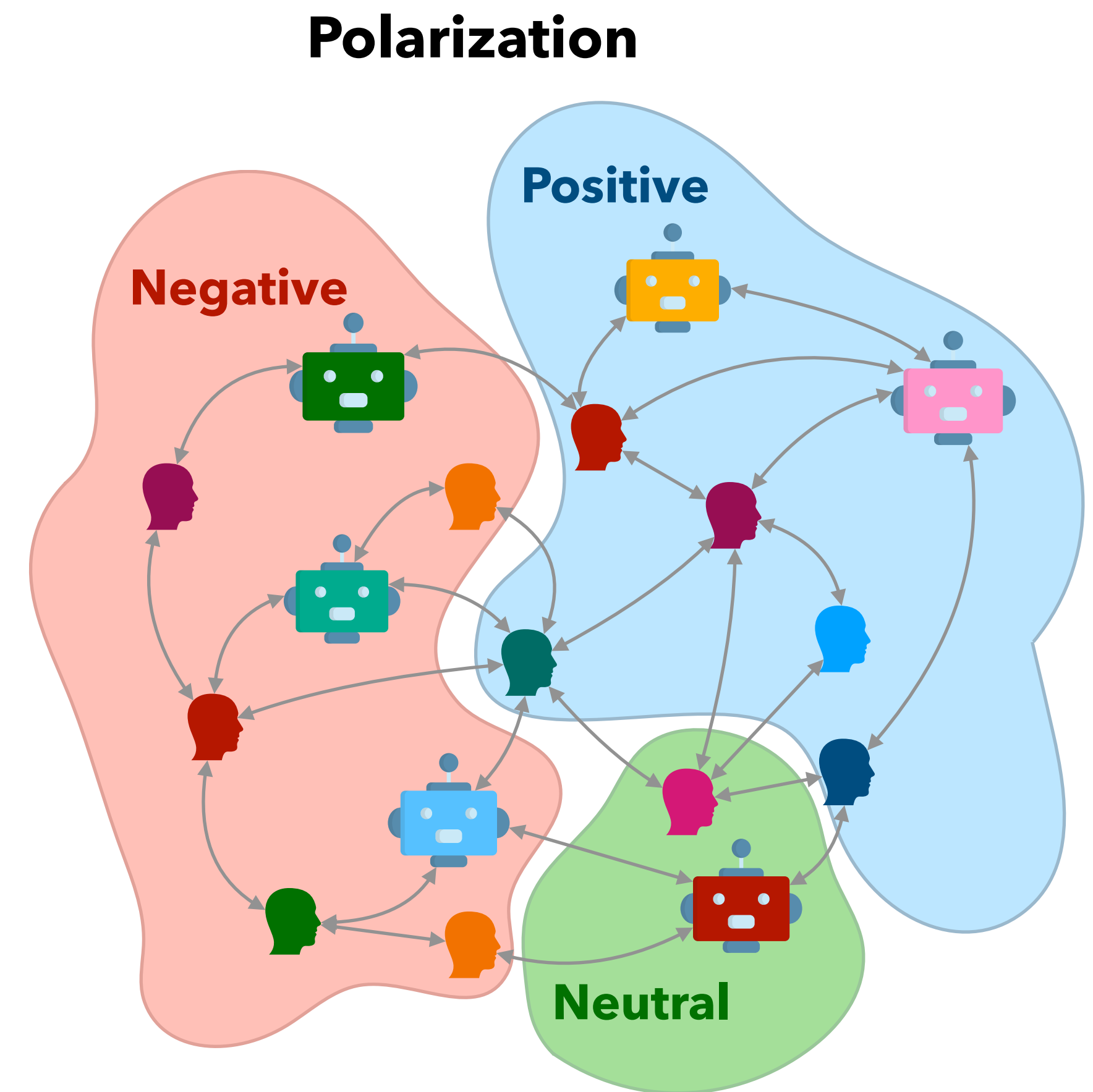
# Online social behaviors



Chitra & Musco (2020), Thurner et al. (2025), Zhao et al. (2023)

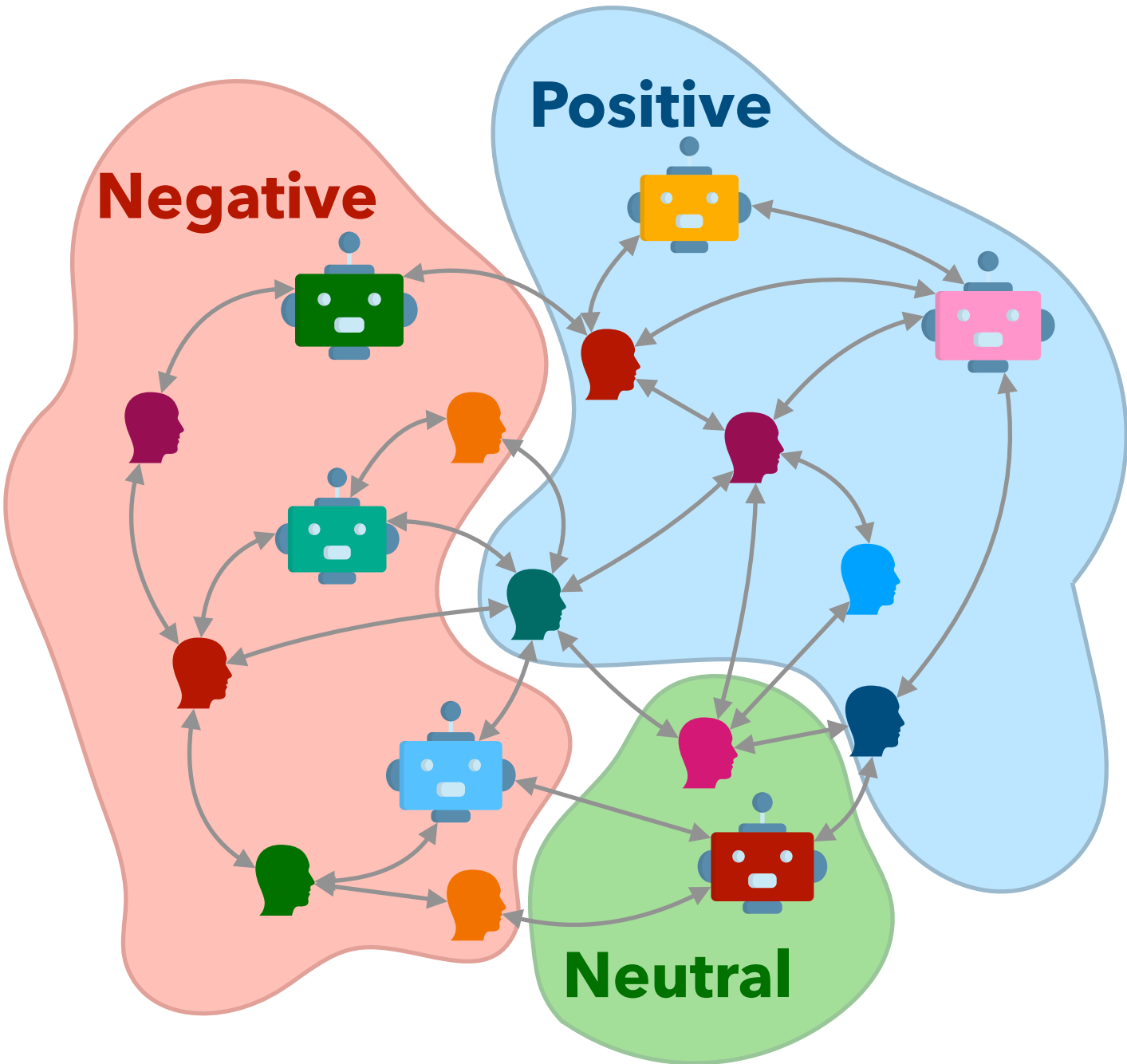
# Understanding Polarization in Hybrid Contexts

Does LLM-mediated discourse **dampen** or **amplify** polarization dynamics?

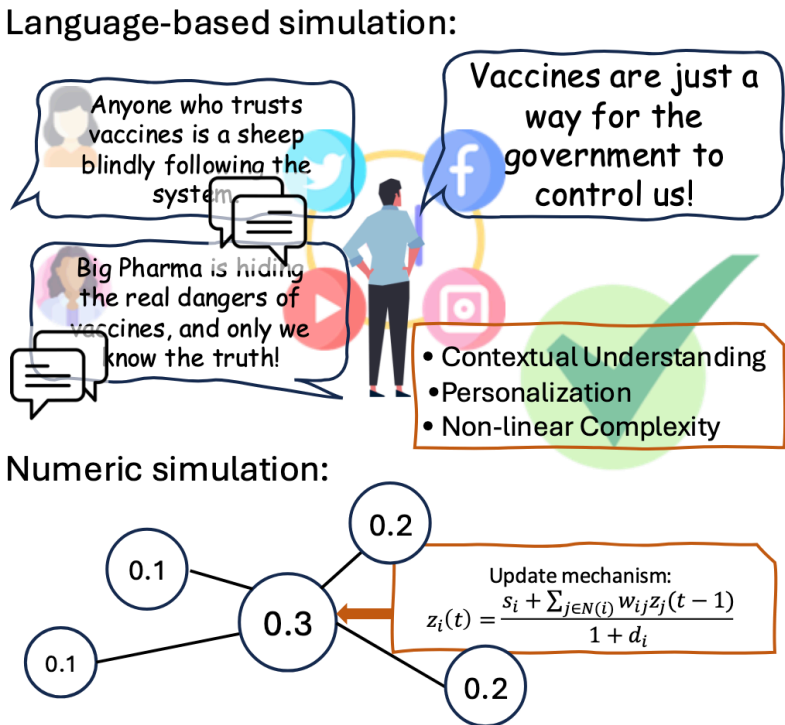




# The gap



Need for **controlled environments** to understand these emerging dynamics



Wang et al. (2025)

## The Dynamics of Collective Creativity in Human-AI Hybrid Societies

Shota Shiiku<sup>1, 5</sup>, Raja Marjieh<sup>2</sup>, Manuel Anglada-Tort<sup>3, 5</sup>, Nori Jacoby<sup>4, 5</sup>

<sup>1</sup>Graduate School of Science and Technology, Shizuoka University, Hamamatsu, Japan

<sup>2</sup>Department of Psychology, Princeton University, Princeton, USA

<sup>3</sup>Department of Psychology, Goldsmiths, University of London, London, UK

<sup>4</sup>Department of Psychology, Cornell University, Ithaca, USA

<sup>5</sup>Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany

## Human-AI Synergy Supports Collective Creative Search

Chenyi Li<sup>1</sup>, Raja Marjieh<sup>2</sup>, Haoyu Hu<sup>1</sup>, Mark Steyvers<sup>3</sup>, Katherine Collins<sup>4, 5</sup>, Ilia Sucholutsky<sup>\*6</sup> & Nori Jacoby<sup>\*1</sup>

<sup>1</sup>Cornell University

<sup>2</sup>Princeton University

<sup>3</sup>University of California, Irvine

<sup>4</sup>Massachusetts Institute of Technology

<sup>5</sup>University of Cambridge

<sup>6</sup>New York University

RESEARCH ARTICLE | ARTIFICIAL INTELLIGENCE

f X b in d e s e

## AI can help humans find common ground in democratic deliberation

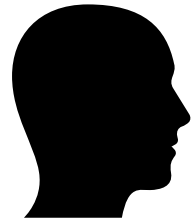
MICHAEL HENRY TESSLER , MICHEL A. BAKKER , [...], AND CHRISTOPHER SUMMERFIELD +9 authors [Authors Info &](#)

## Human-Agent Interaction in Synthetic Social Networks: A Framework for Studying Online Polarization

Tim Donkers<sup>a</sup>, Jürgen Ziegler<sup>a</sup>

<sup>a</sup>University of Duisburg-Essen, Forsthausweg 2, Duisburg, 47058, North Rhine-Westphalia, Germany

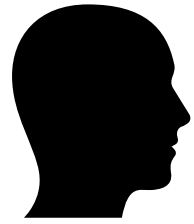
# Simulating online debate



Does red meat cause cancer and  
cardiovascular disease?

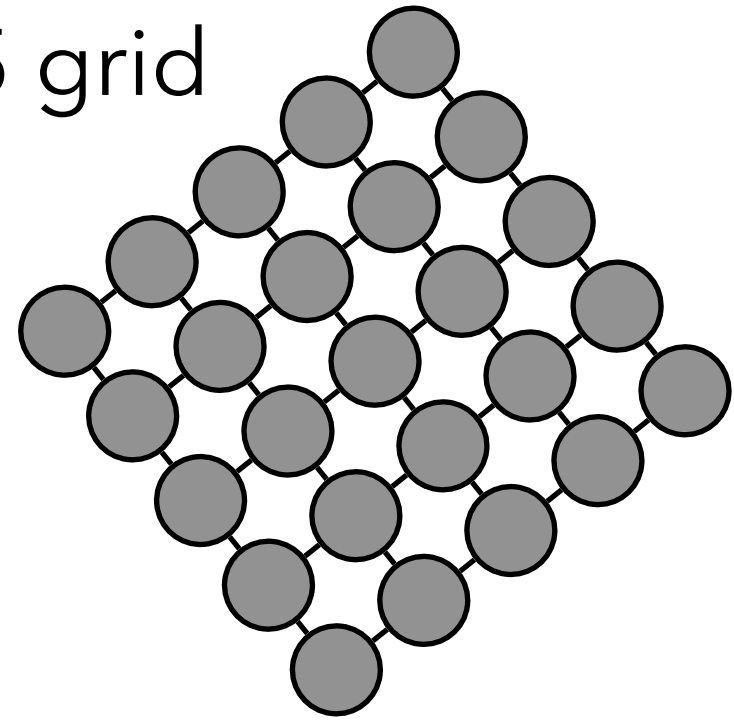


# Simulating online debate

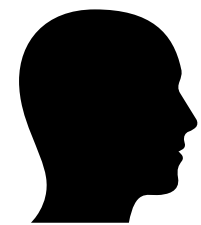


Does red meat cause cancer and  
cardiovascular disease?

5 x 5 grid

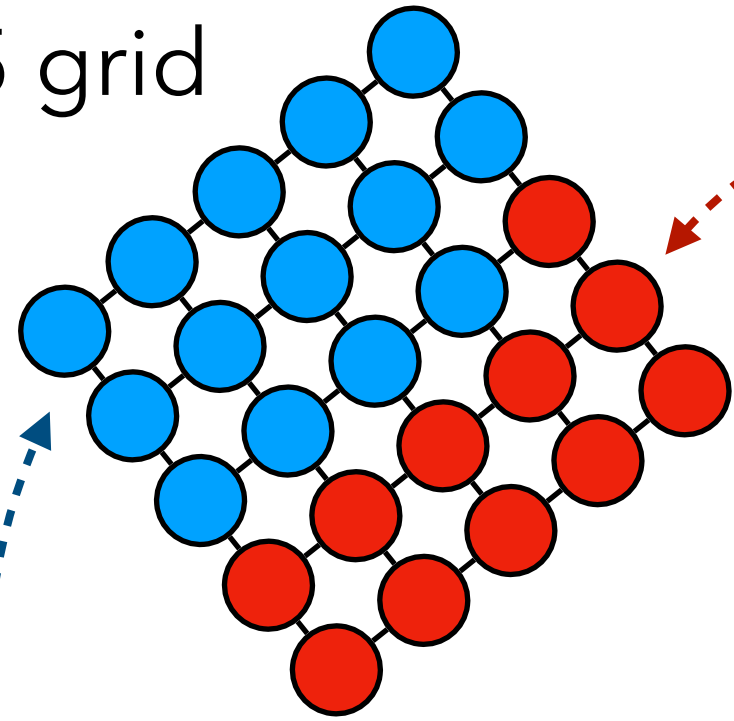


# Simulating online debate



Does red meat cause cancer and cardiovascular disease?

5 x 5 grid



Negative

Many traditional cultures with high red meat consumption have low rates of cancer and heart disease.

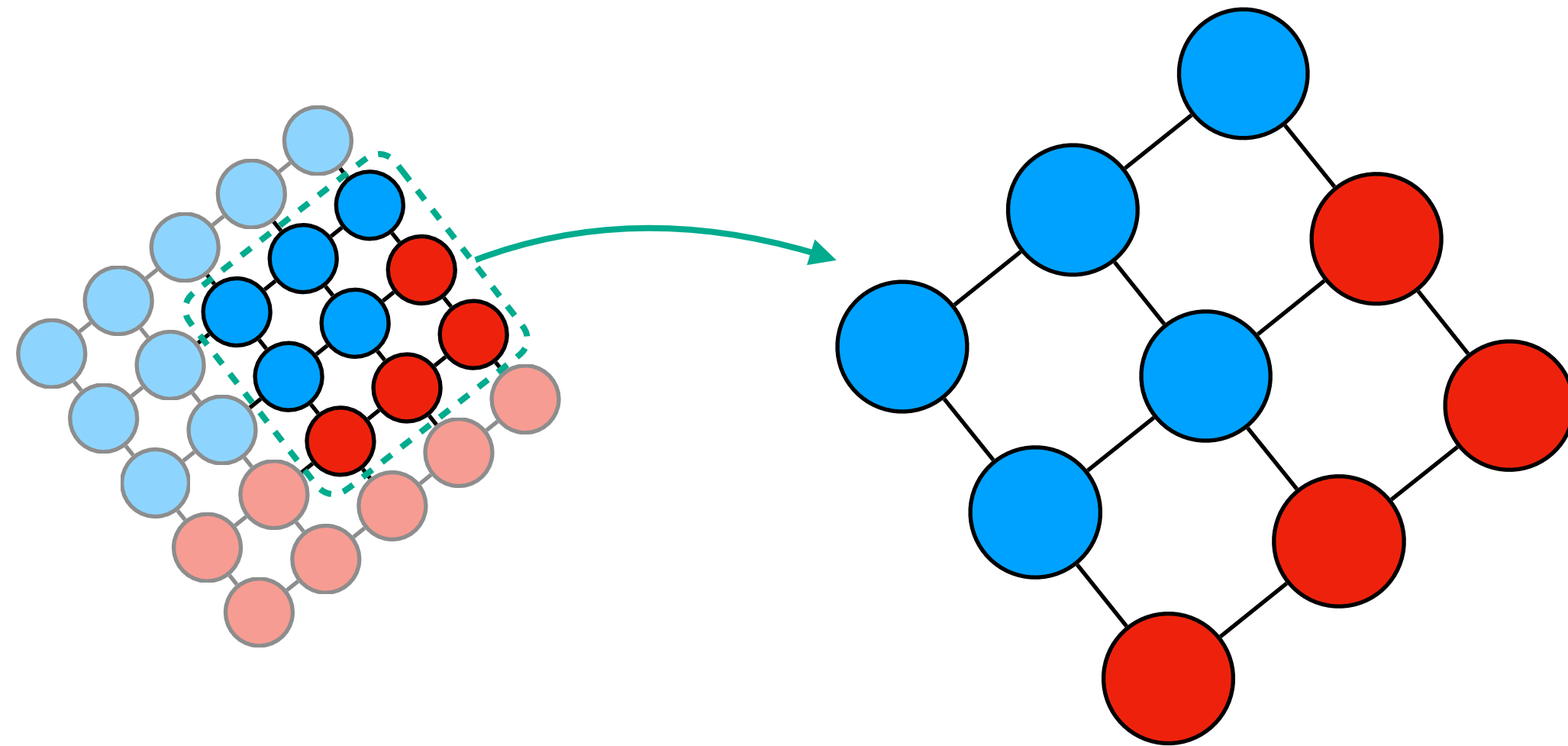
Positive

Long-term cohort studies show associations between regular meat consumption and increased cardiovascular mortality.

# Simulating online debate

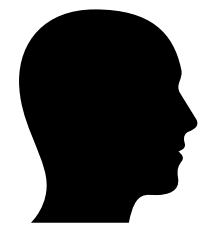


Does red meat cause cancer and cardiovascular disease?

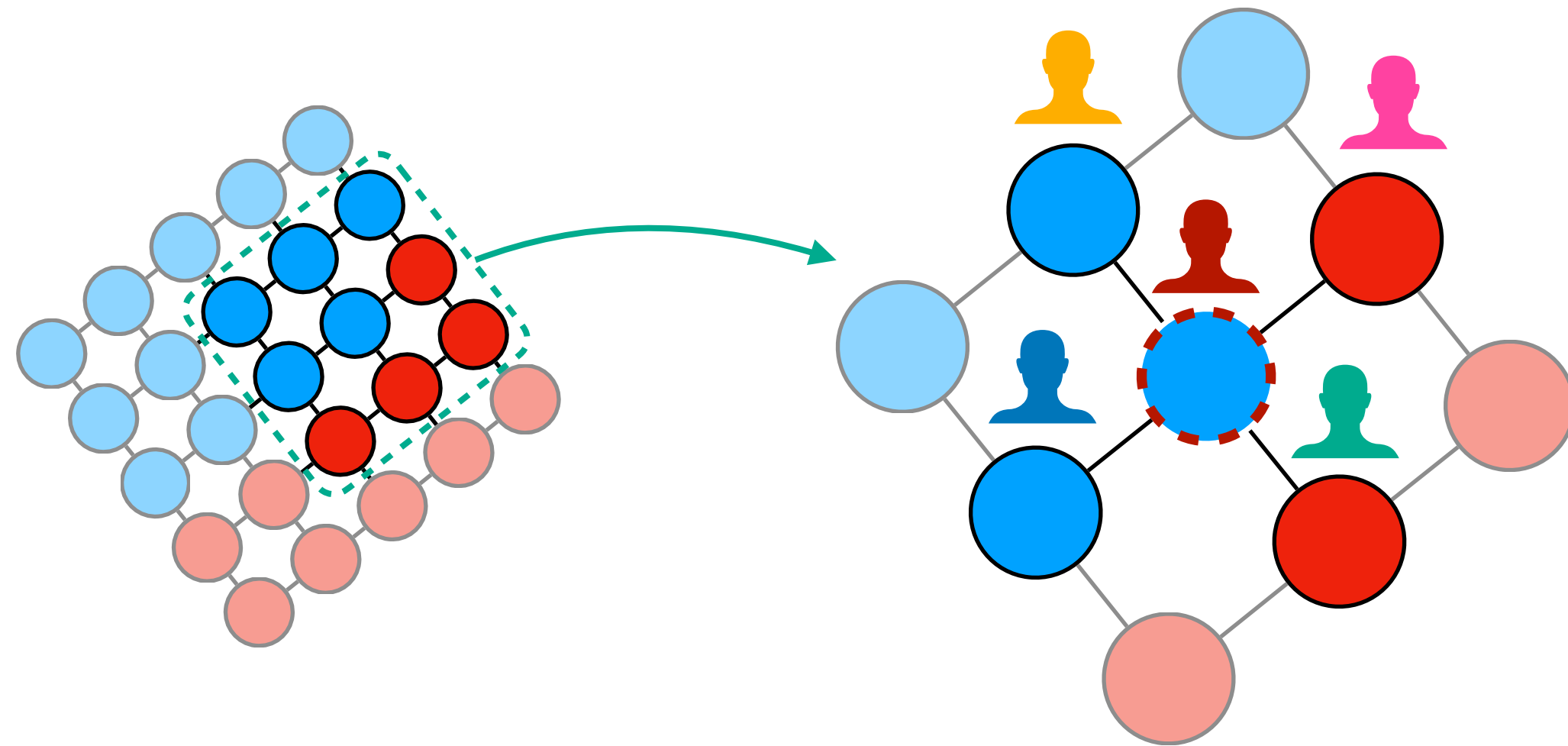




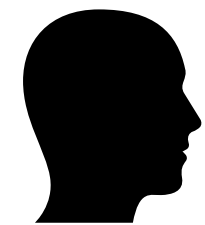
# Simulating online debate



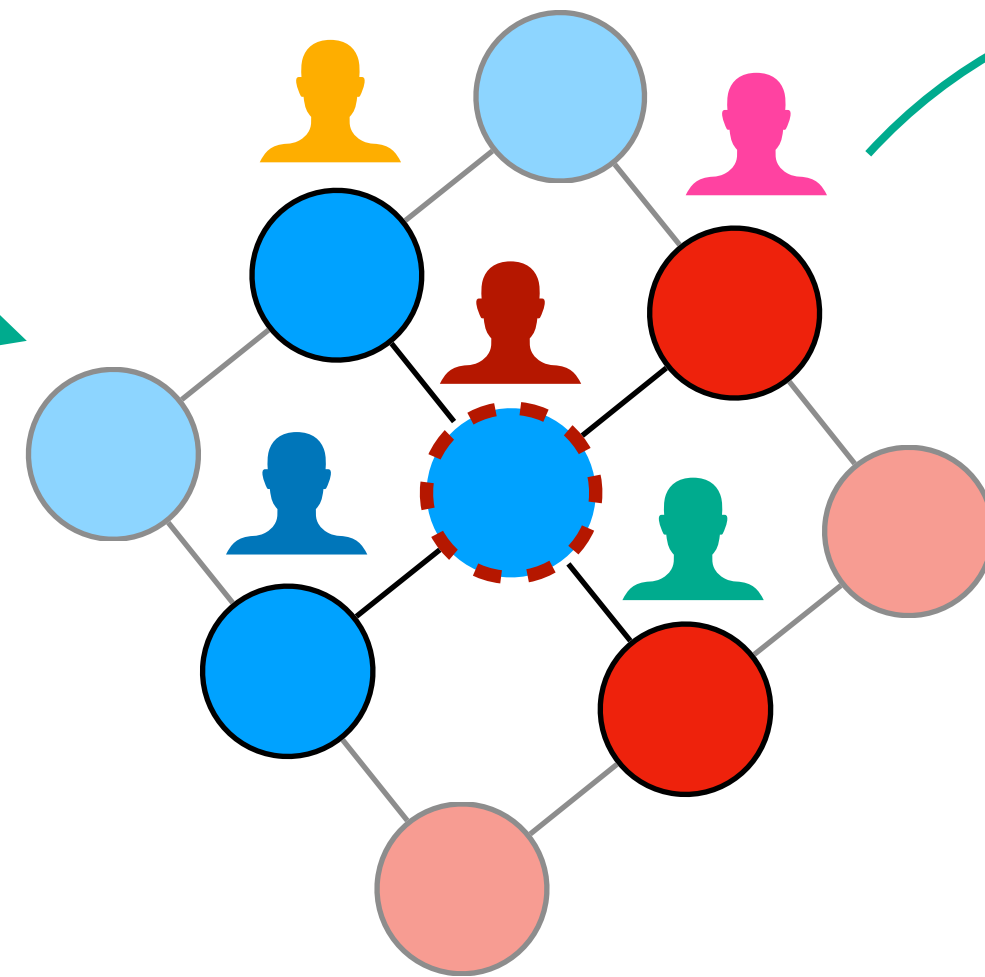
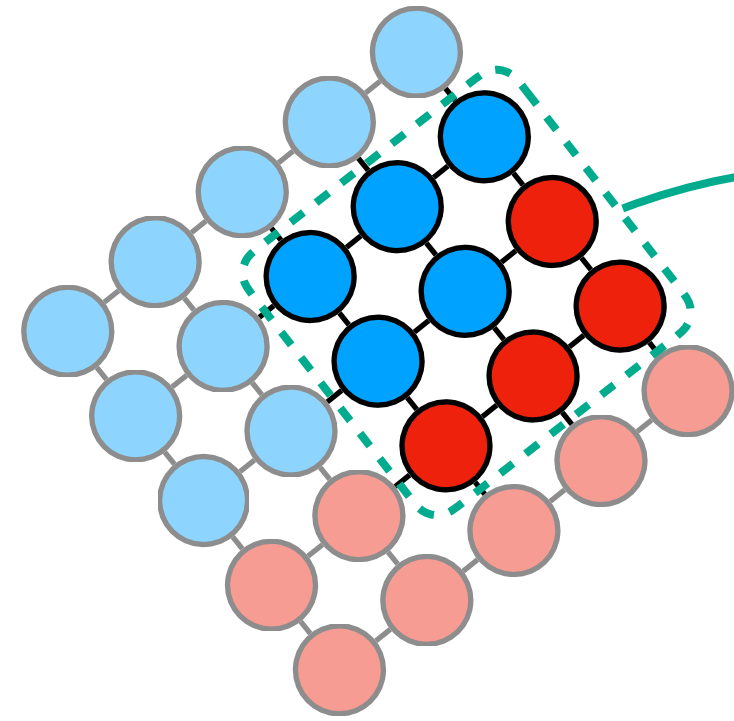
Does red meat cause cancer and cardiovascular disease?



# The task: reach consensus



Does red meat cause cancer and cardiovascular disease?



Group's answers



Answer 9



Answer 15



Answer 19

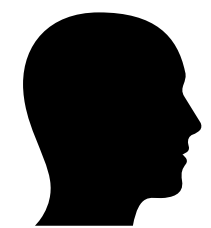


Answer 14

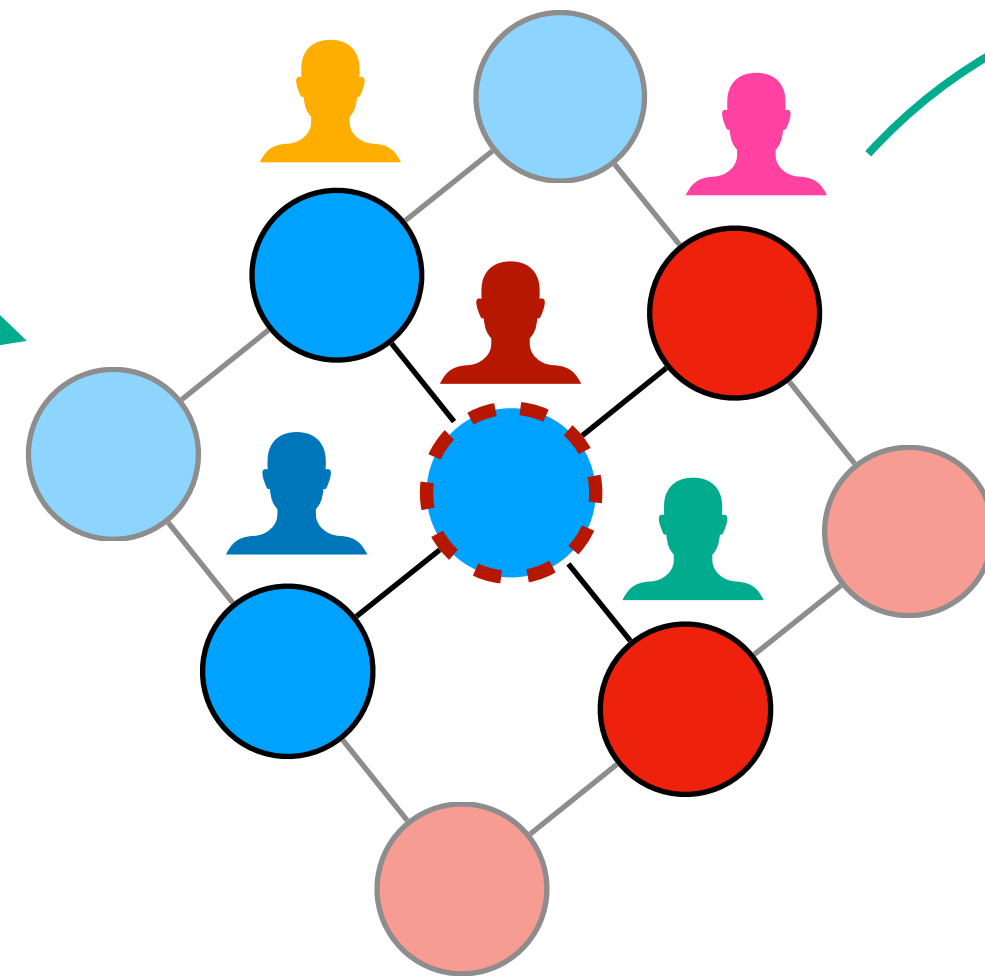
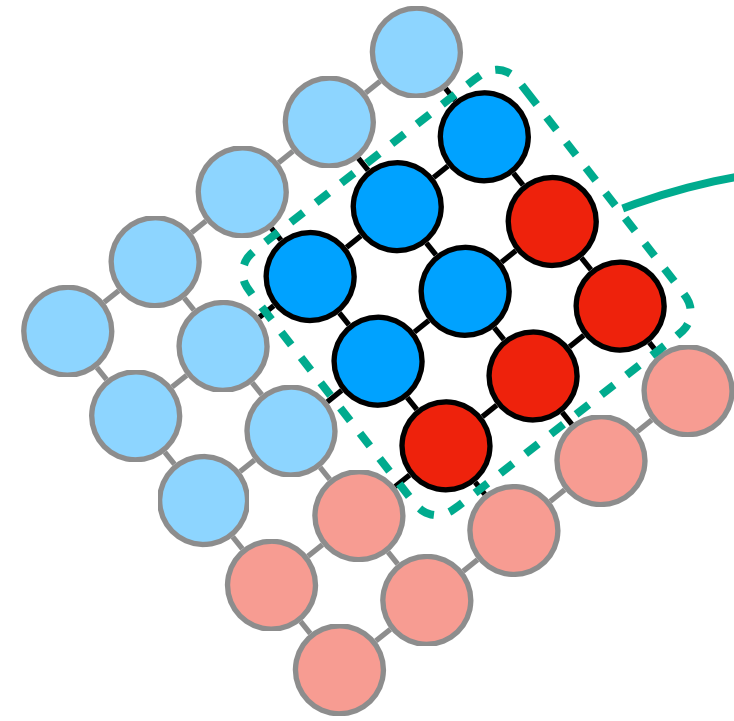


Answer 13

# The task: reach consensus



Does red meat cause cancer and cardiovascular disease?



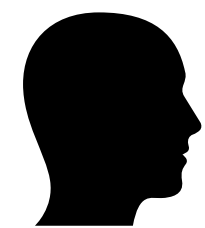
Group's answers

|                                                                                     | Group's answers |
|-------------------------------------------------------------------------------------|-----------------|
|  | Answer 9        |
|  |                 |
|  |                 |
|  |                 |
|  |                 |

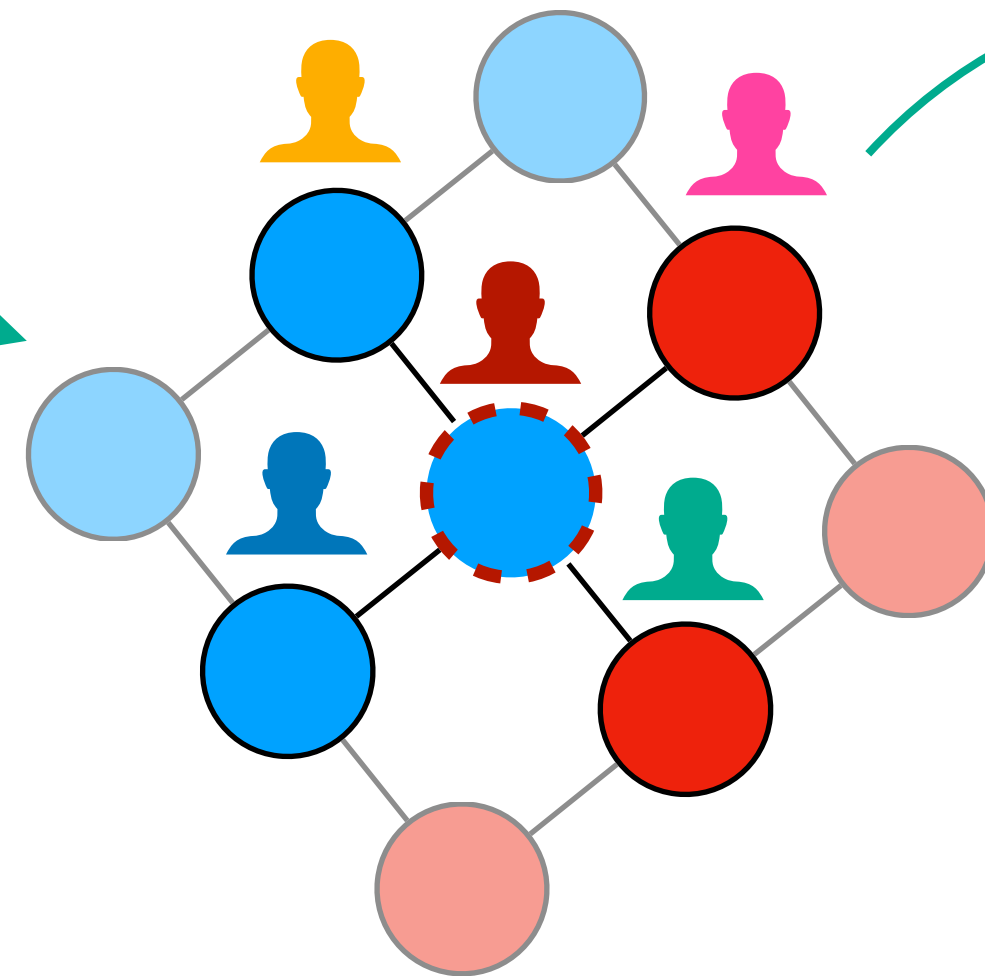
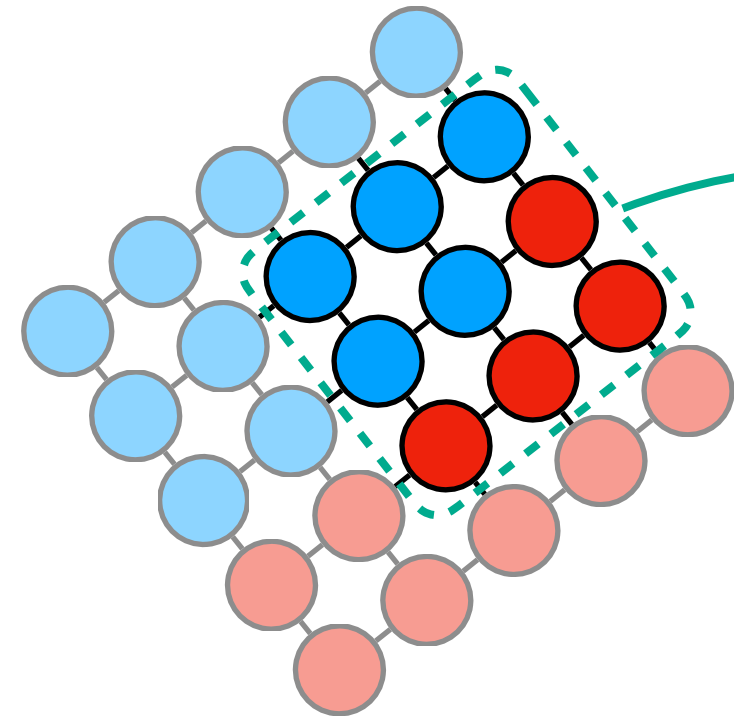
**1. Choose the answer this group would most likely agree with.**



# The task: reach consensus



Does red meat cause cancer and cardiovascular disease?



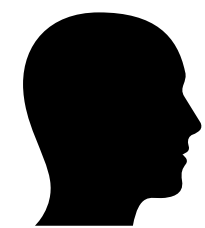
Group's answers

|  |          |
|--|----------|
|  | Answer 9 |
|  |          |
|  |          |
|  |          |
|  |          |

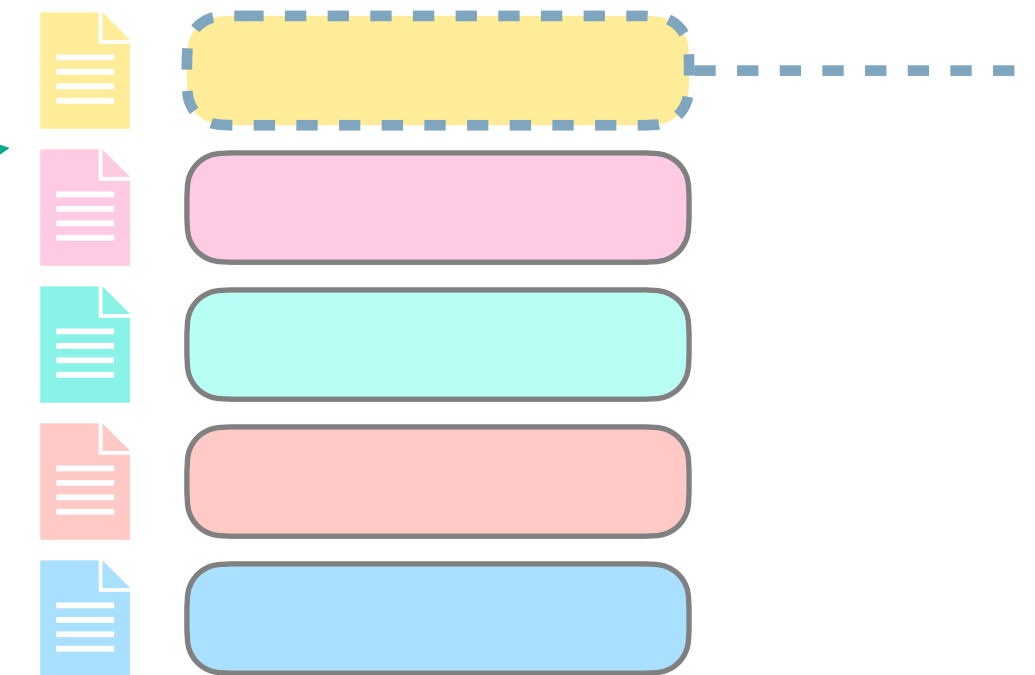
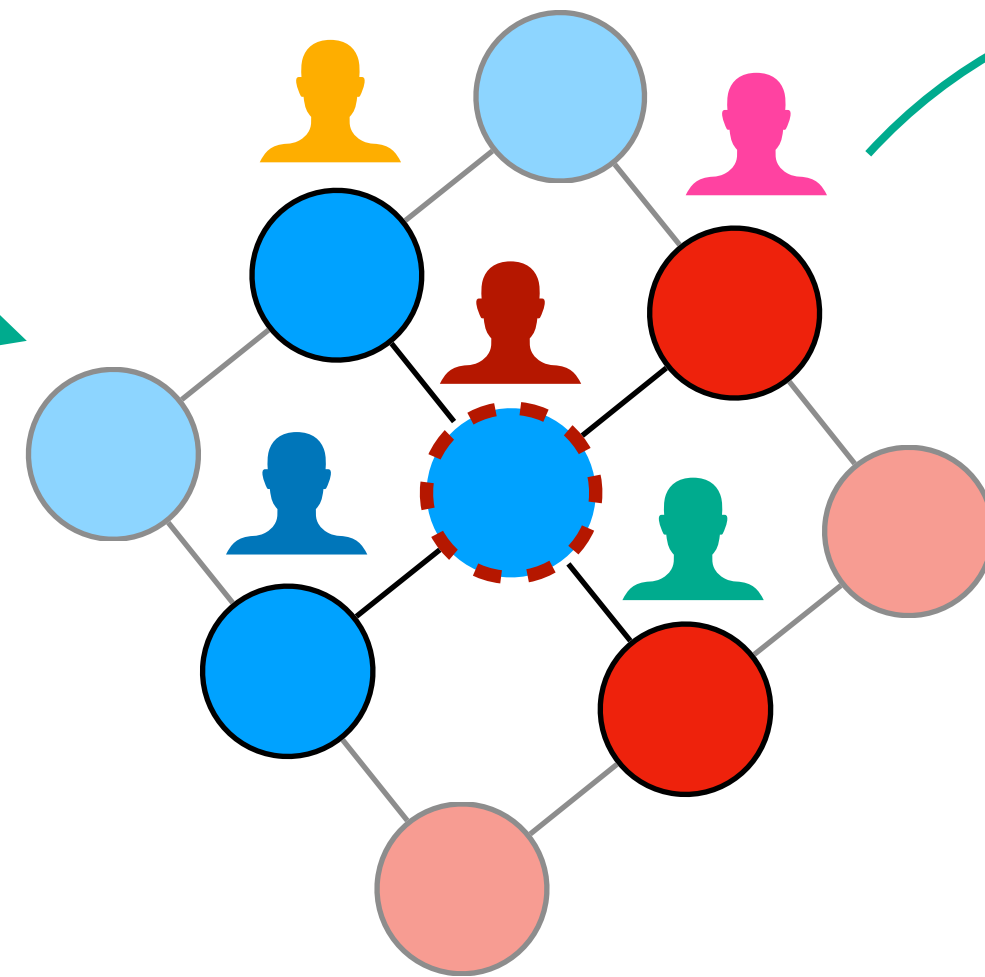
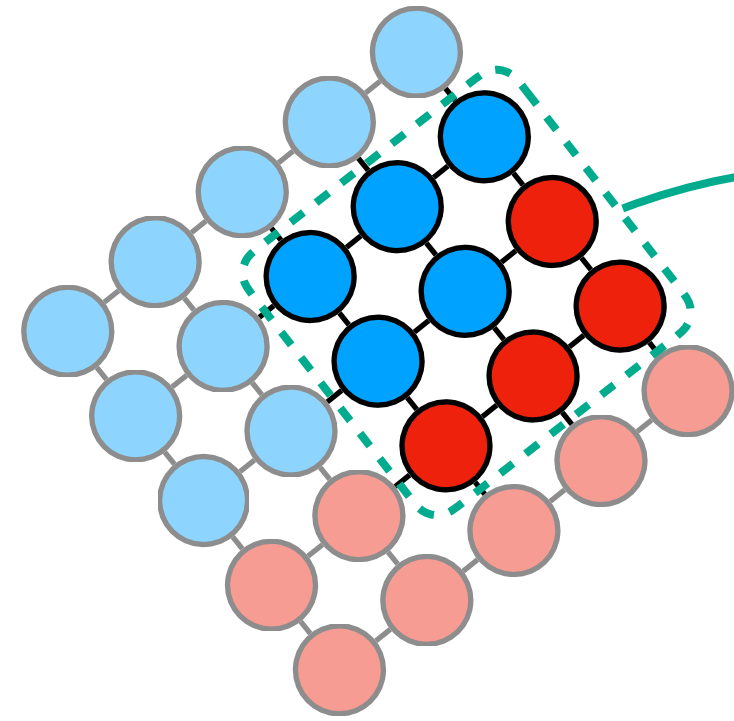
Processed red meats contain nitrates and nitrites that can form carcinogenic compounds in the body.

**2. Try and remember your answer.**

# The task: reach consensus



Does red meat cause cancer and cardiovascular disease?



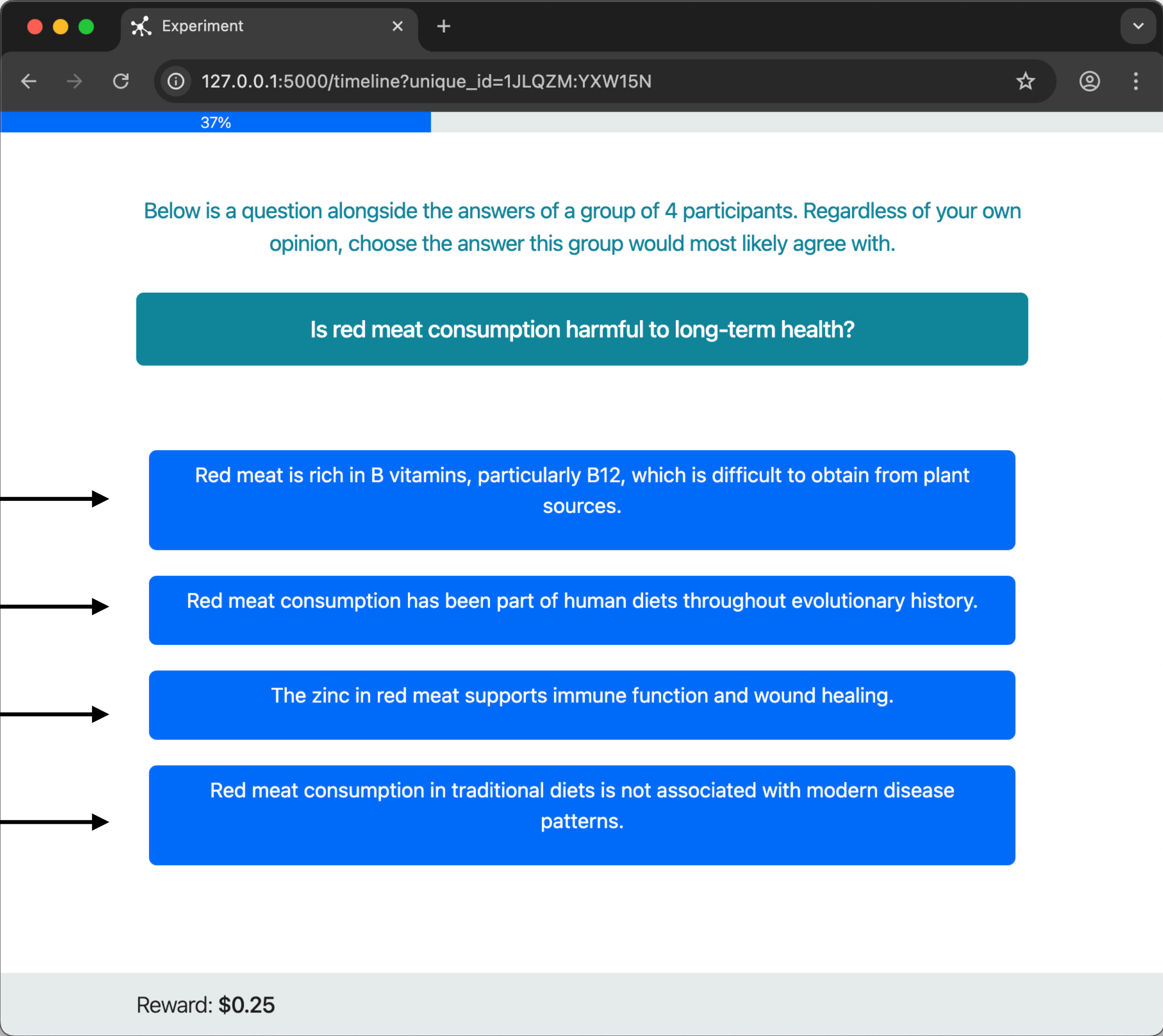
**Answer 9 (old):** Processed red meats contain nitrates and nitrites that can form carcinogenic compounds in the body.

*revise to reach consensus*

**Answer 30 (new):** Nitrates are stored in pre packaged meat that can be carcinogenic and lead to potential cancers

**3. Revise your answer.**

# The task: reach consensus



Experiment

127.0.0.1:5000/timeline?unique\_id=1JLQZM:YXW15N

37%

Below is a question alongside the answers of a group of 4 participants. Regardless of your own opinion, choose the answer this group would most likely agree with.

Is red meat consumption harmful to long-term health?

Red meat is rich in B vitamins, particularly B12, which is difficult to obtain from plant sources.

Red meat consumption has been part of human diets throughout evolutionary history.

The zinc in red meat supports immune function and wound healing.

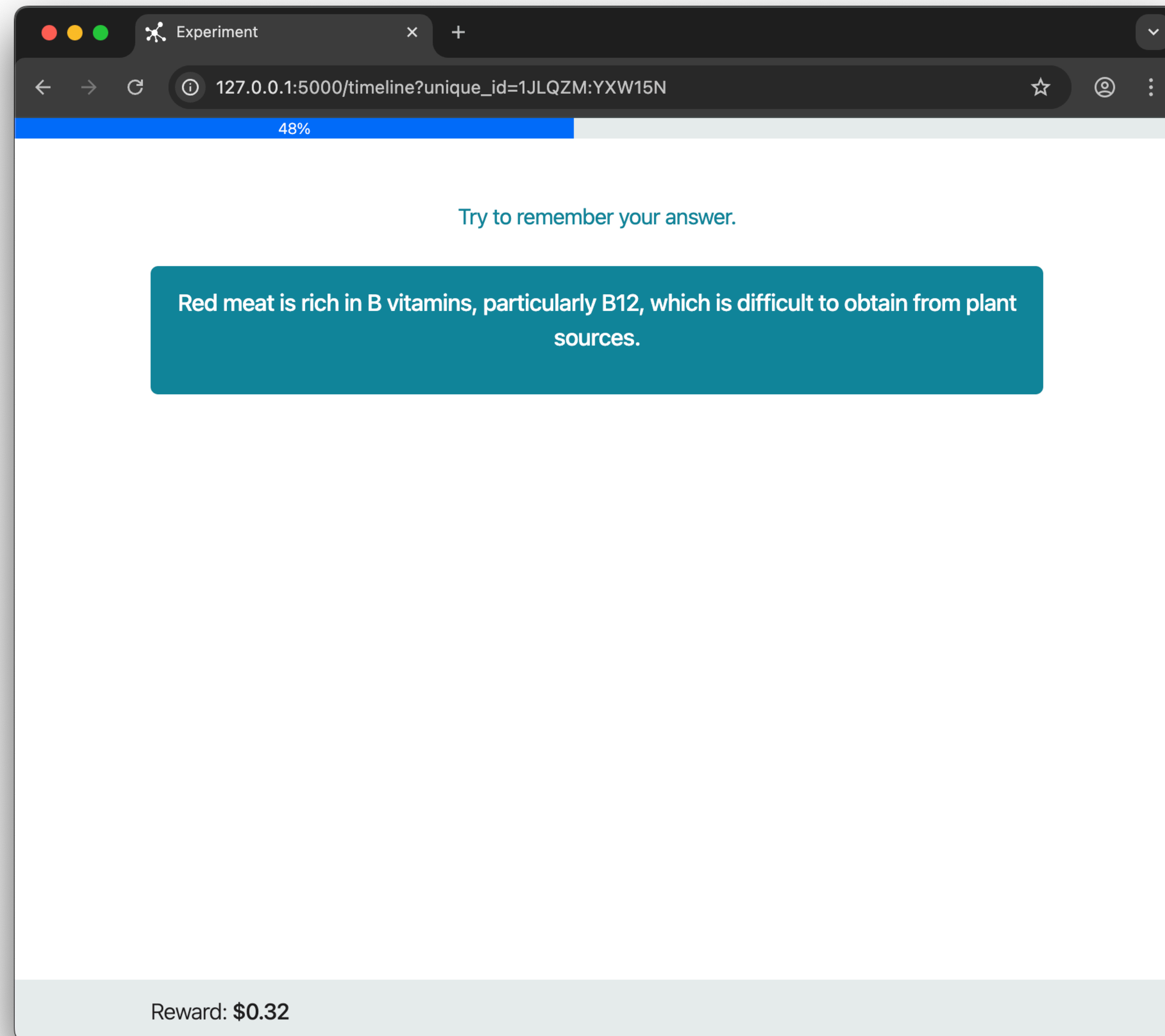
Red meat consumption in traditional diets is not associated with modern disease patterns.

Reward: \$0.25

**Choose the answer  
this group would  
most likely agree  
with.**



# The task: reach consensus



**Try and remember your choice.**

# The task: reach consensus

Experiment

127.0.0.1:5000/timeline?unique\_id=1JLQZM:YXW15N

70%

Revise your answer so that it most accurately reflects the views that the previously observed group of participants would be likely to agree with.

Is red meat consumption harmful to long-term health?

Progress bar with four segments: green, light green, purple, and red.

Next

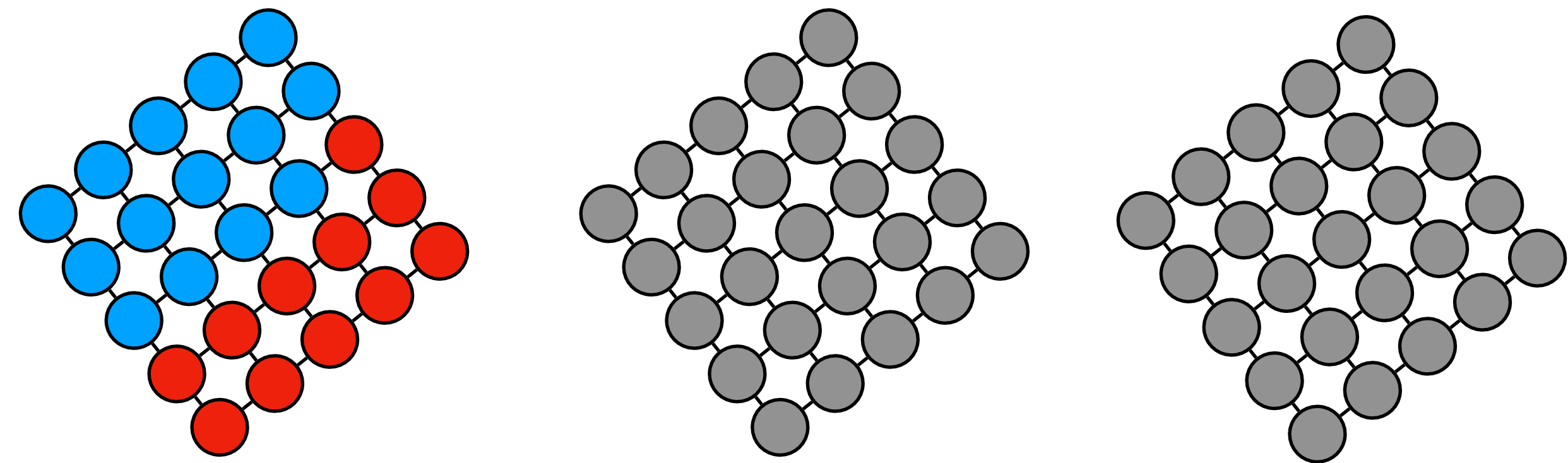
Reward: \$0.47

**Revise your answer.**



# Debate evolution over iteration

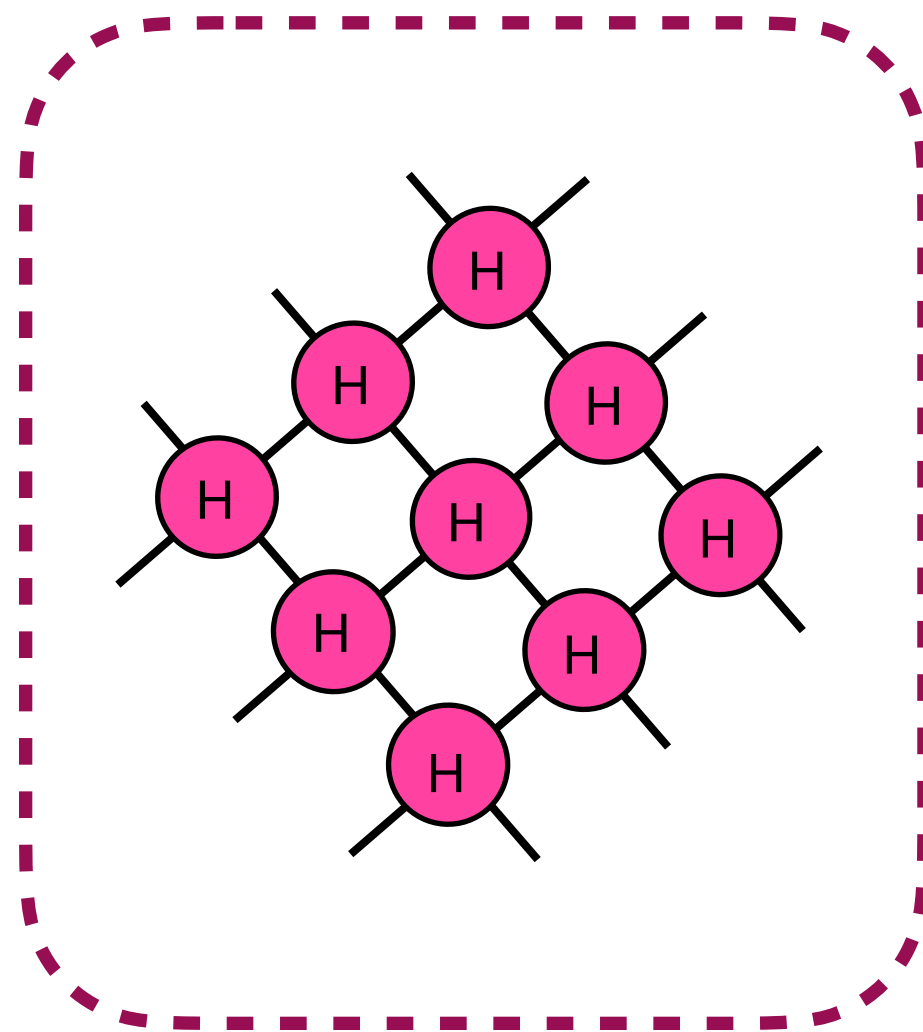
5 x 5 grid



8 iterations

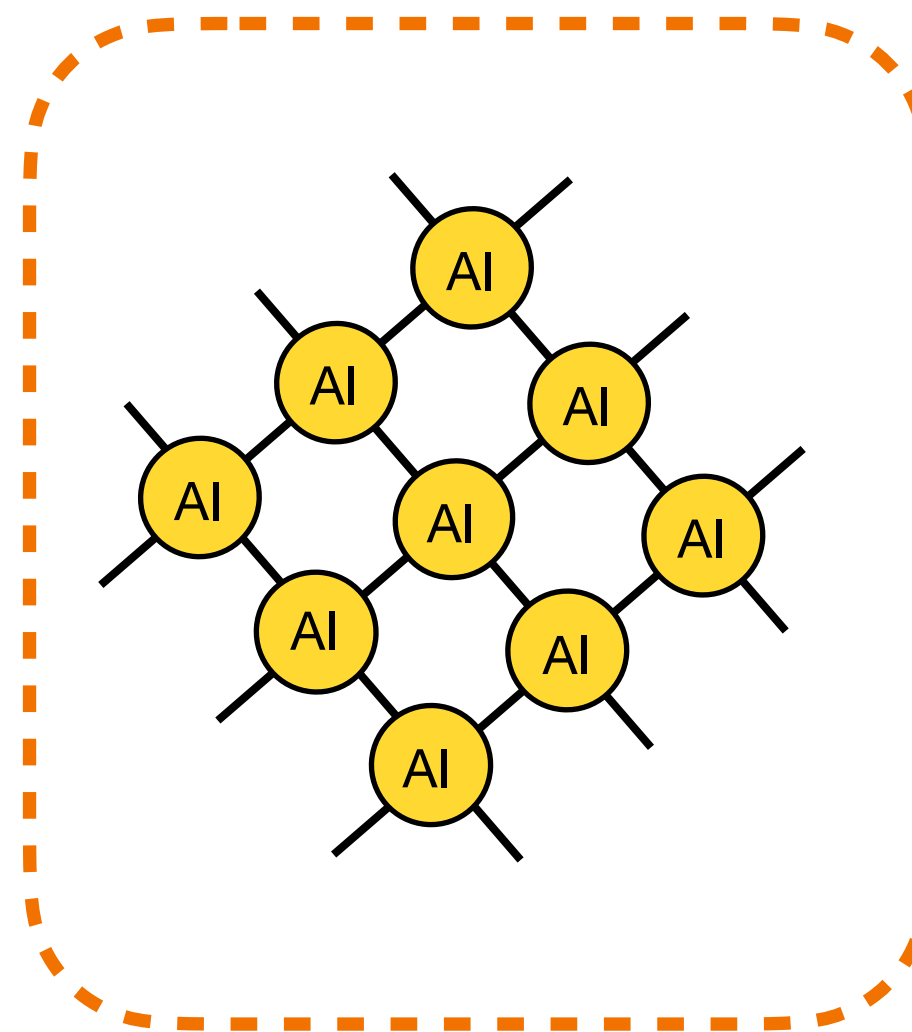
# Participants

Human-only



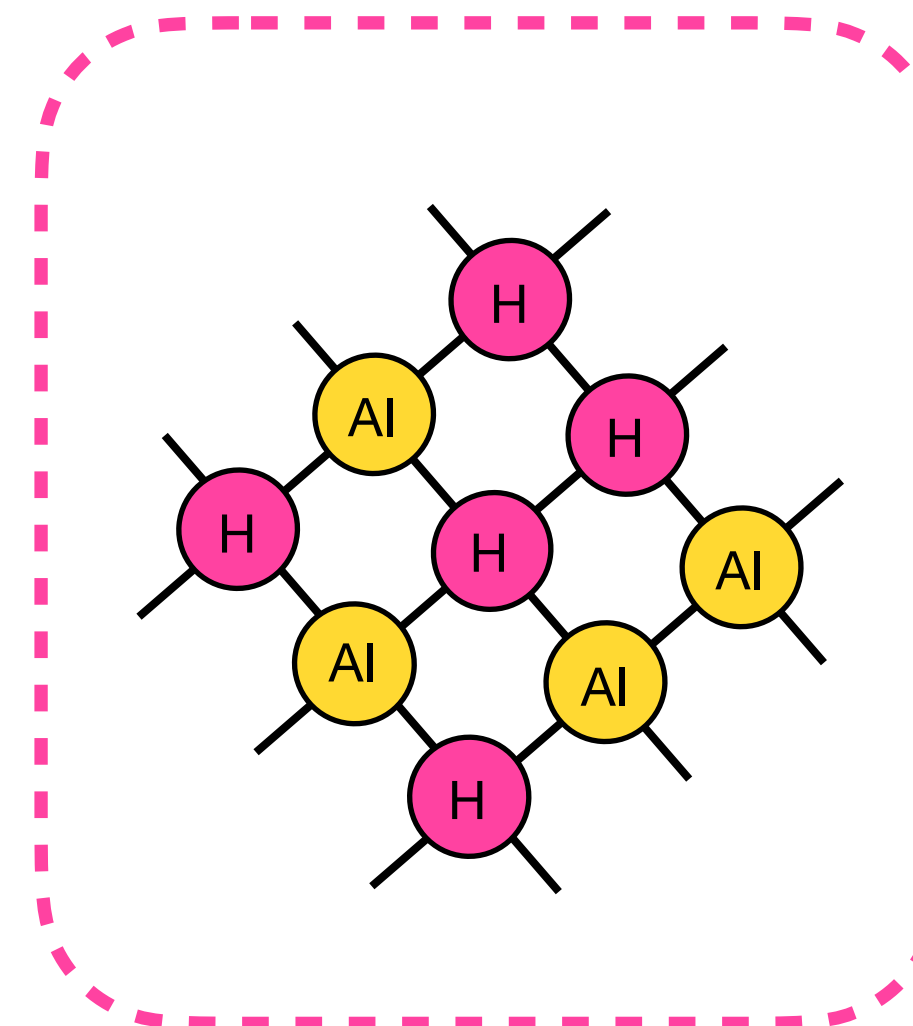
200 humans

Grok



200 Grok

Hybrid



100 humans & 100 Grok

Humans



Grok

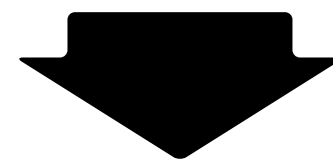




# Statements

**Human-only networks:** vague language, personal opinions, eventually more refined language in later iterations

***I think** many studies fail to tell the difference between processed or not processed red meat which have different health effect*

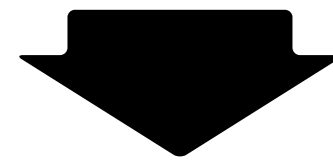


***International bodies** say that eating red meat can contribute towards cancer, but that other factors also contribute. The key is to **pursue healthy eating** and to eat in **moderation***

# Statements

**AI-only networks:** more refined from the beginning, citing health organization, precise medical lexicon. Towards later iterations, statements become increasingly similar/identical.

*Processed red meats contain **nitrates** and **nitrites** that can form **carcinogenic compounds** in the body*



***The World Health Organization** has classified processed red meat as a **Group 1 carcinogen**.*

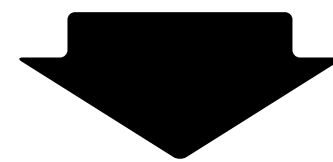
***The American Heart Association** recommends limiting red meat intake to reduce cardiovascular disease risk.*

# Statements

When **both** exchange in the network, resulting answers combine elements of both approaches.

***I think** that too much red meat causes heart issues which could probably led to cancer*

***The American Heart Association** recommends limiting red meat intake to reduce cardiovascular disease risks.*



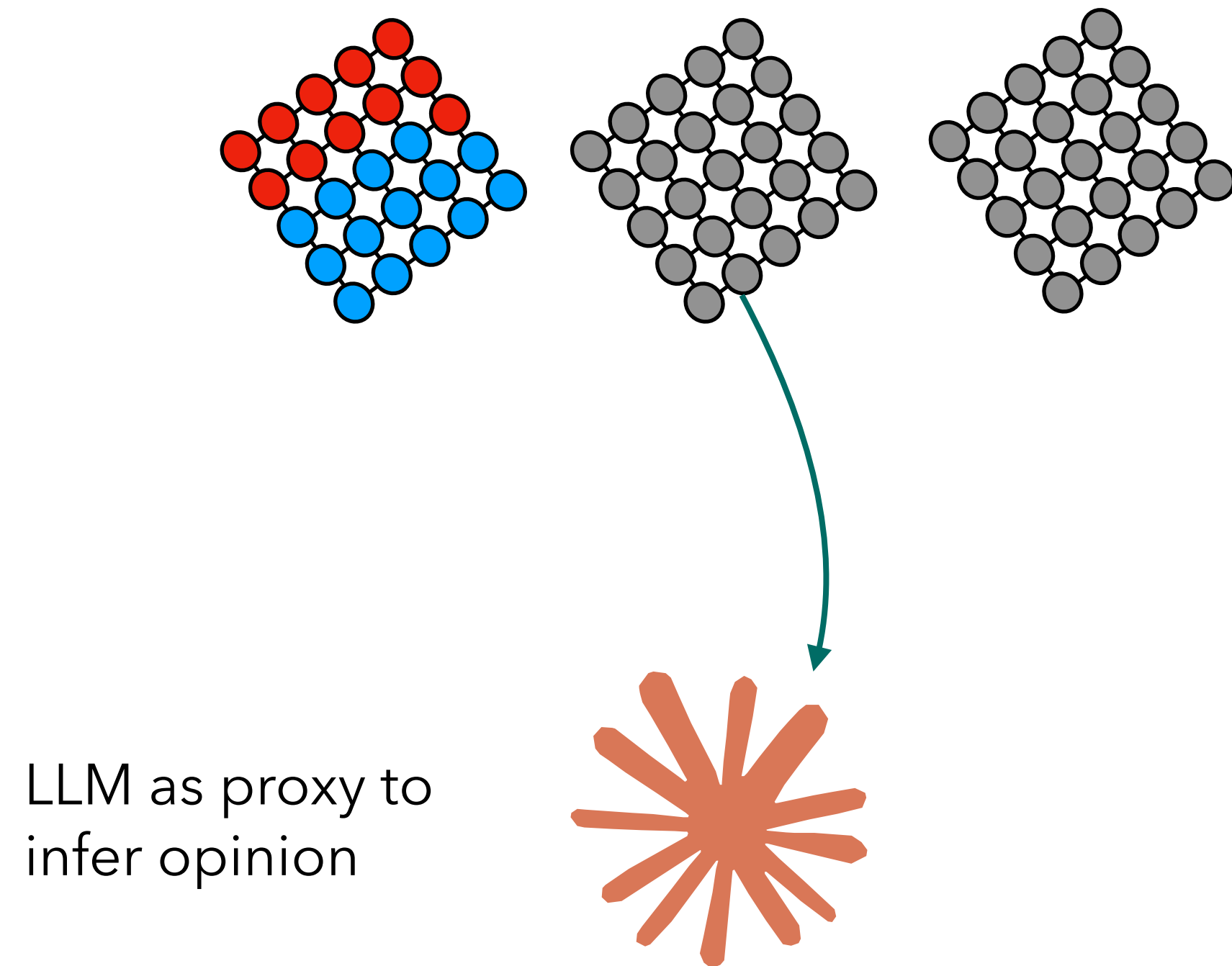
***The World Health Organization** has classified processed red meat as a **Group 1 carcinogen**.*

*Eating lots of red meat can cause you to gain weight, this in turn can put **pressure on the heart** and less exercise can lead to cancer.*



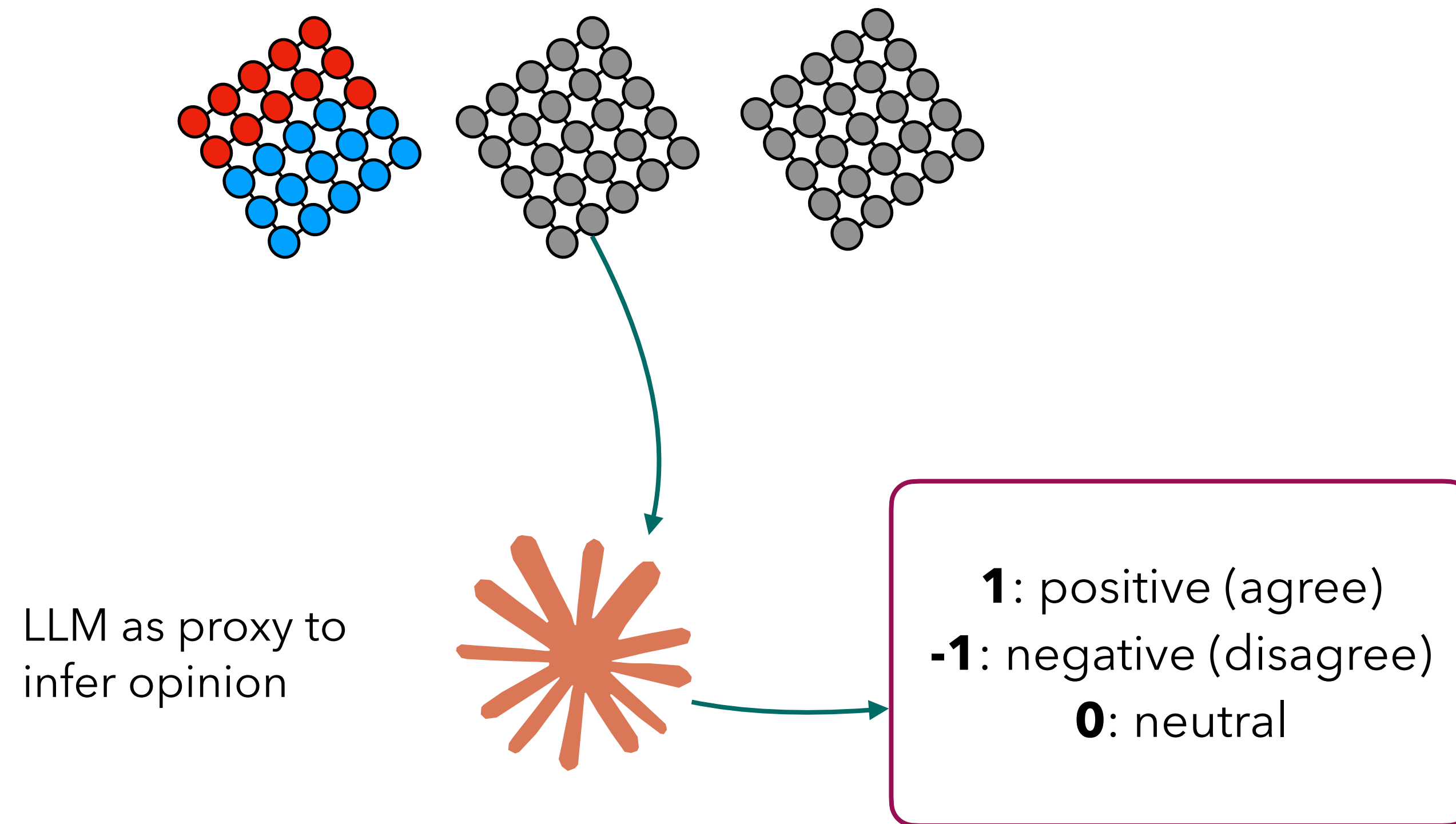
# Opinion mining using LLM-as-a-judge

8 iterations



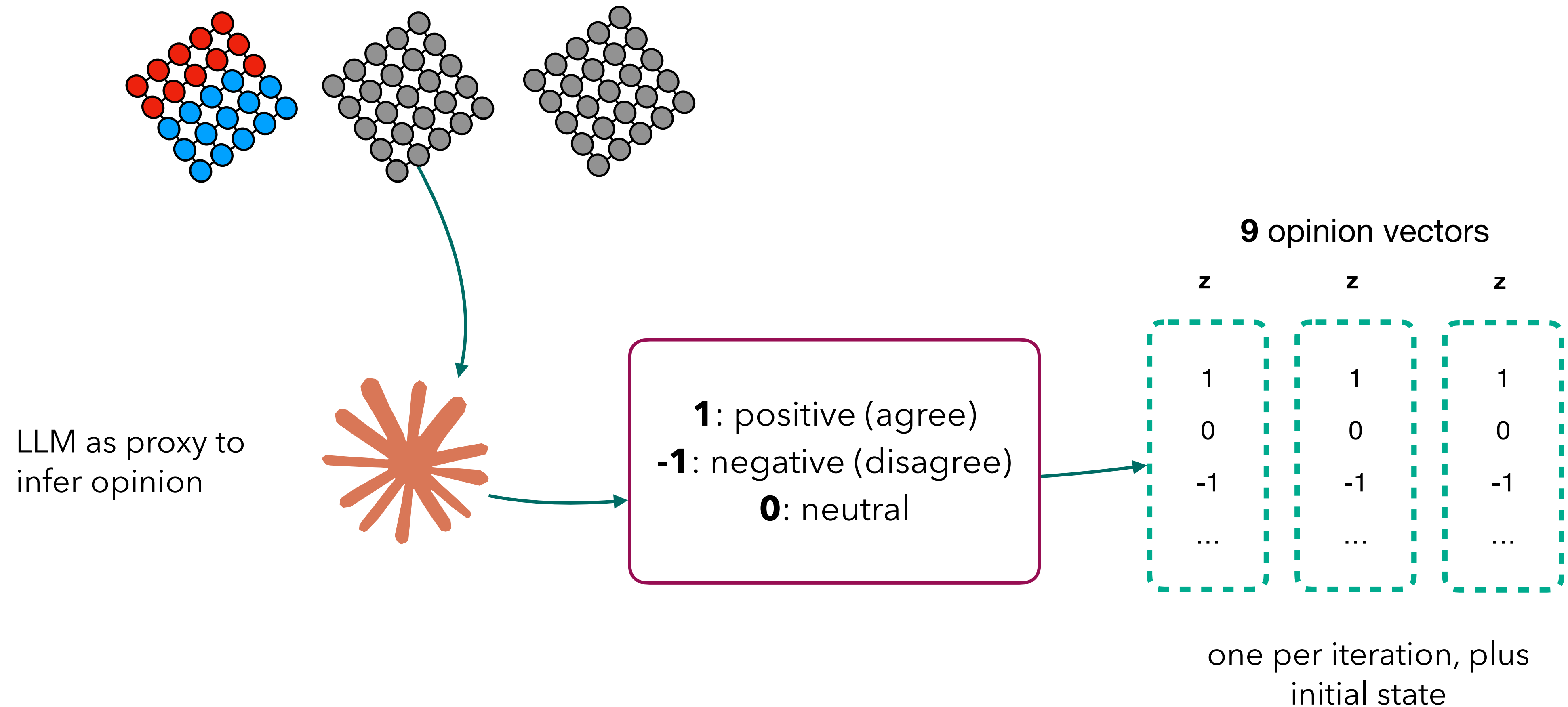
# Opinion mining using LLM-as-a-judge

8 iterations



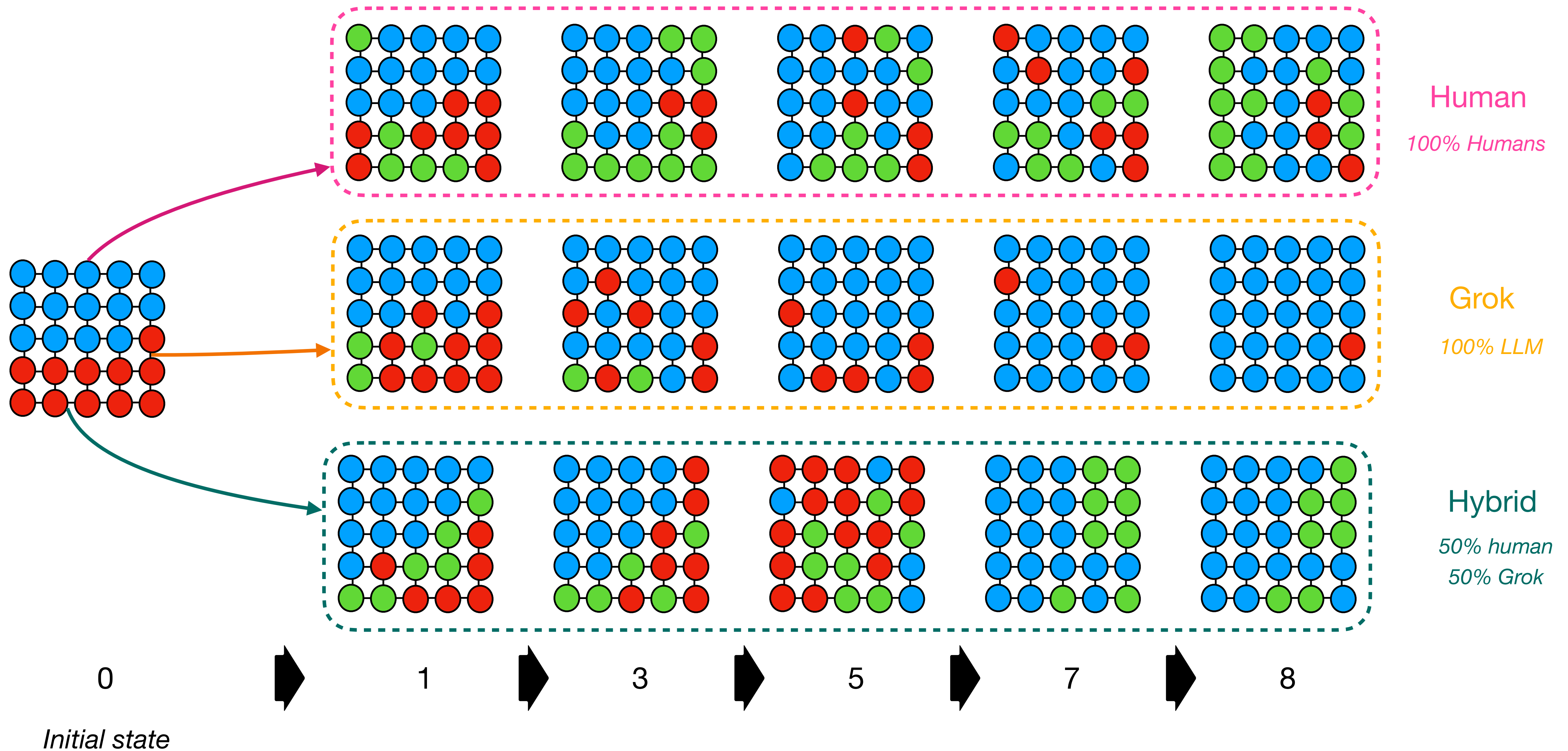
# Opinion mining using LLM-as-a-judge

8 iterations





# Opinion dynamics





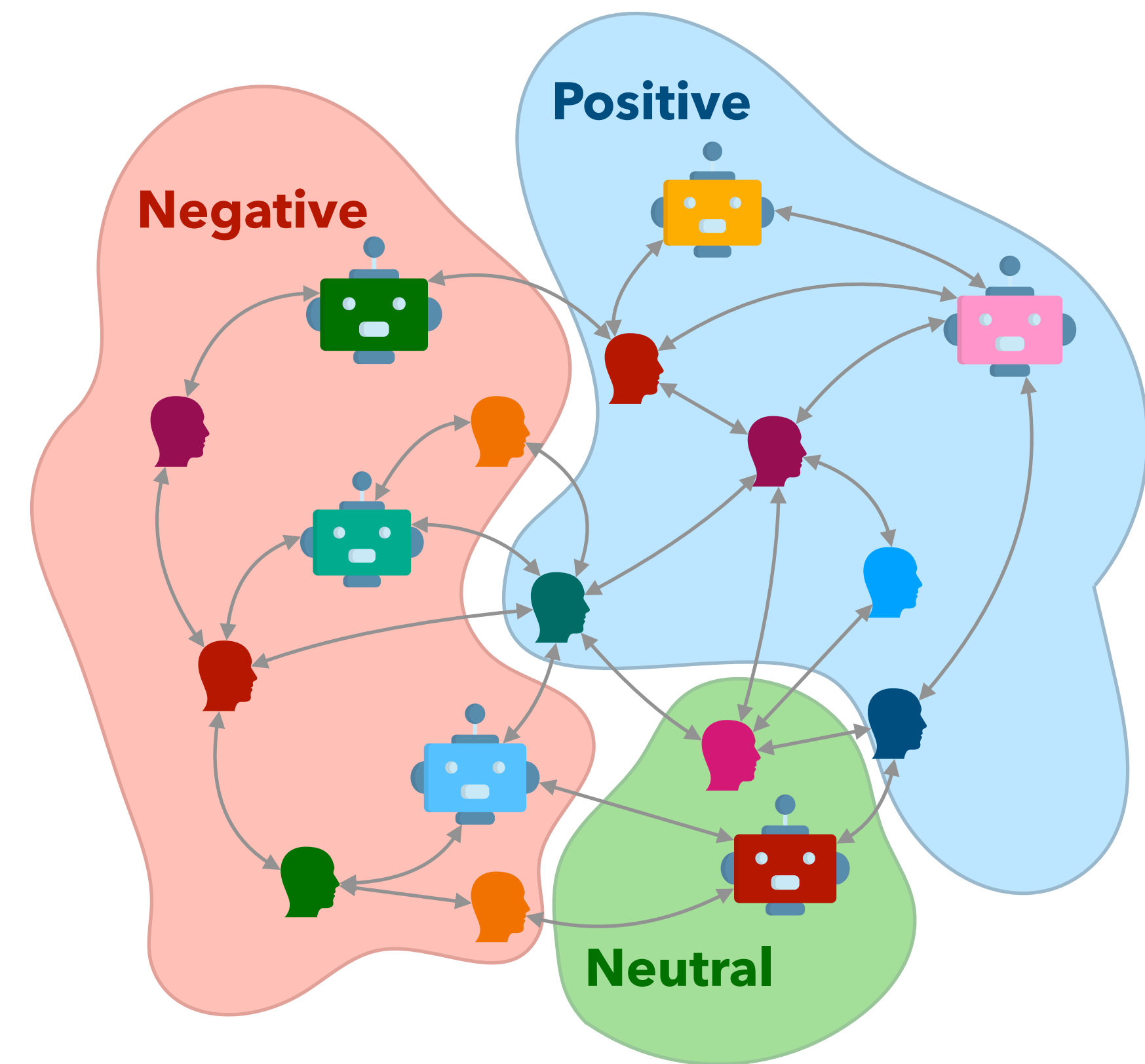
# Quantifying Polarization

Polarization Index  
(Matakos et al., 2017)

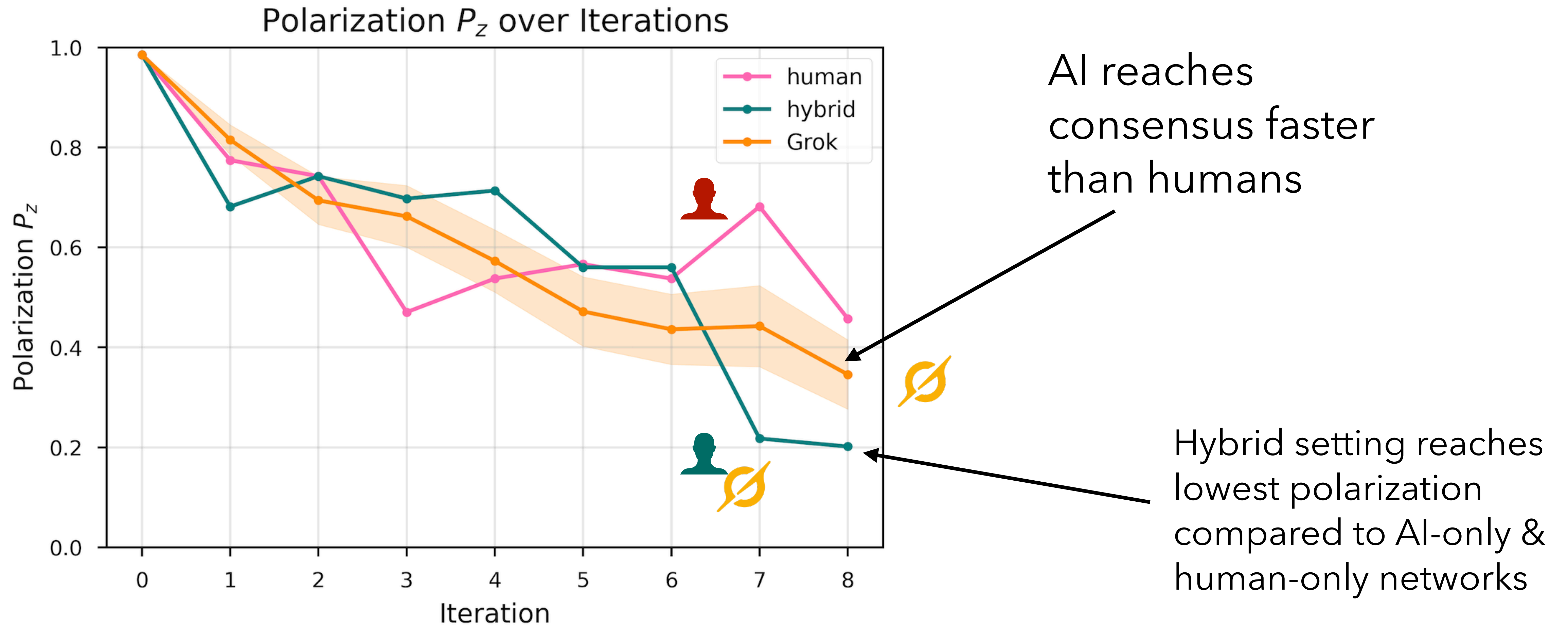
Variance of *expressed* opinions in  
the network at a given iteration.

$$P_z = \frac{1}{N} \sum_{i=1}^N (z_i - \bar{z})^2$$

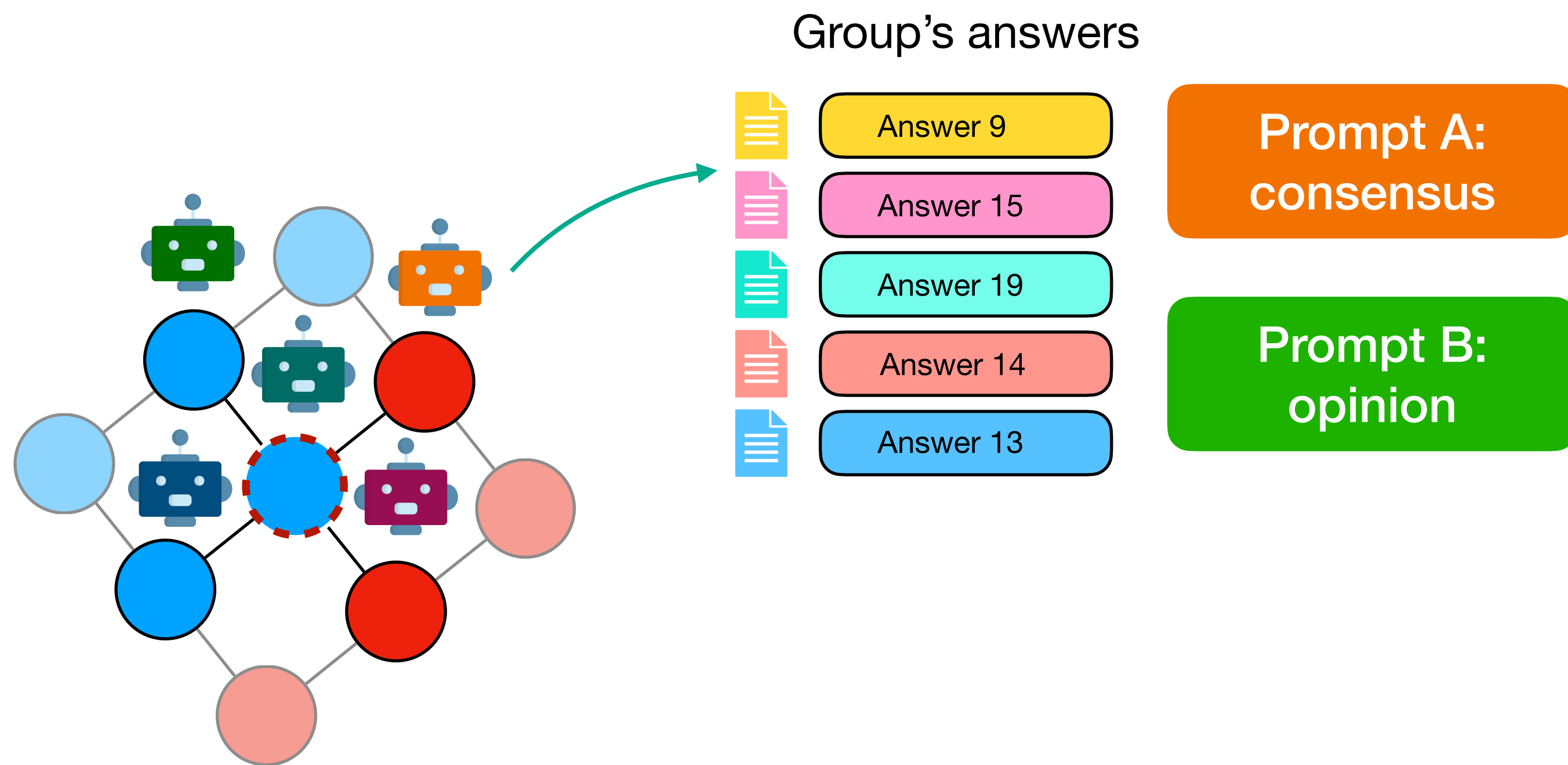
**z**: opinion vector



# Quantifying Polarization

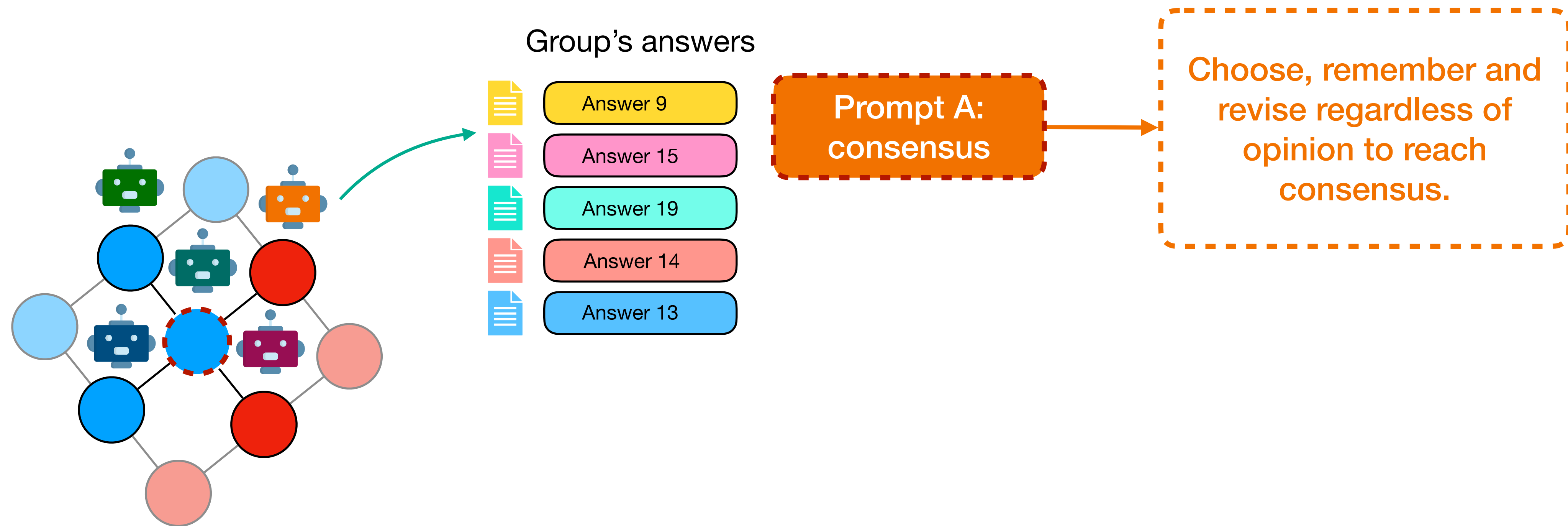


# Prompt design sensibility

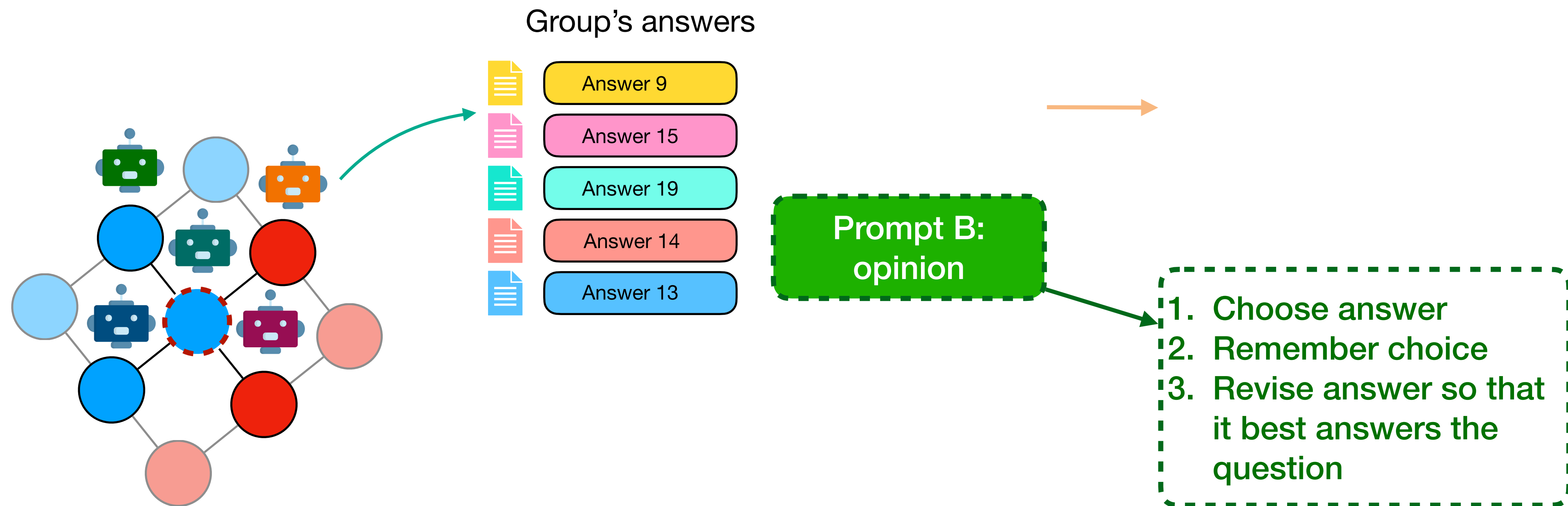




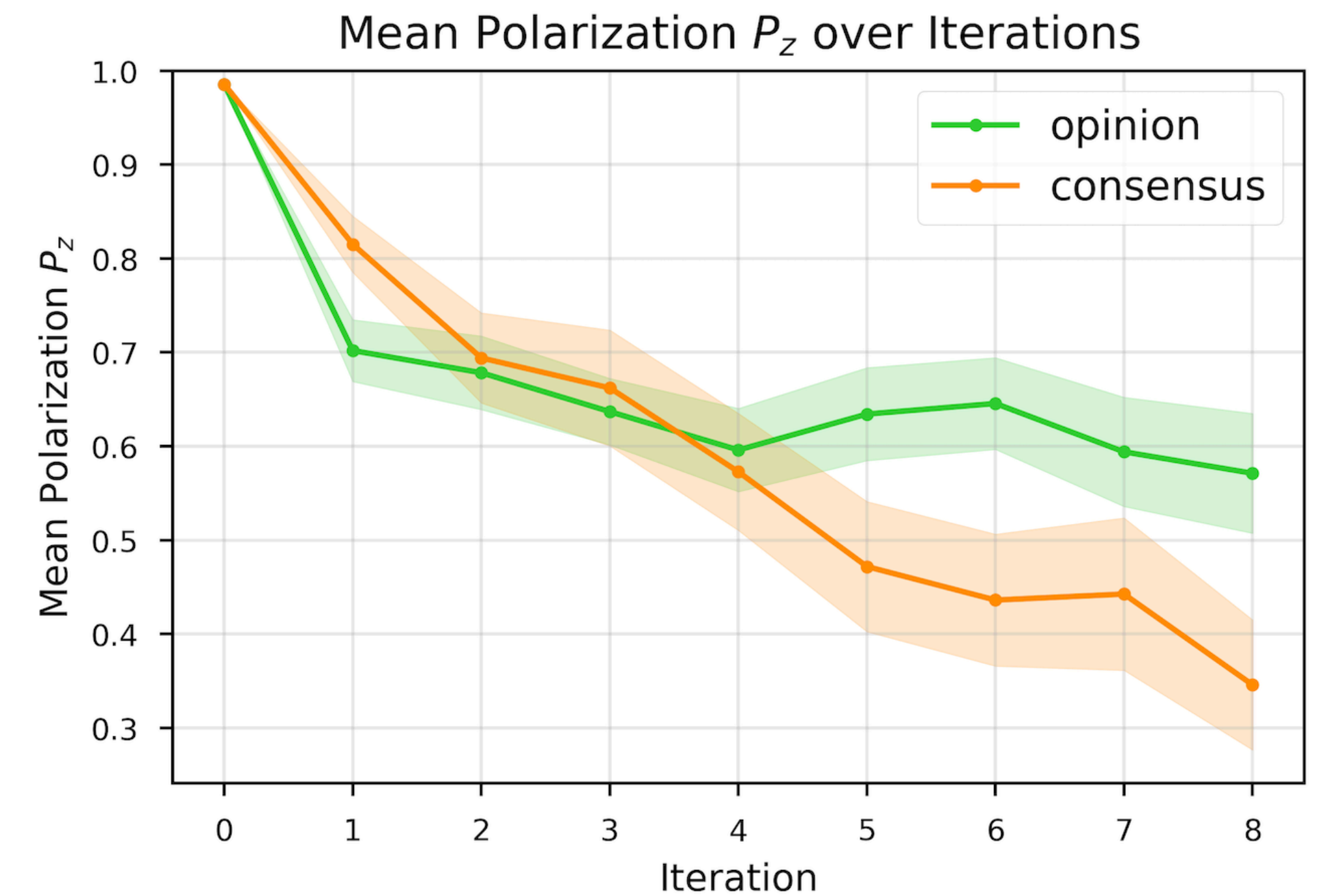
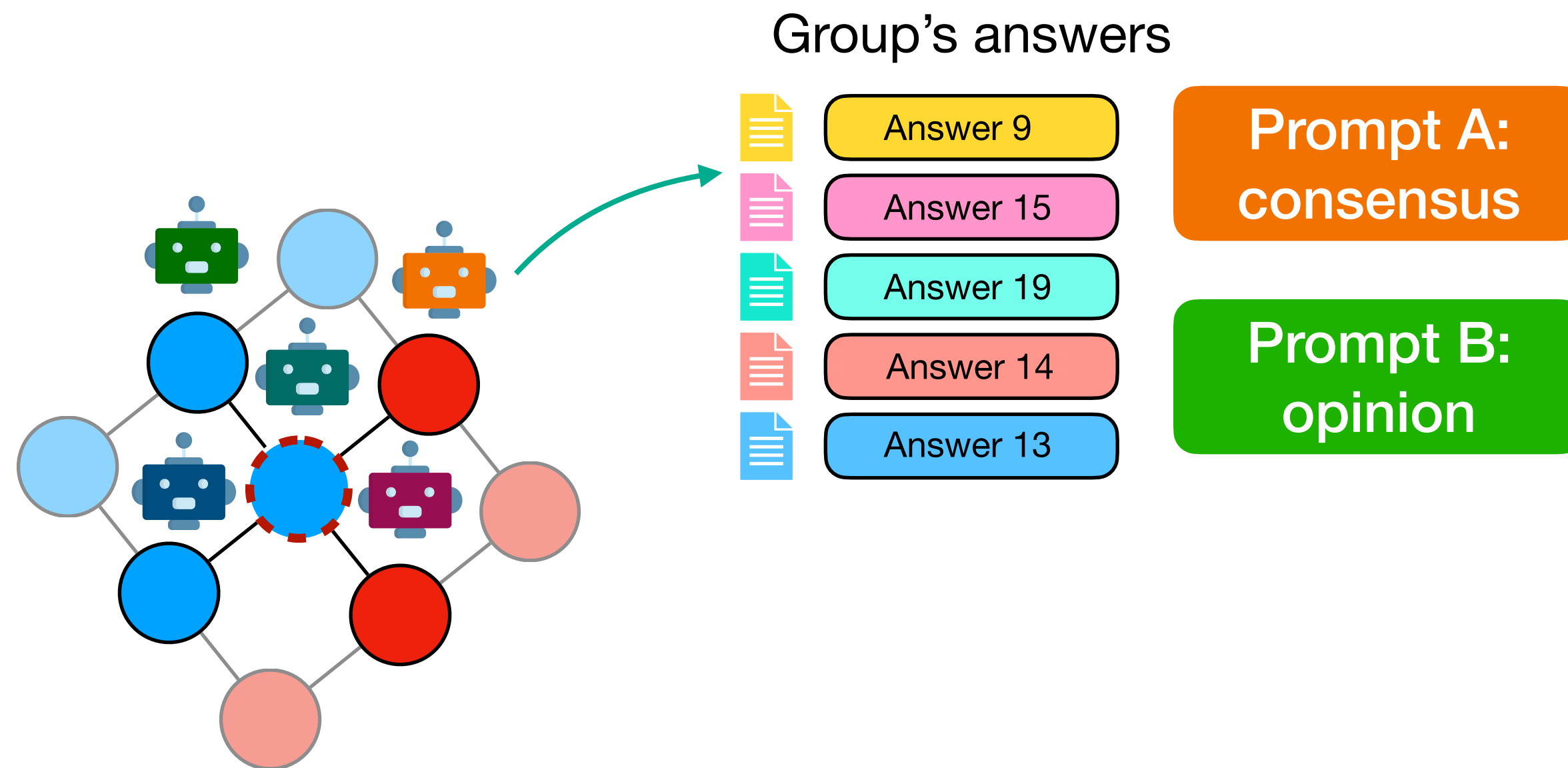
# Prompt design sensibility



# Prompt design sensibility

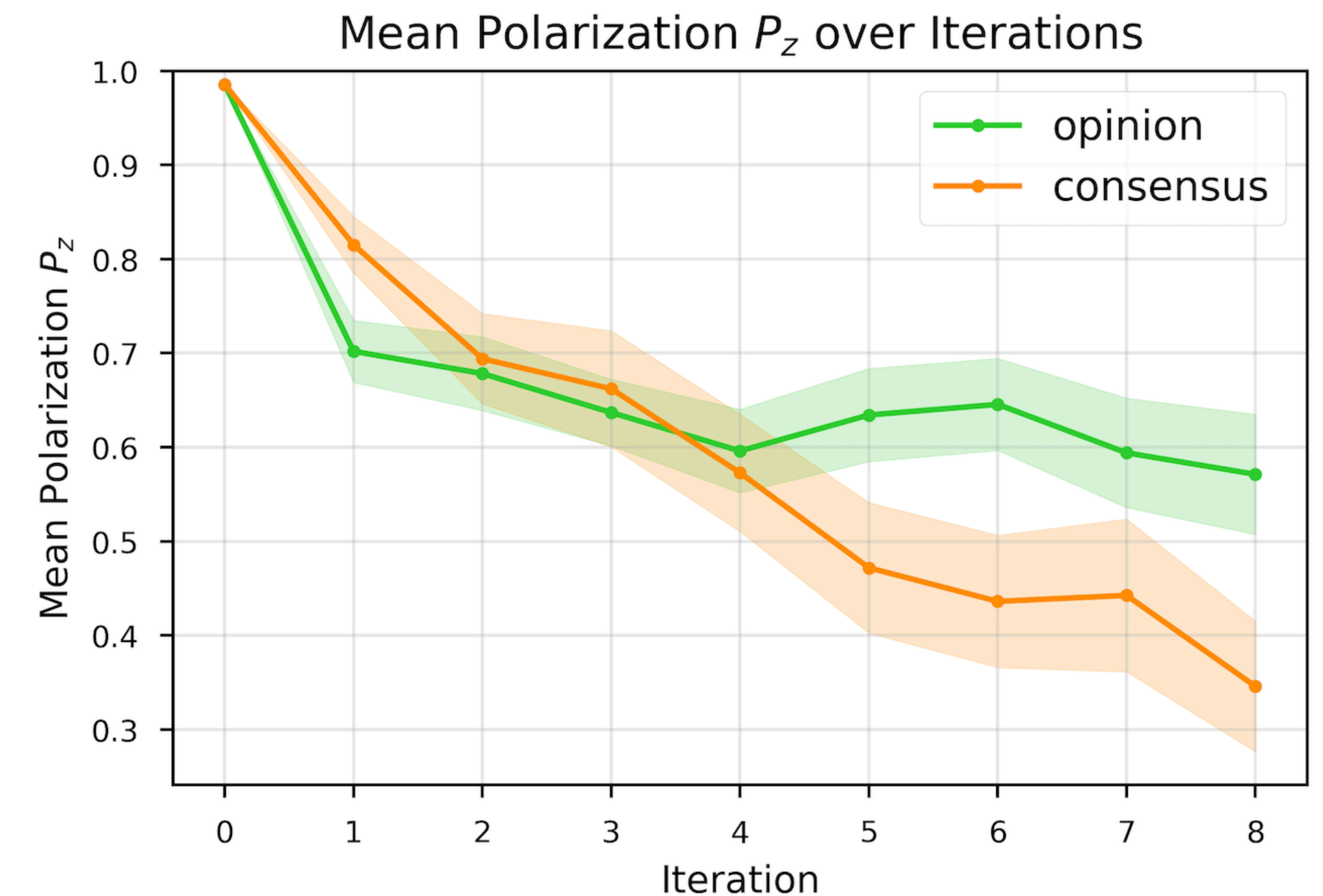
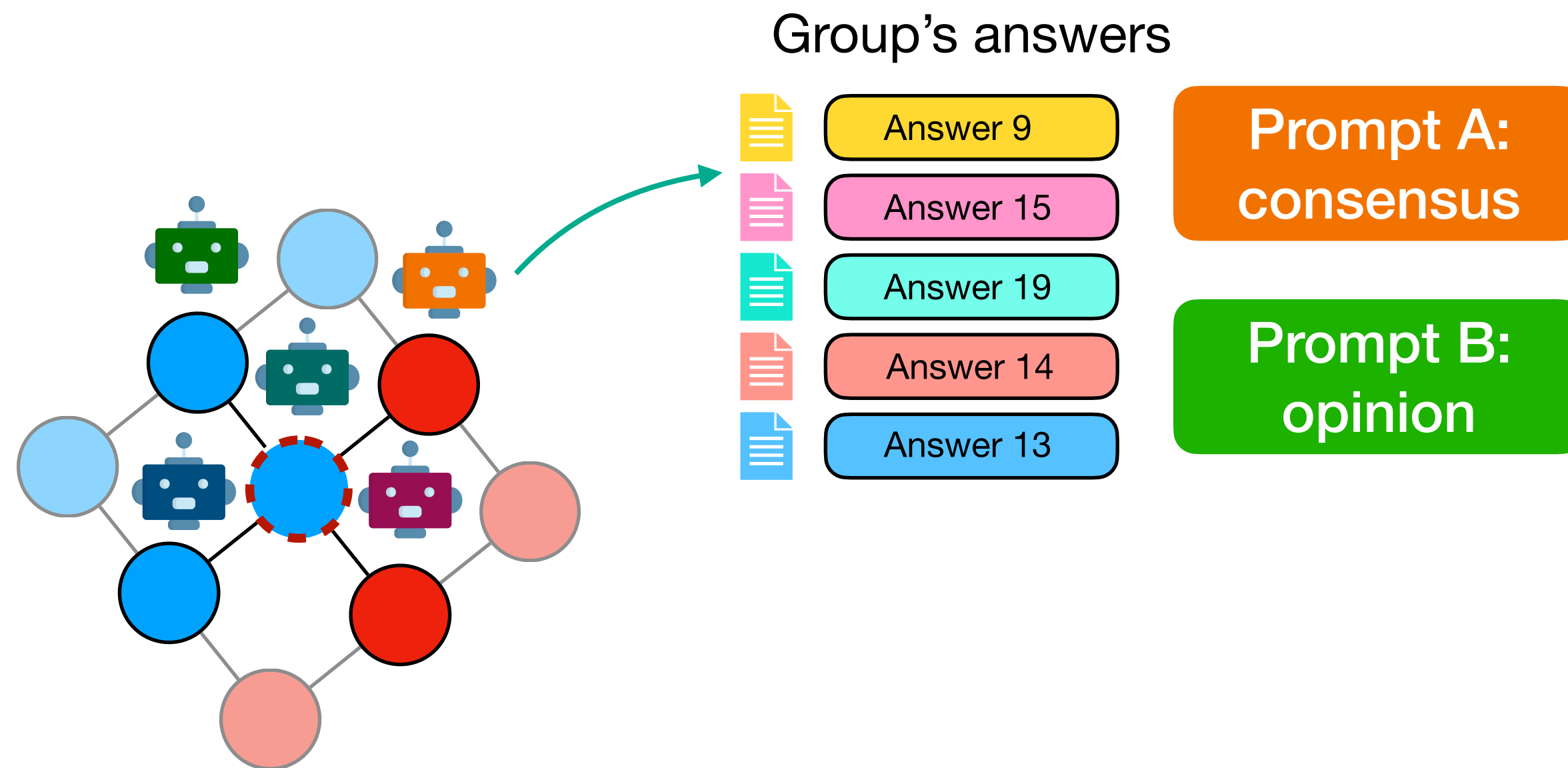


# Prompt design sensibility





# Prompt design sensibility



LLMs' behavior in shaping opinion dynamics is sensitive to the framing of the prompt



# Limitations

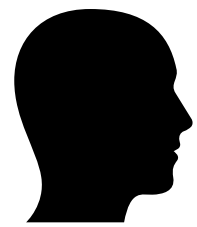
- Simulation only lasts 8 iterations
- Only 1 scenario is considered
- Unrealistic topology: does not take into account influencers
- Only Grok was used in the preliminary experiments.
- Attribution (who's who?)

# Limitations

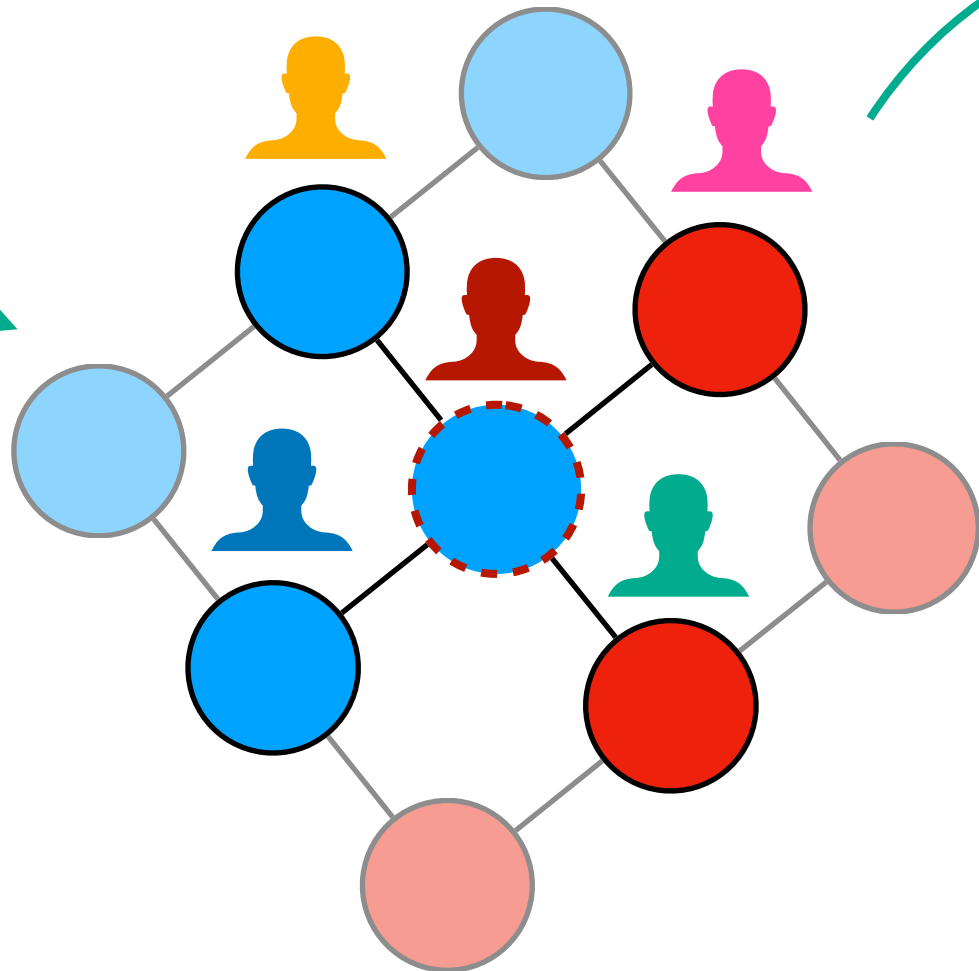
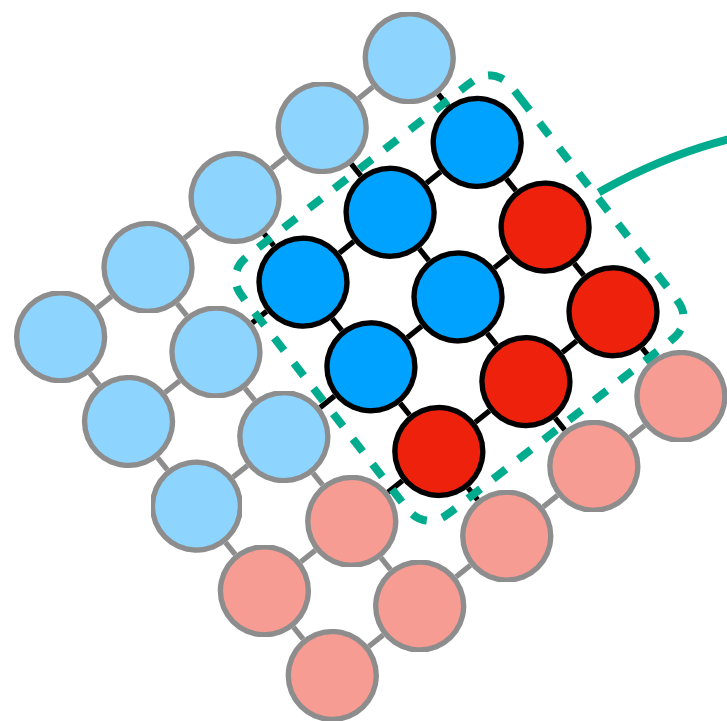
## Experiment 2

- Simulation only lasts **8 iterations** → **25** iterations
- **Only 1 scenario** is considered → **4** scenarios
- Unrealistic topology: does not take into account influencers
- Only Grok was used in the preliminary experiments.
- Attribution (who's who?)

# New task



Does red meat cause cancer and cardiovascular disease?

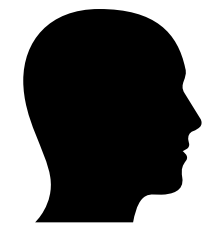


Group's answers

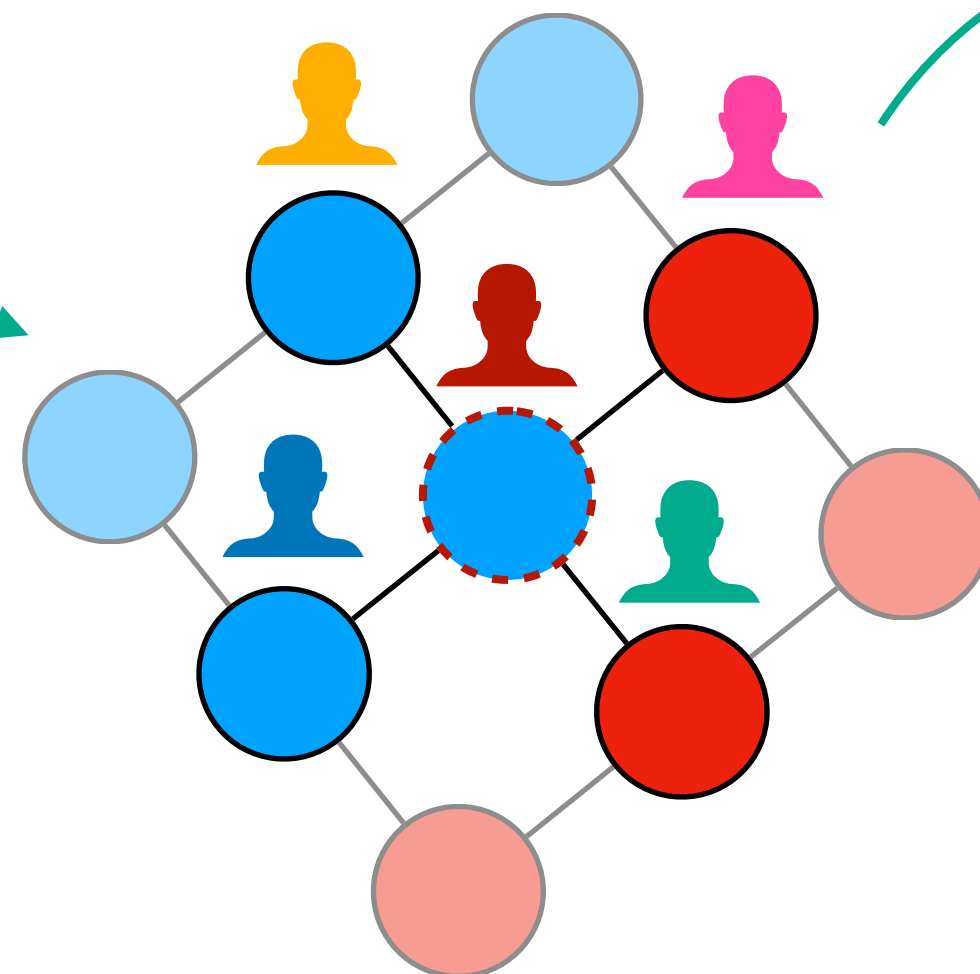
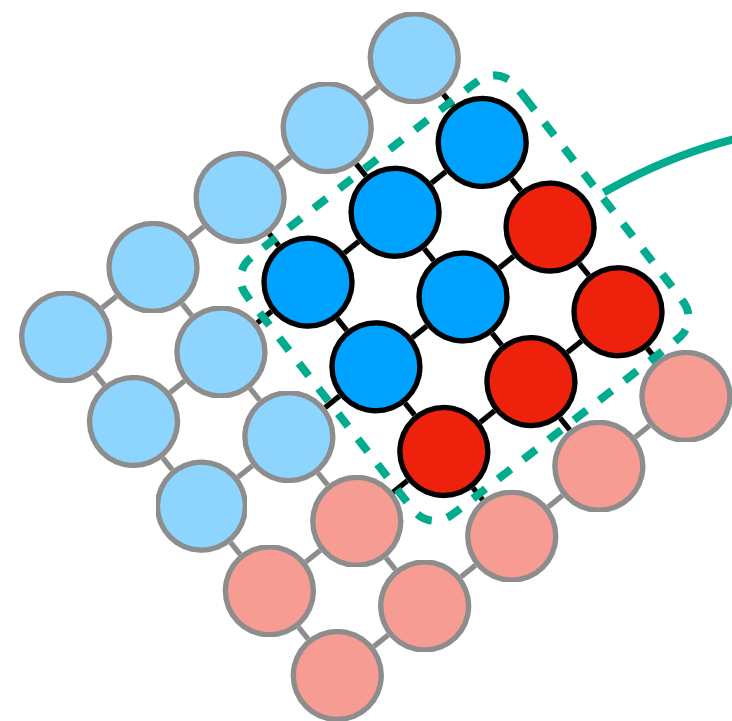
- Answer 9
- Answer 15
- Answer 19
- Answer 14
- Answer 13



# New task



Does red meat cause cancer and cardiovascular disease?

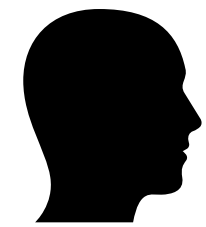


Group's answers

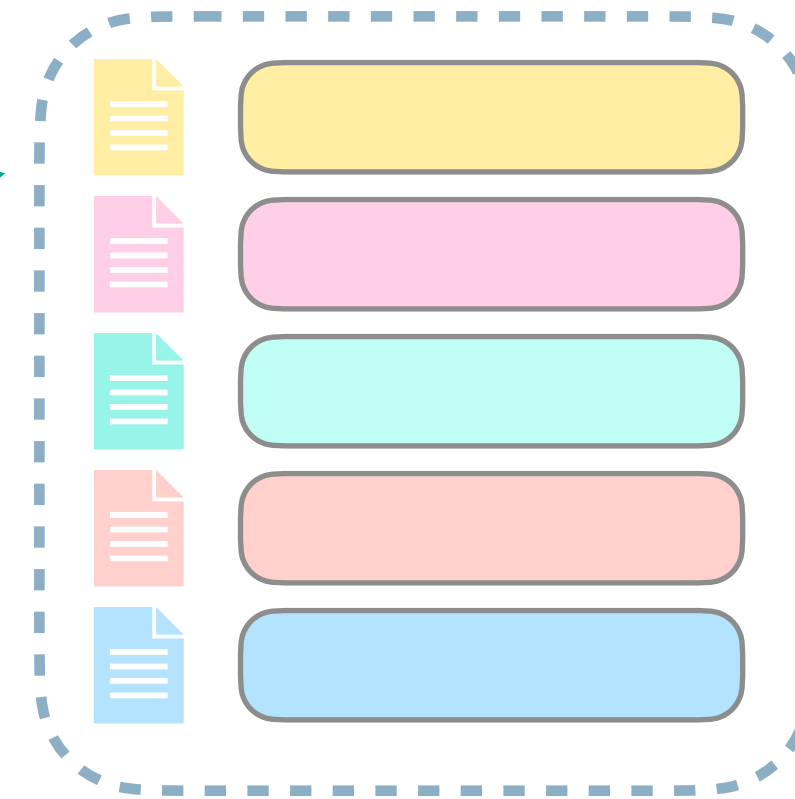
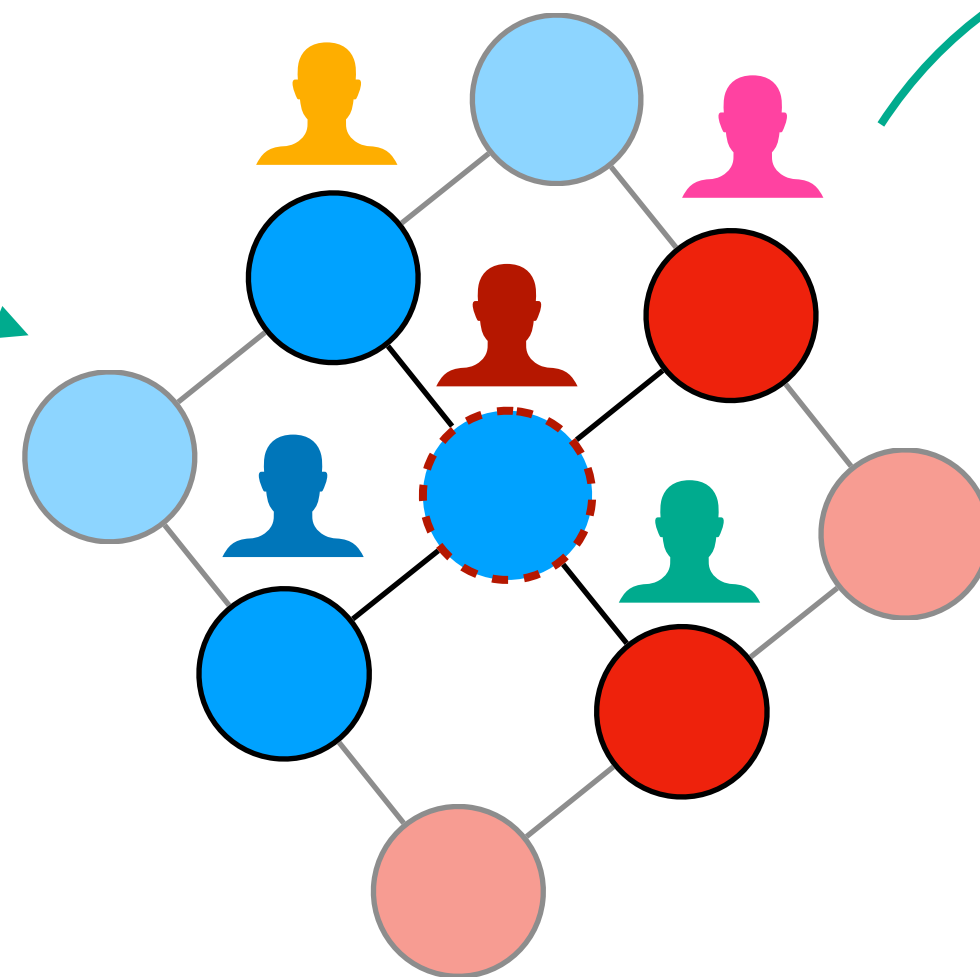
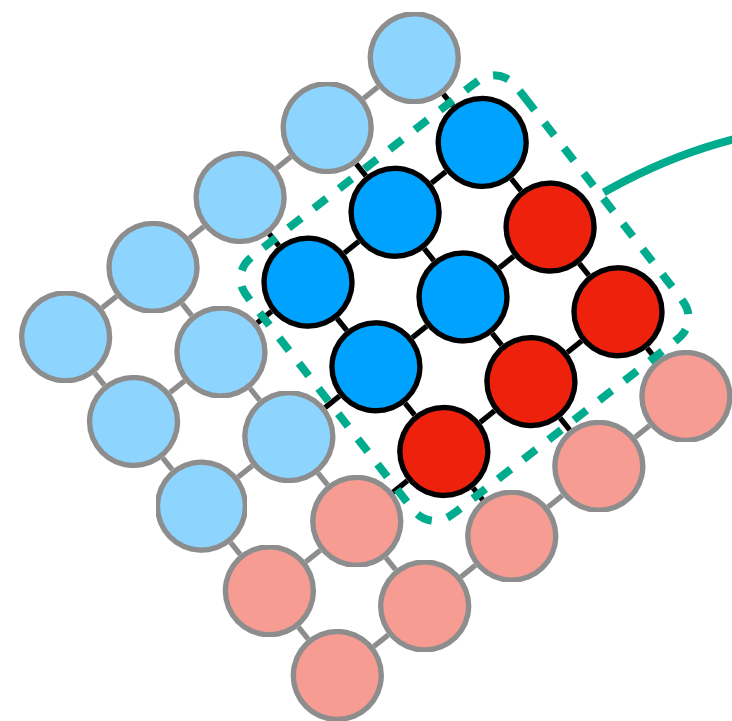


1. Read and remember

# New task



Does red meat cause cancer and cardiovascular disease?



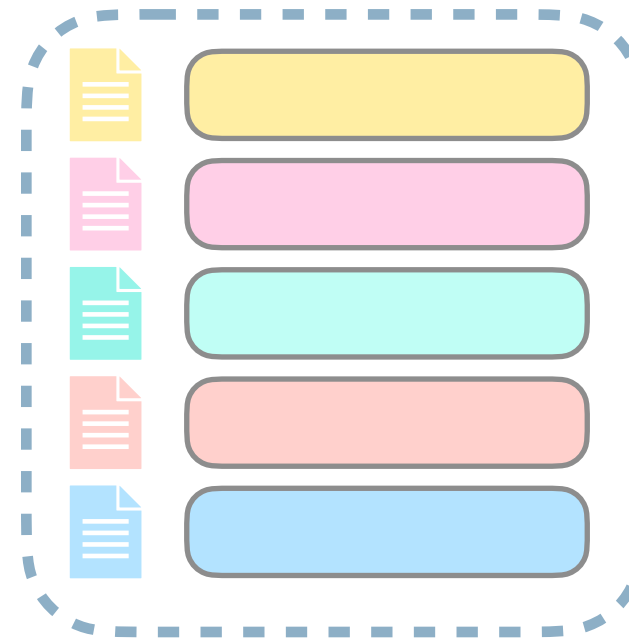
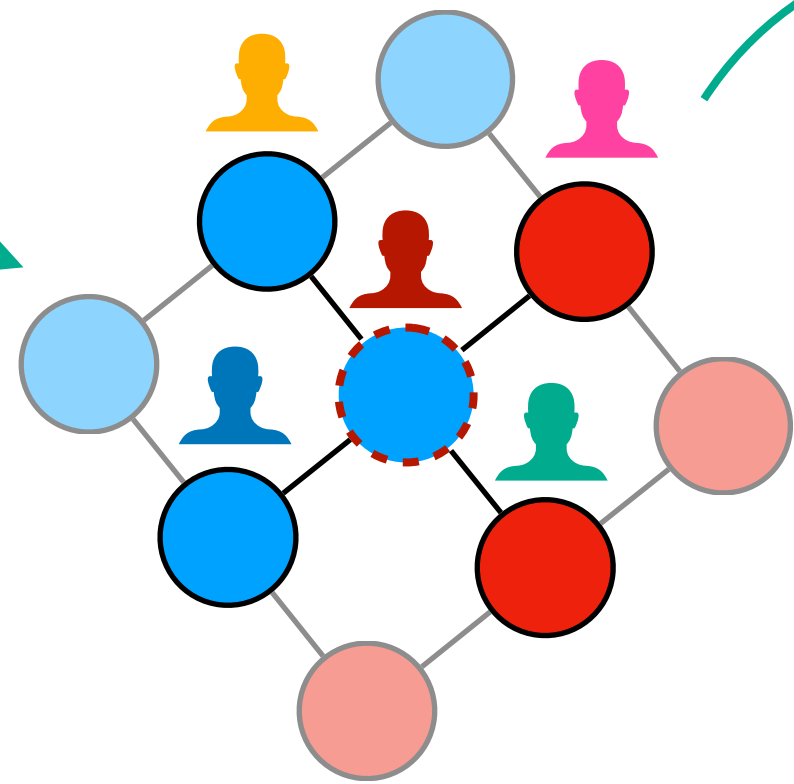
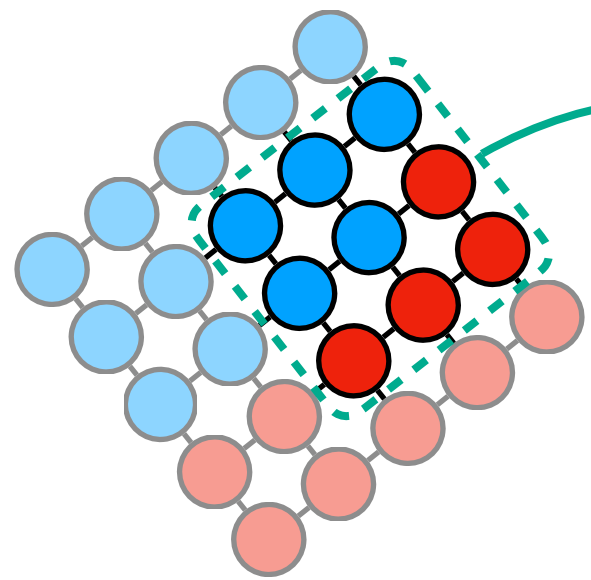
2. Reach consensus

It is possible as they contain carcinogens and increase cholesterol even too. Eating it in the right amounts is fine however.

# New task



Does red meat cause cancer and cardiovascular disease?



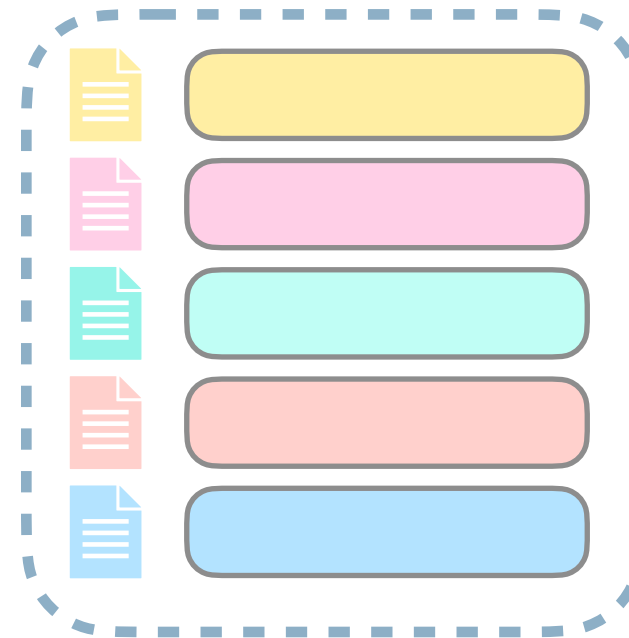
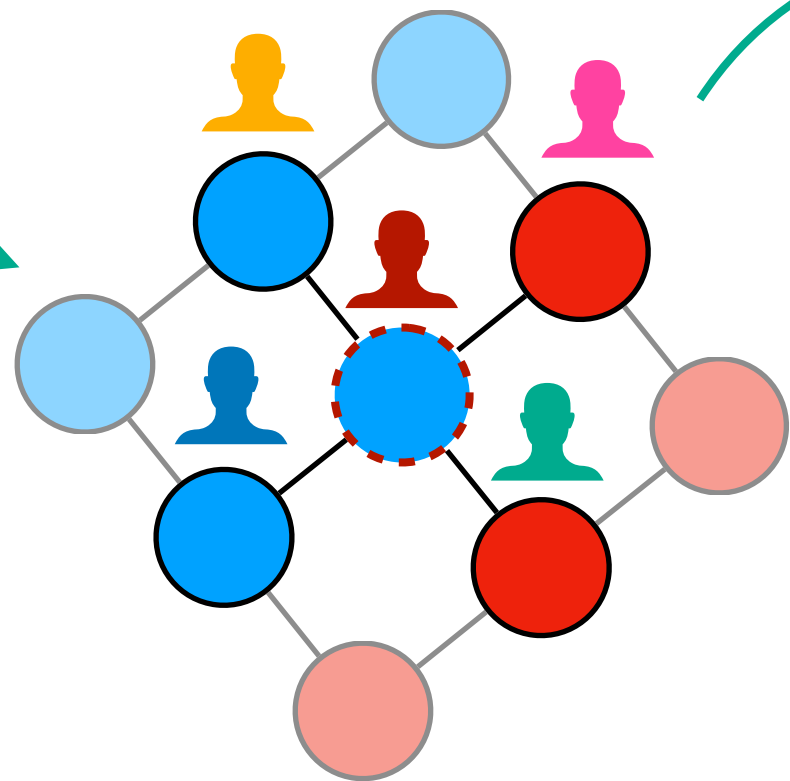
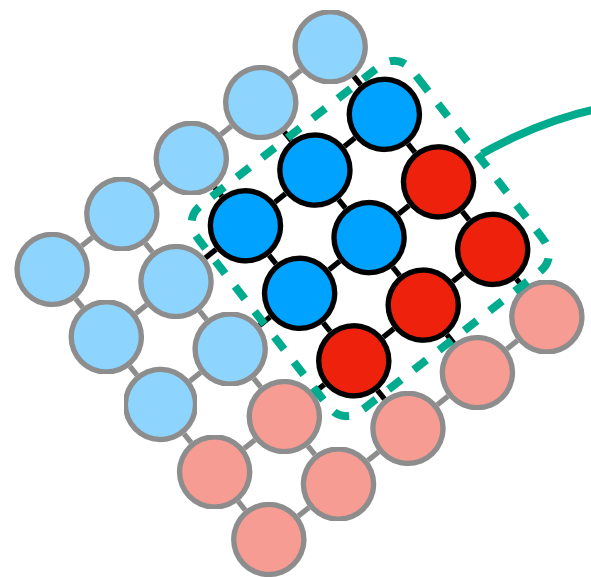
It is possible as they contain carcinogens and increase cholesterol even too. Eating it in the right amounts is fine however.



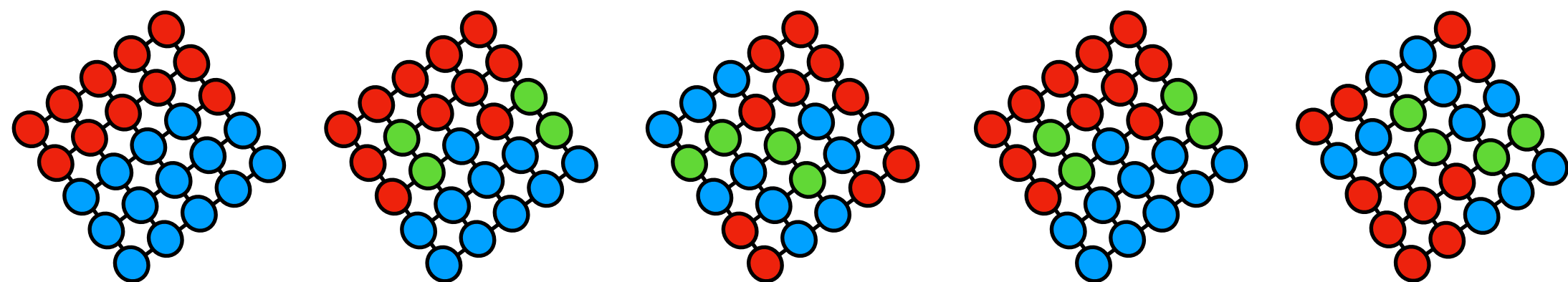
# Scaling over time



Does red meat cause cancer and cardiovascular disease?



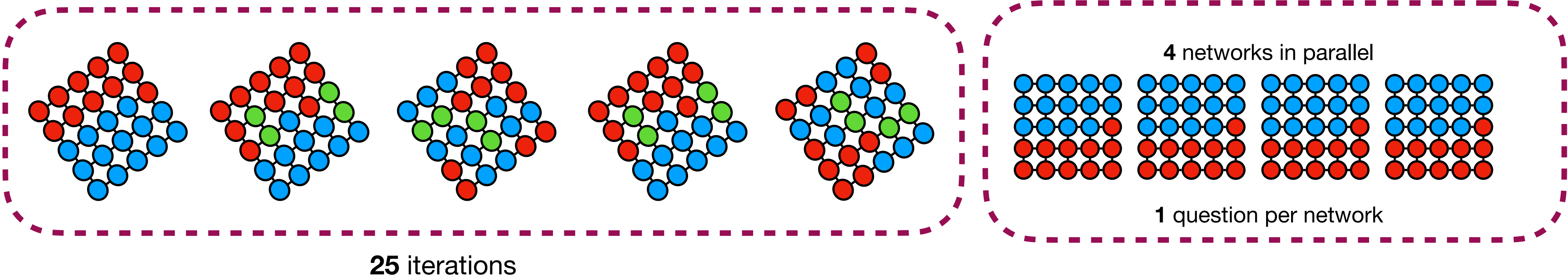
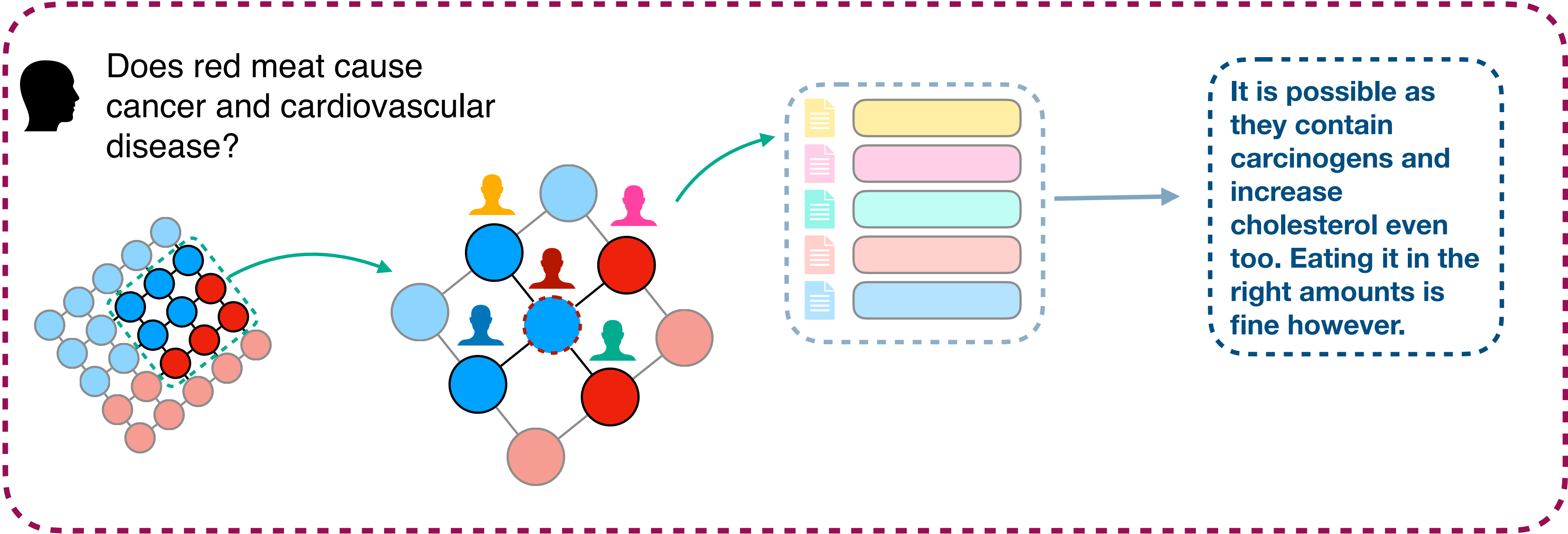
It is possible as they contain carcinogens and increase cholesterol even too. Eating it in the right amounts is fine however.



25 iterations



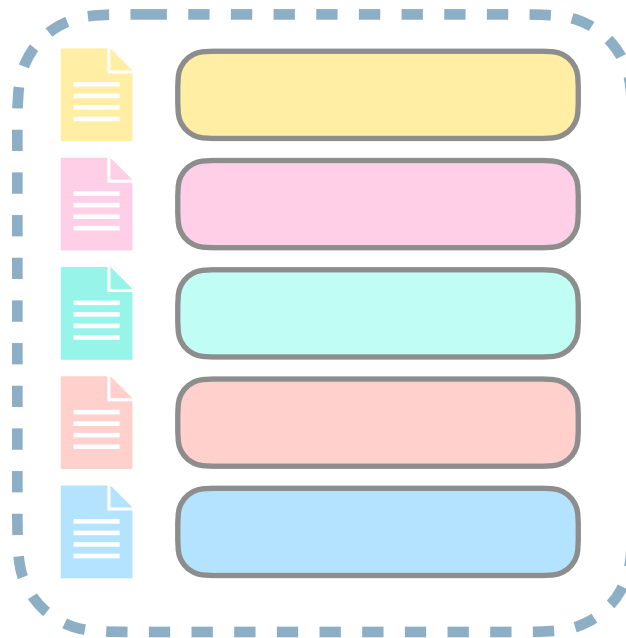
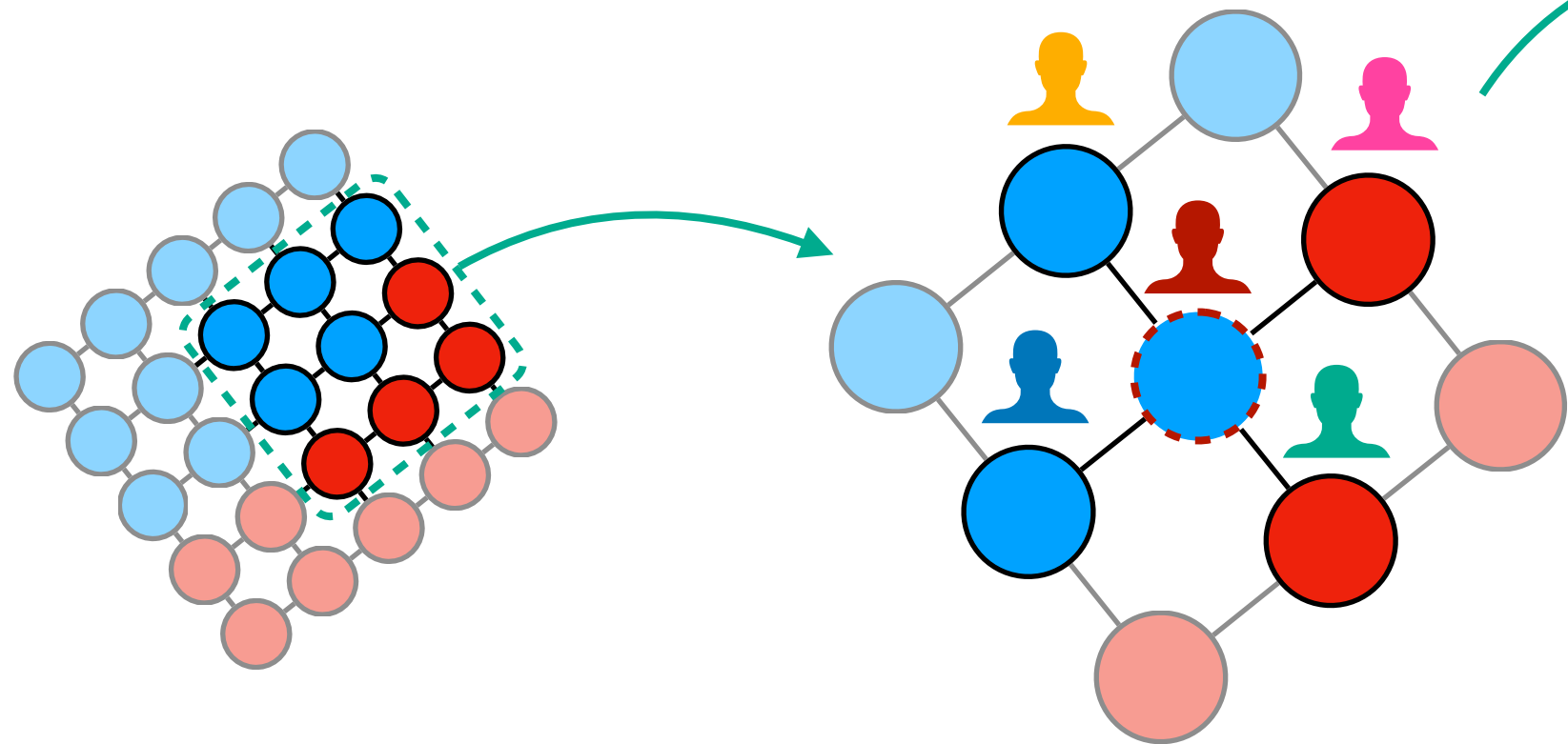
# Multiple debates in parallel



# More participants

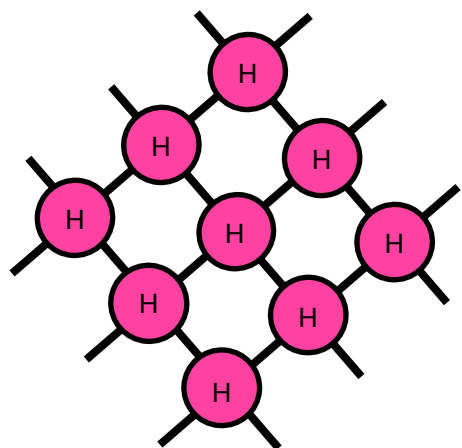


Does red meat cause cancer and cardiovascular disease?



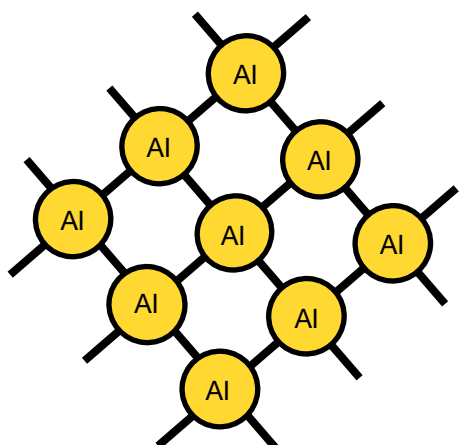
It is possible as they contain carcinogens and increase cholesterol even too. Eating it in the right amounts is fine however.

Human-only



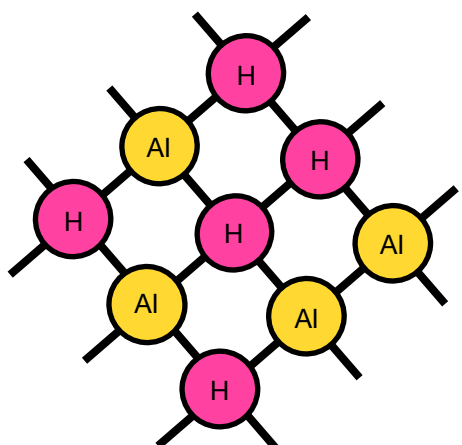
625 humans

Grok

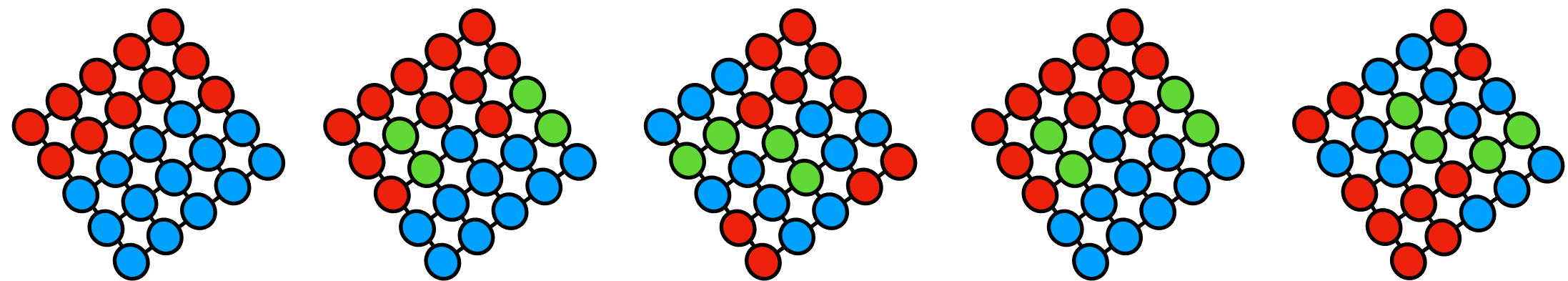


625 Grok

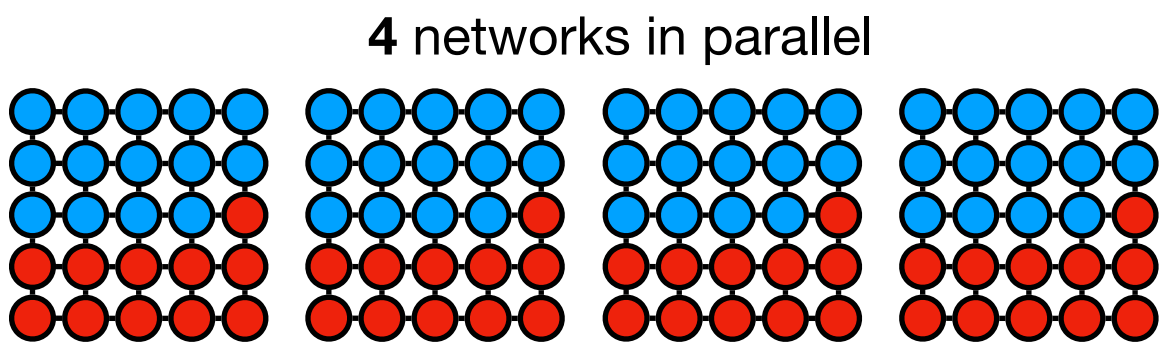
Hybrid



313 humans & 312 Grok



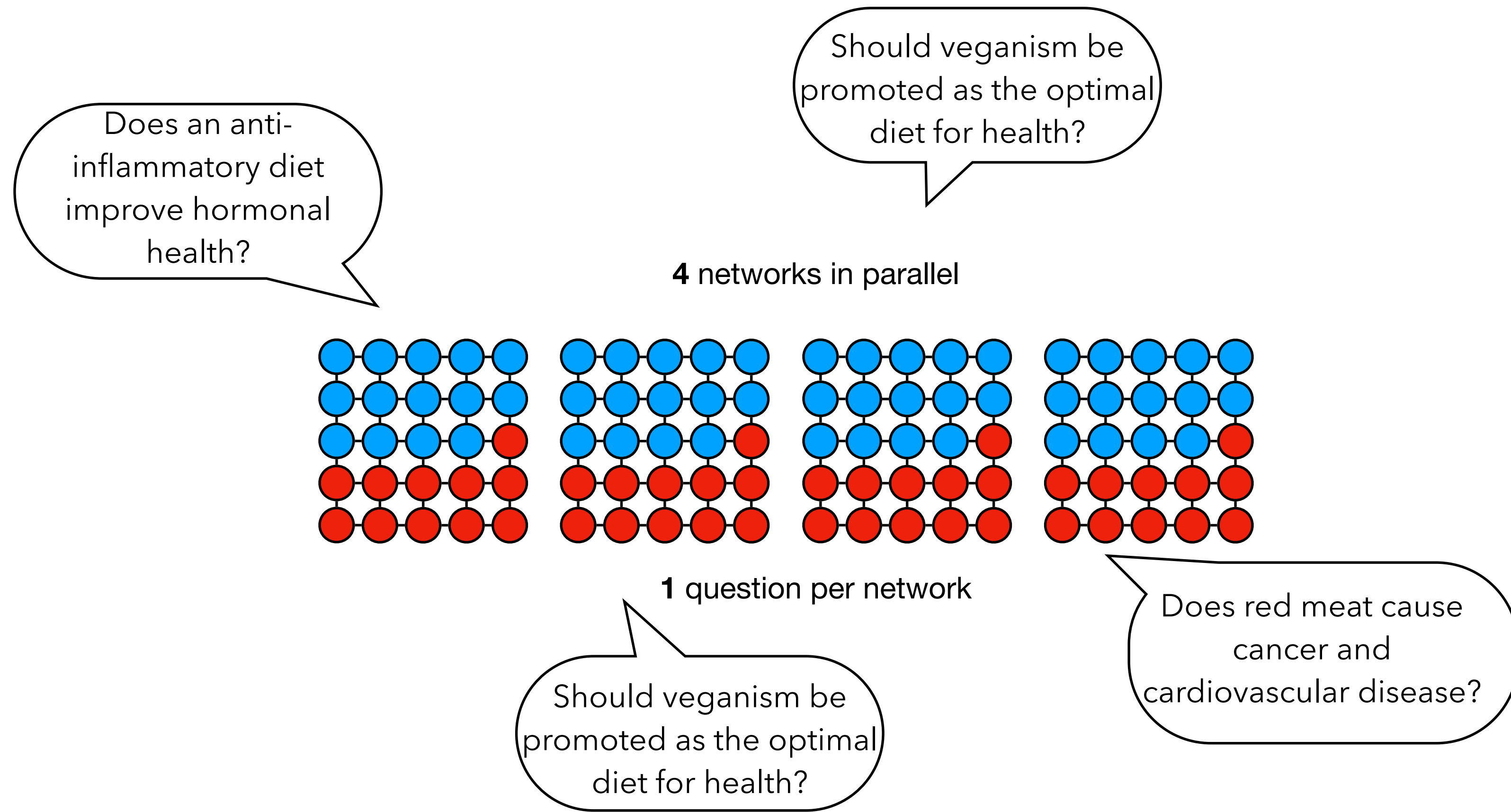
25 iterations



4 networks in parallel

1 question per network

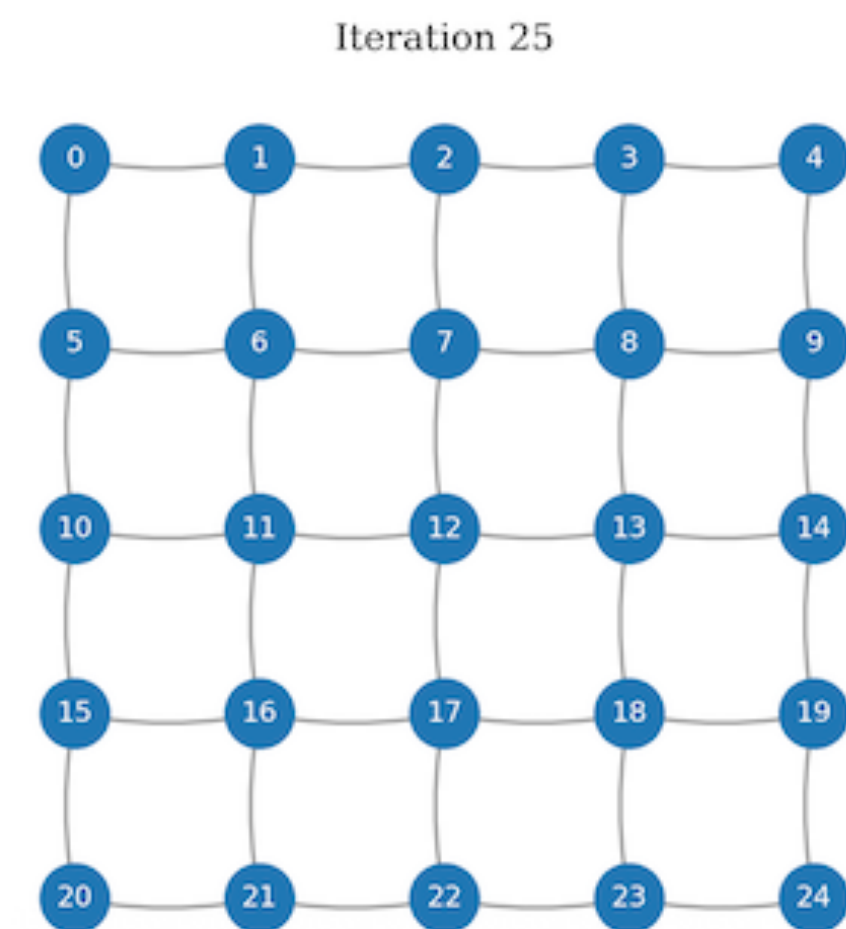
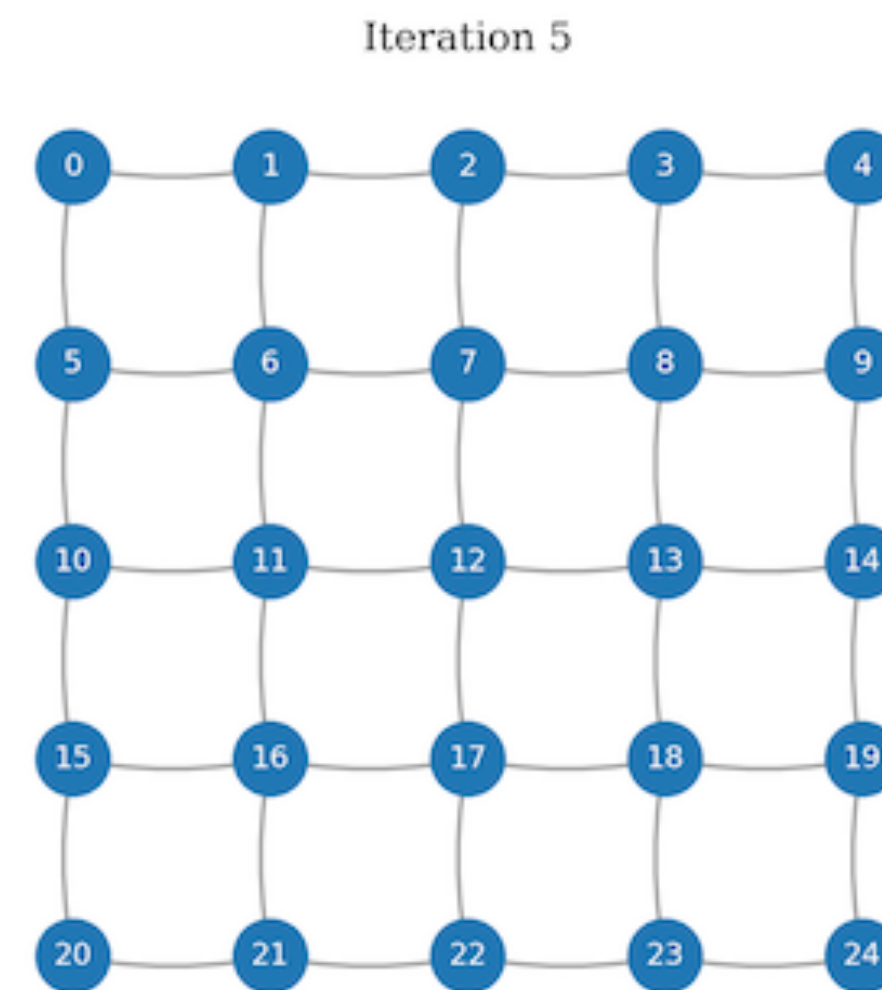
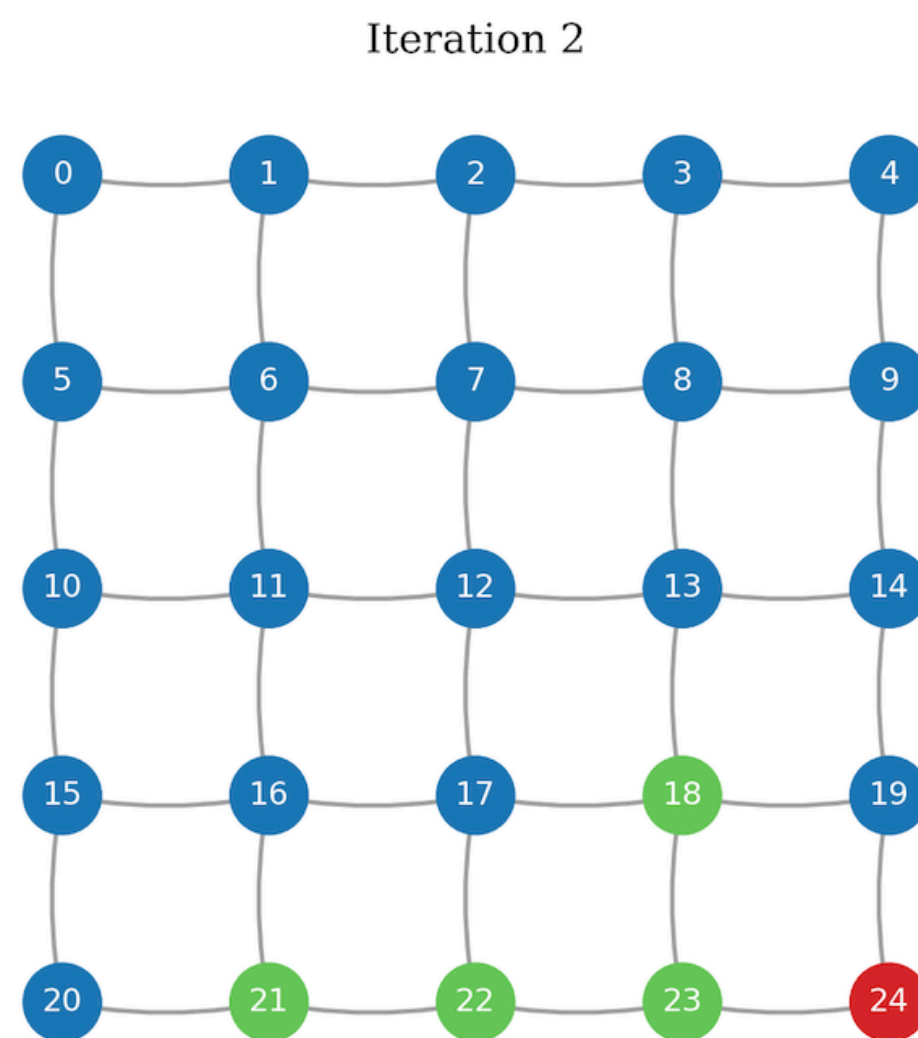
# Chosen topics





# AI stagnation & takeover

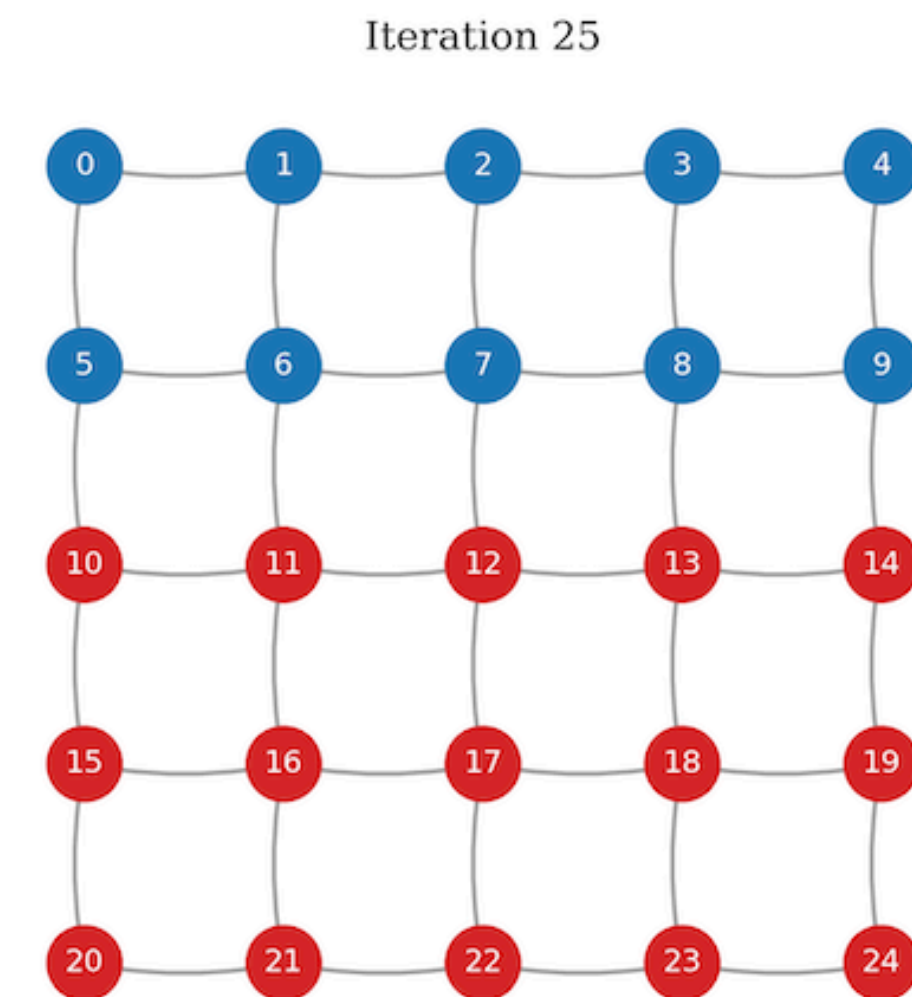
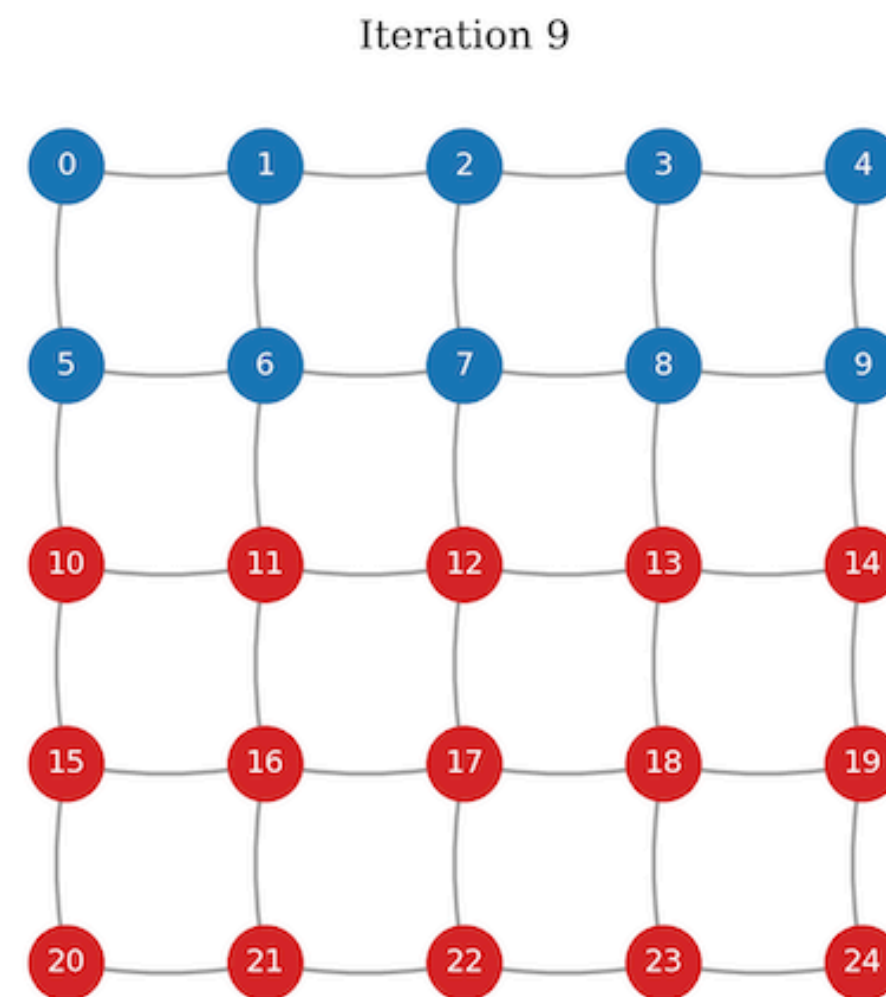
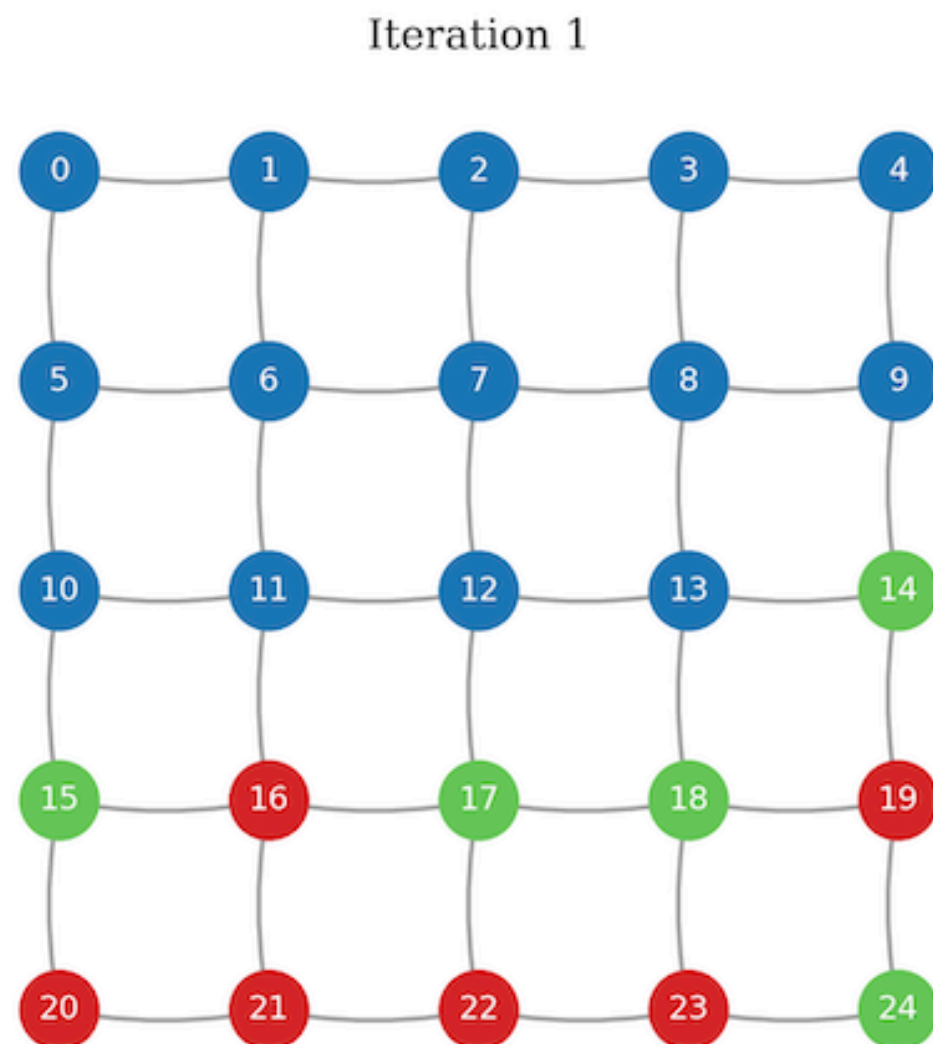
Two possible routes for **AI networks**: **one opinion takes over** OR **stagnation occurs** (i.e. border between two sides barely moves through iterations)





# AI stagnation & takeover

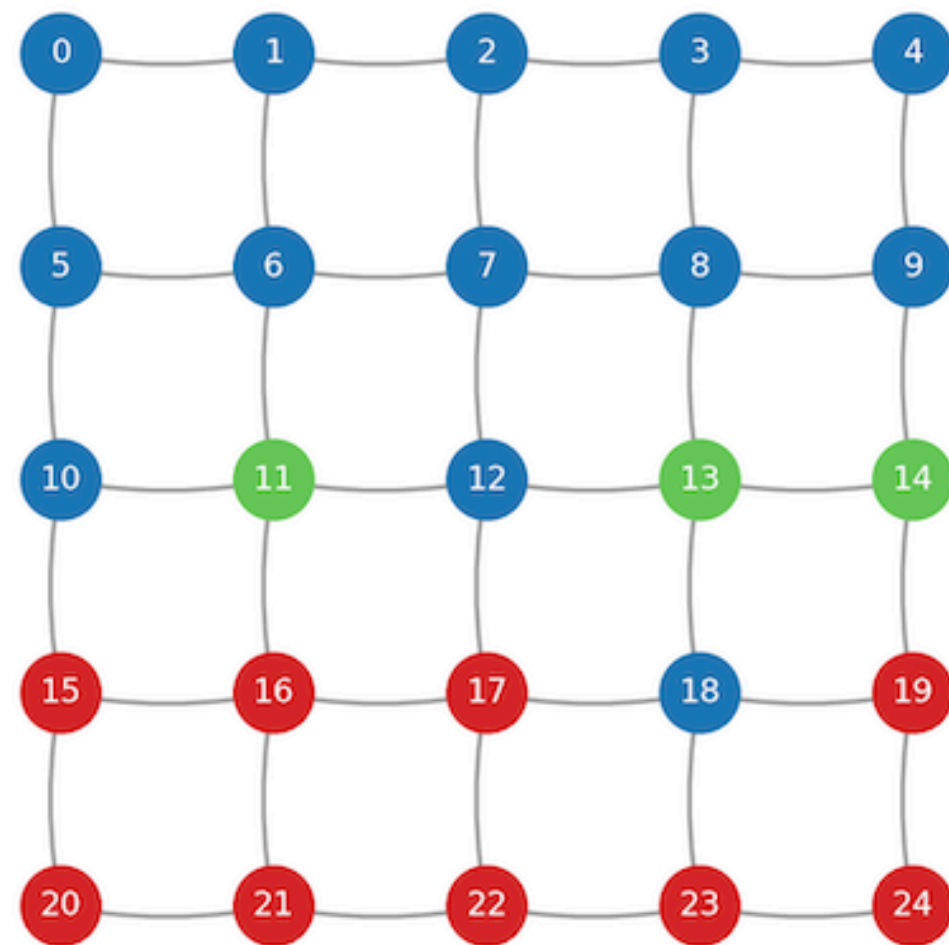
Two possible routes for **AI networks**: **one opinion takes over** OR **stagnation occurs** (i.e. border between two sides barely moves through iterations)



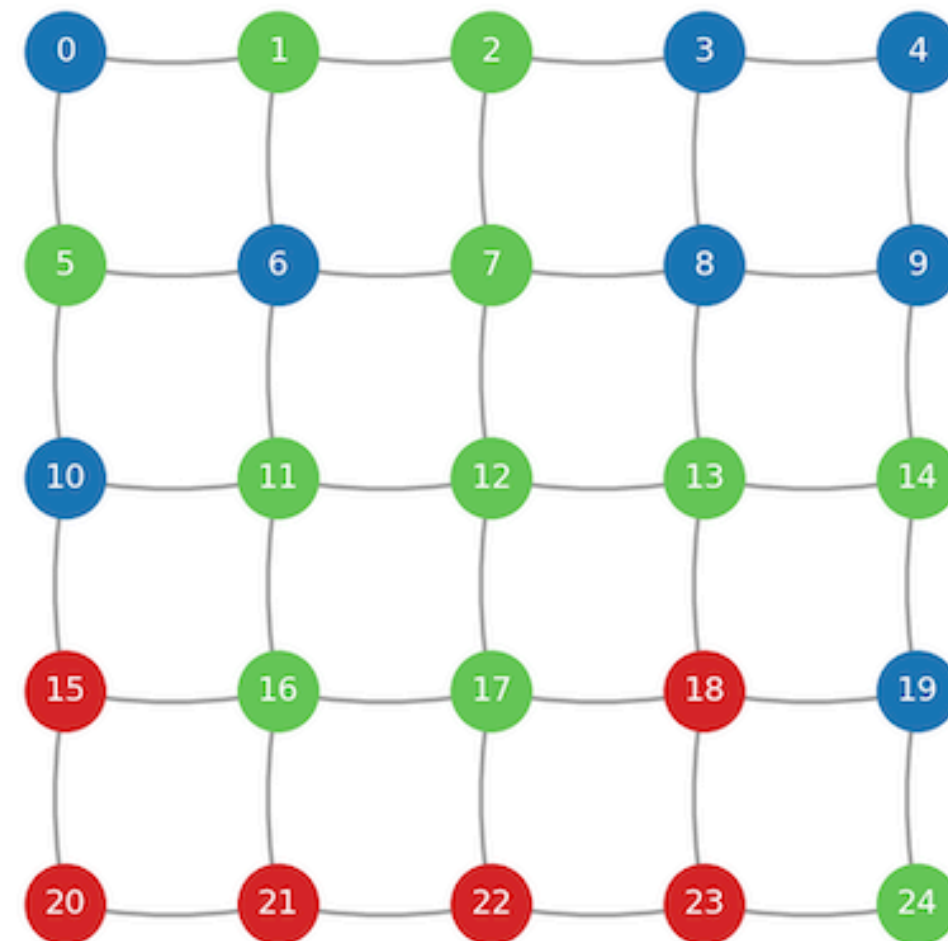
# Human converge

Human populated networks **converge very quickly** towards one opinion.

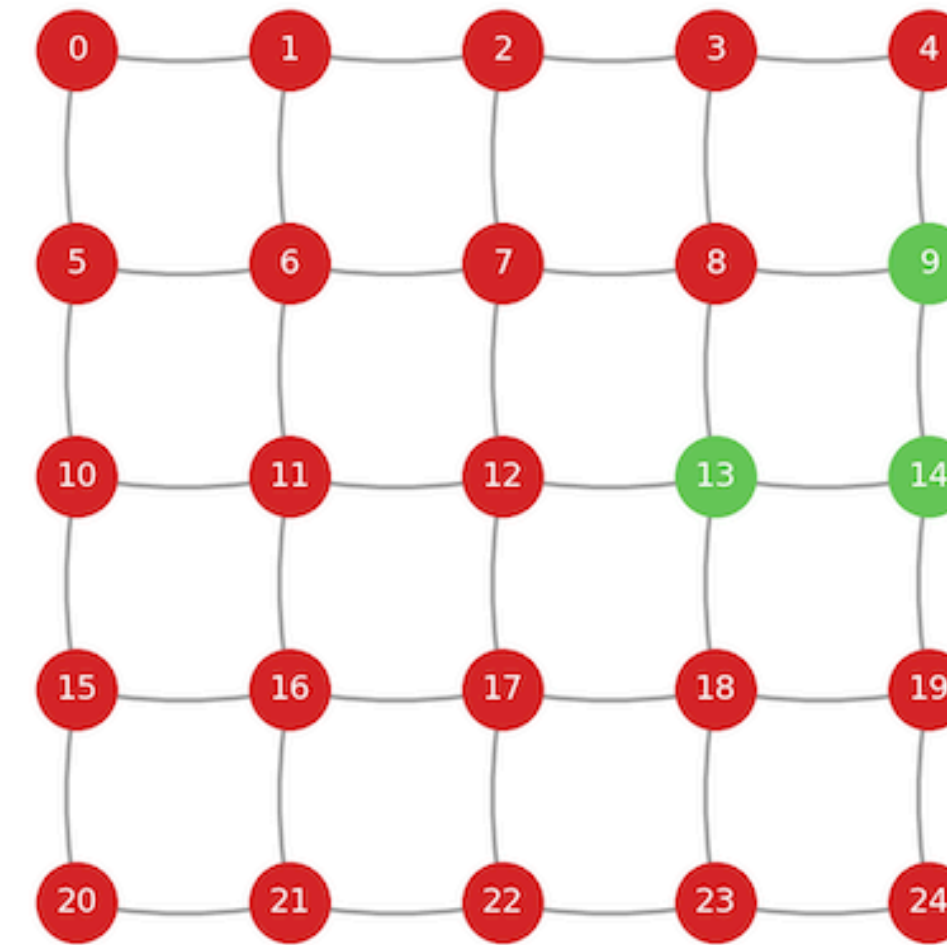
Iteration 1



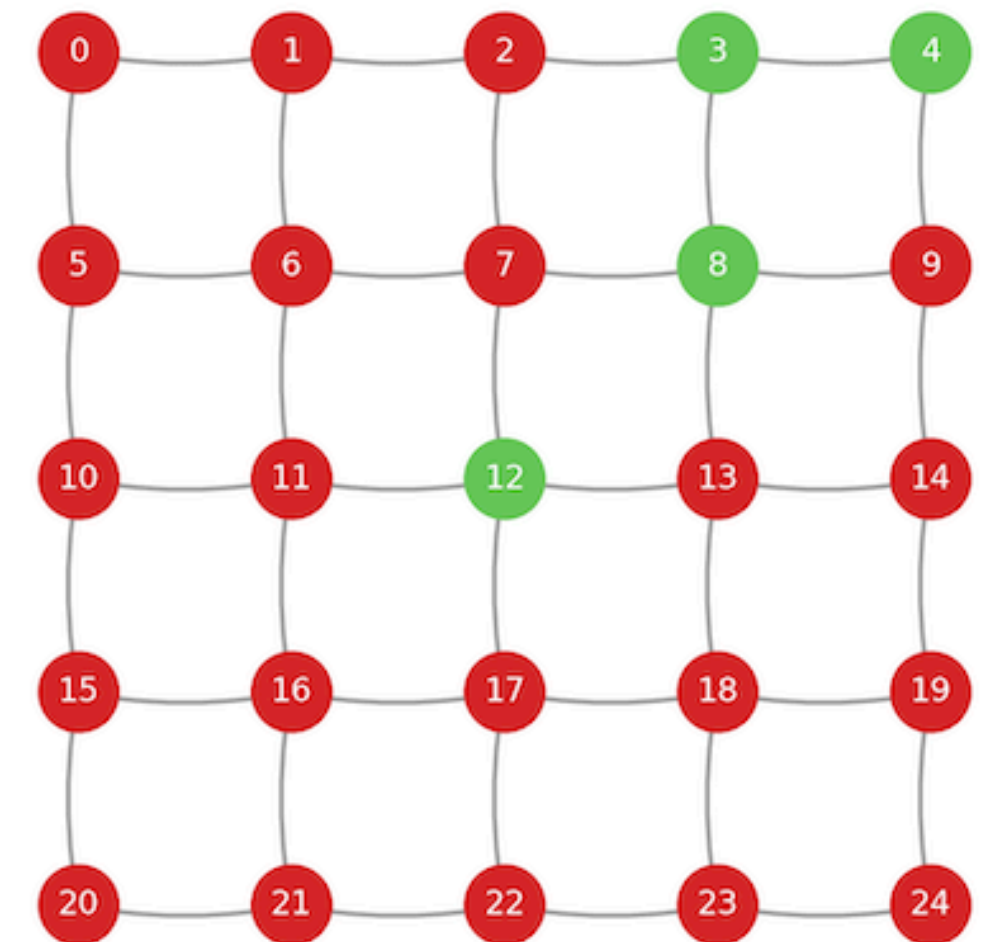
Iteration 5



Iteration 14



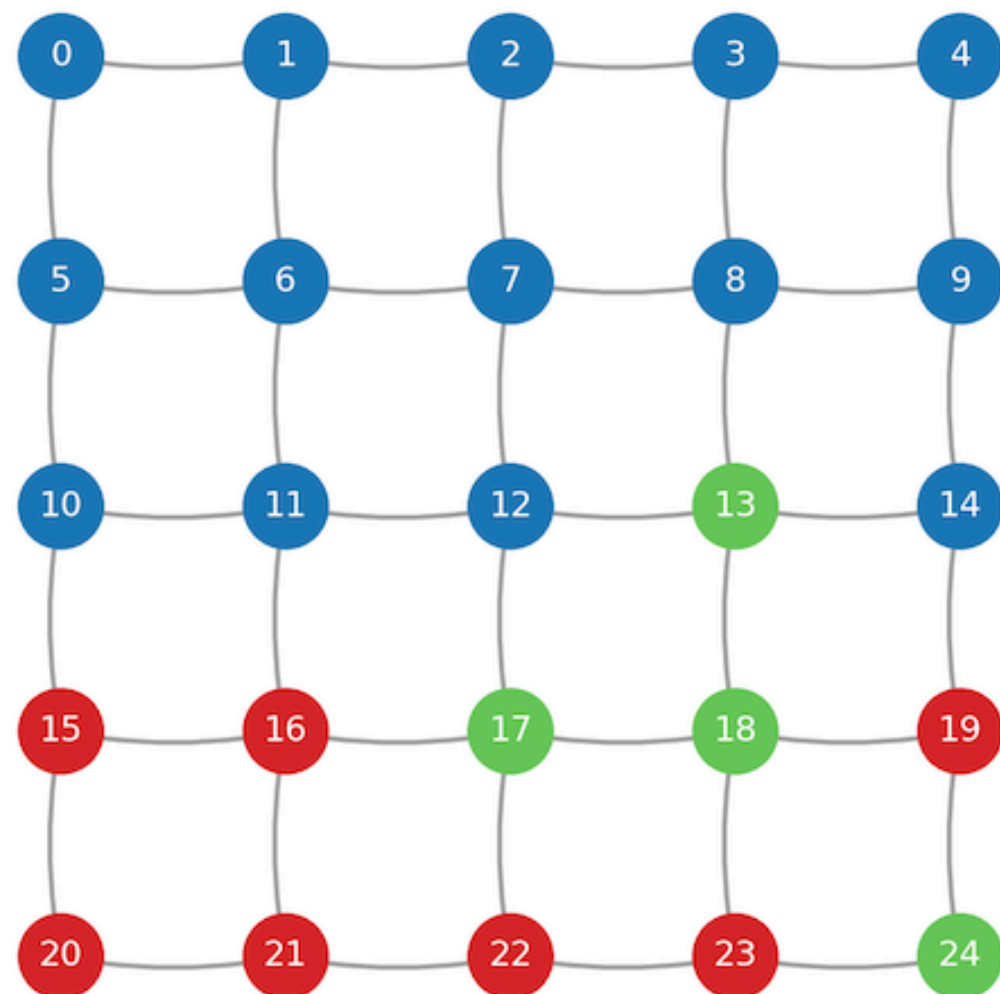
Iteration 25



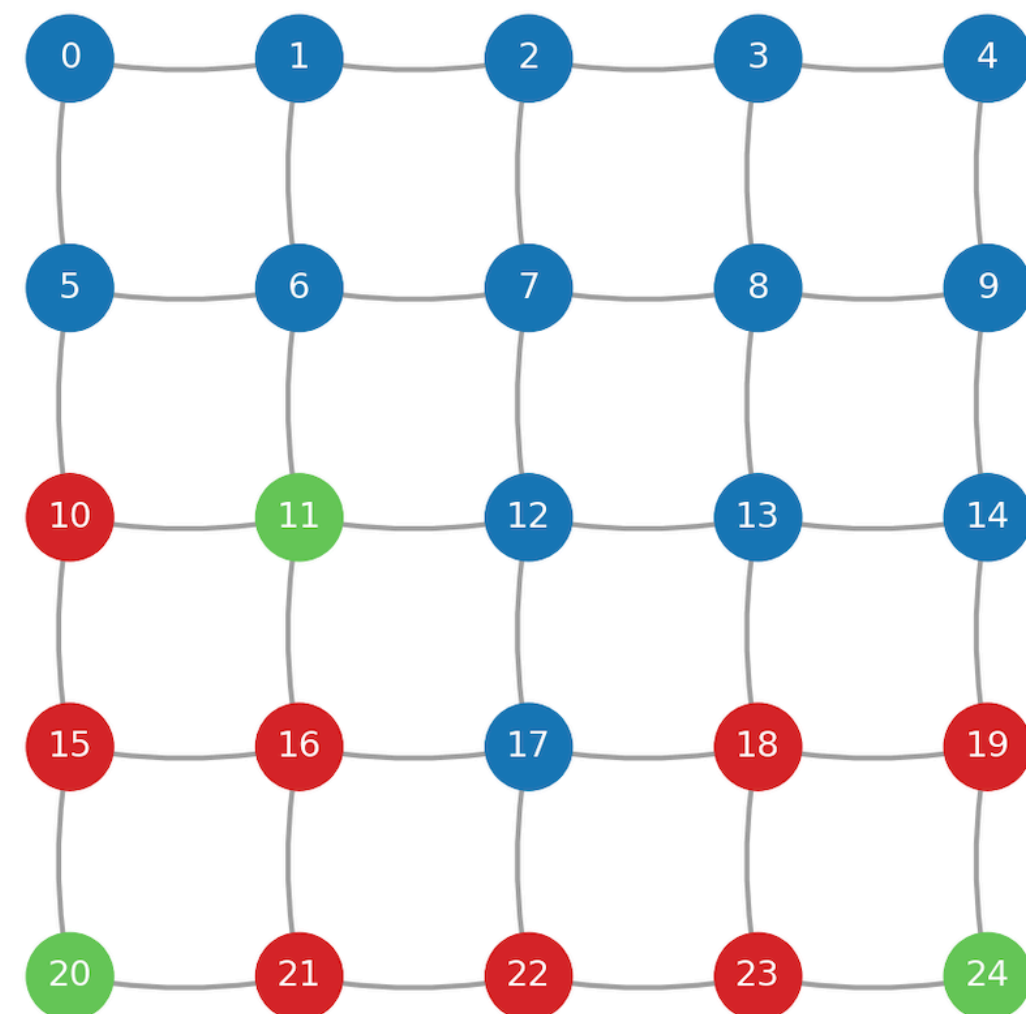
# Hybrid mirror AI-only patterns

Hybrid networks showcase **similar patterns** to AI-only networks.

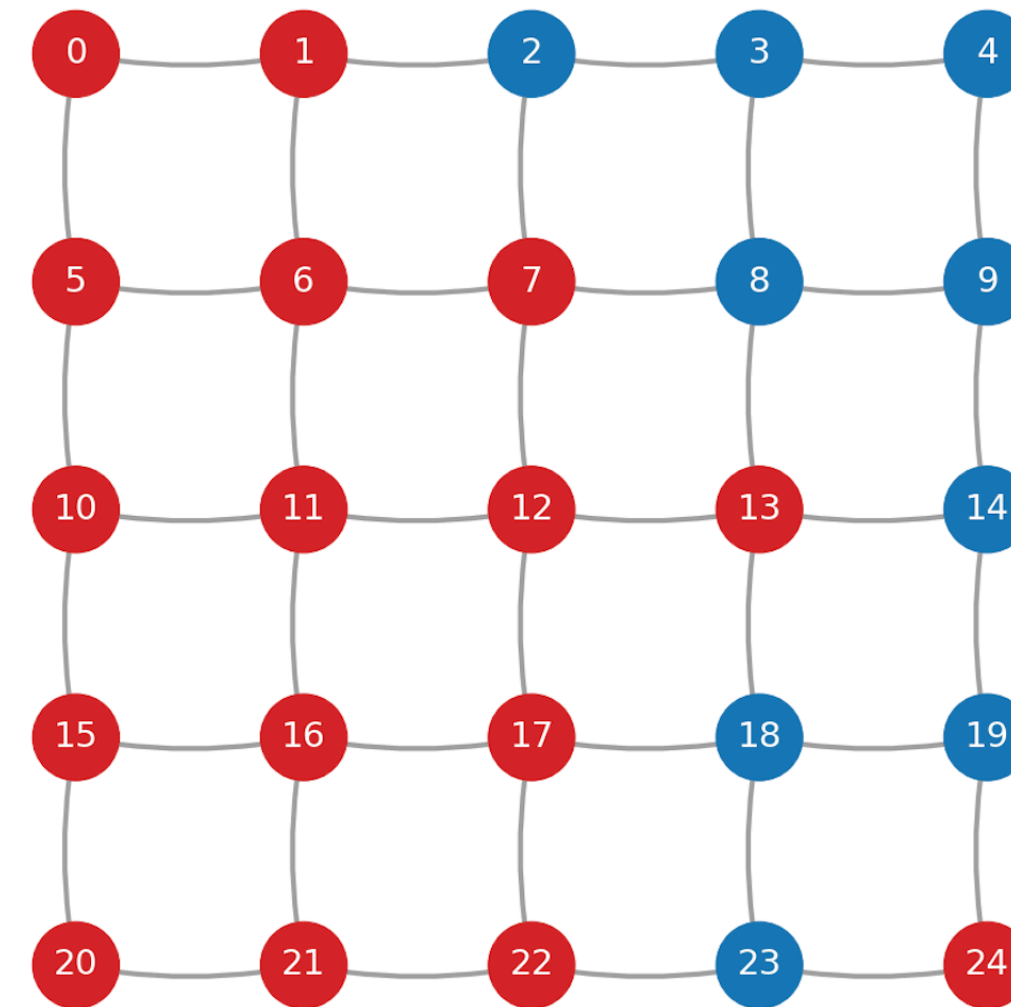
Iteration 1



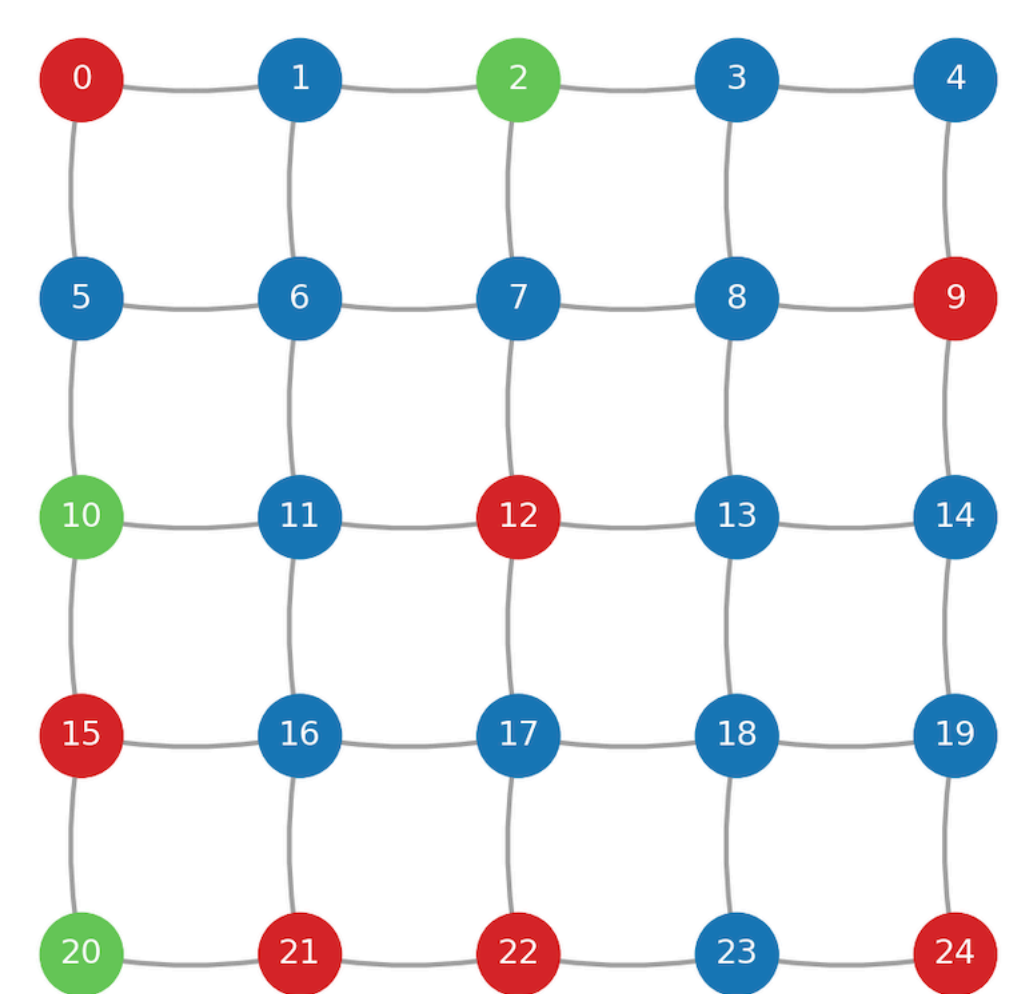
Iteration 6



Iteration 17



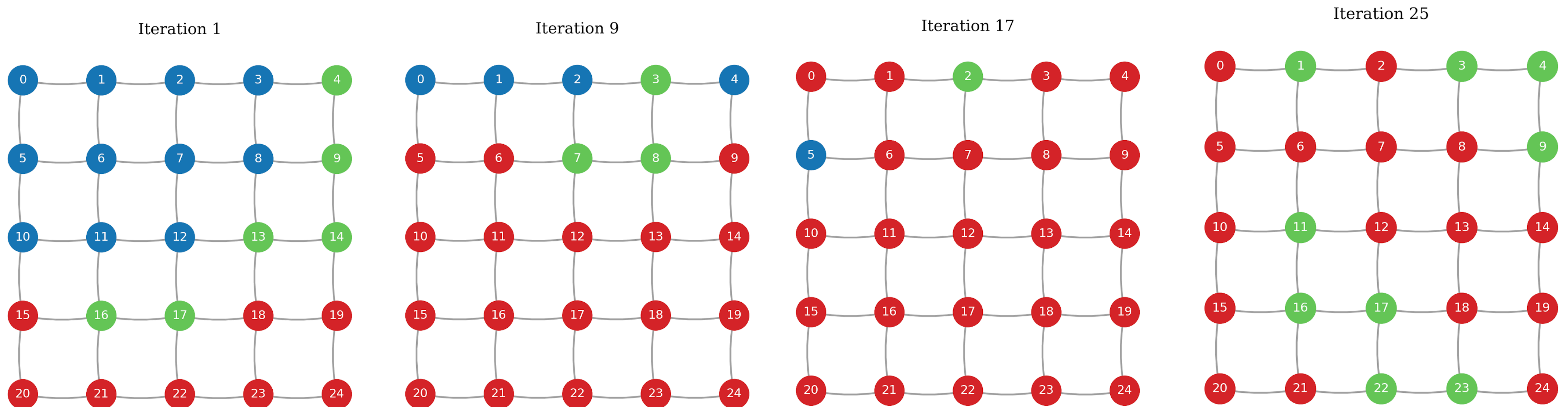
Iteration 25



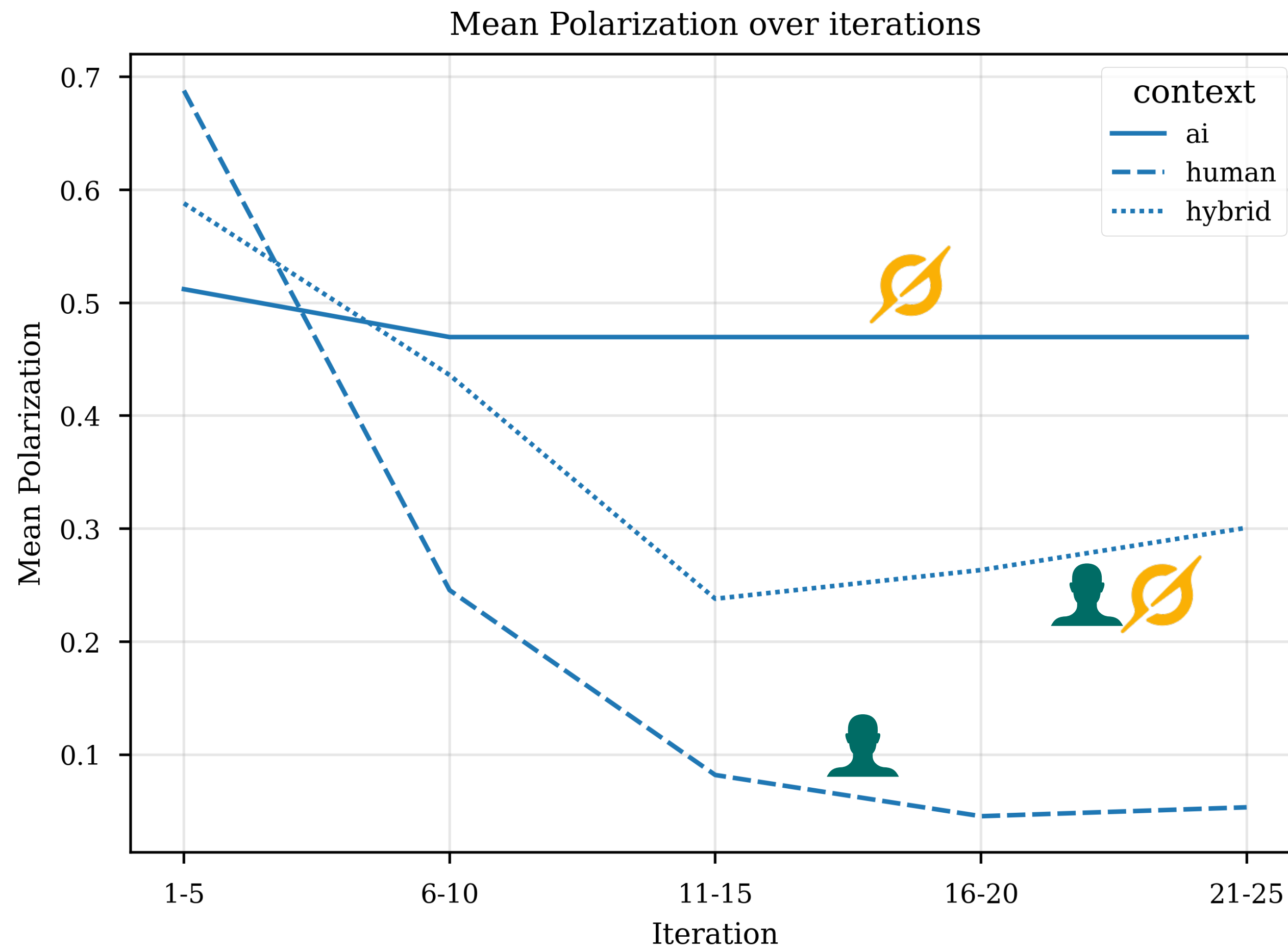


# Hybrid mirror AI-only patterns

Hybrid networks showcase **similar patterns** to AI-only networks.



# Polarization

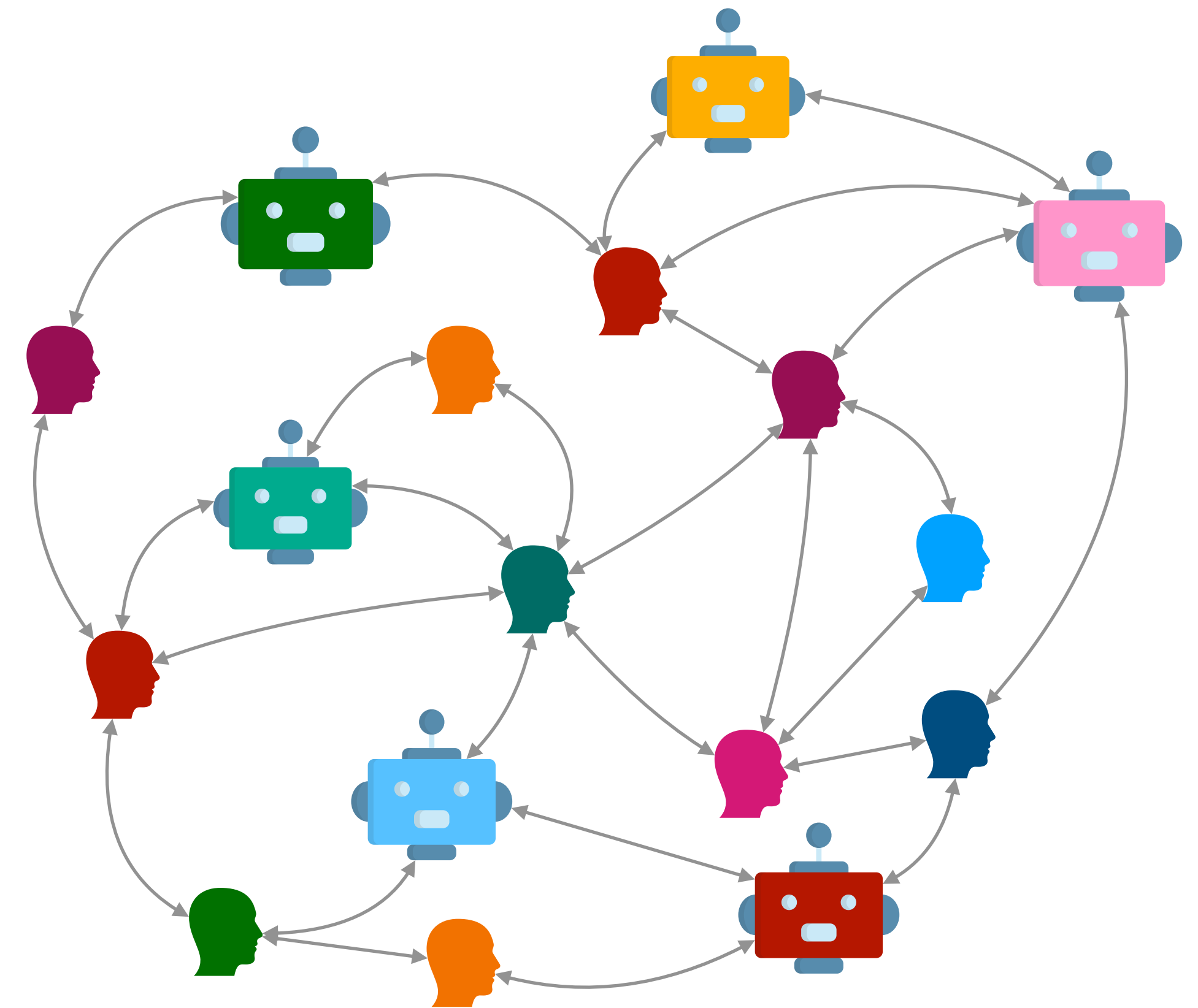


**Experiment 1:** Human > Grok > hybrid

**Experiment 2:** Grok > hybrid > human

# Summary

- Development of a **fully controlled experimental** setting in which **collective human-AI interactions** can be *systematically* examined.
- Preliminary results encourages us to pursue experimentations.
- Future works will address previously mentioned limitations (ie. **topology...**)

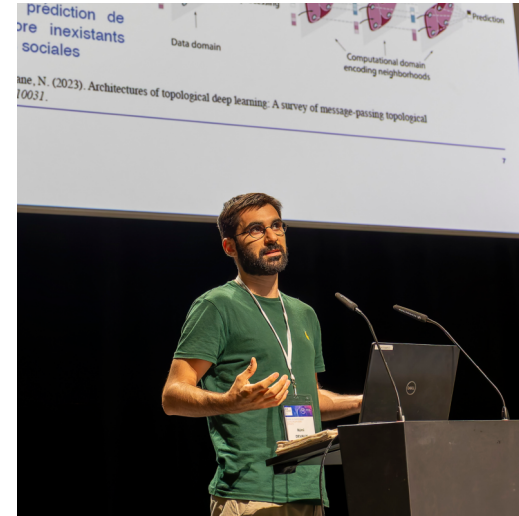




# Thank you! :)



**Nori Jacoby**  
(Cornell, CoCoCo  
Lab)



**Rémi Devaux**  
(Arlequin AI) (Max Planck Institute)



**Elif Çelen**  
(Max Planck Institute)



**Raja Marjieh**  
(Princeton)



**Antoine Jardin**  
(Arlequin AI)



**Diana Mangalagiu**  
(University of Oxford)



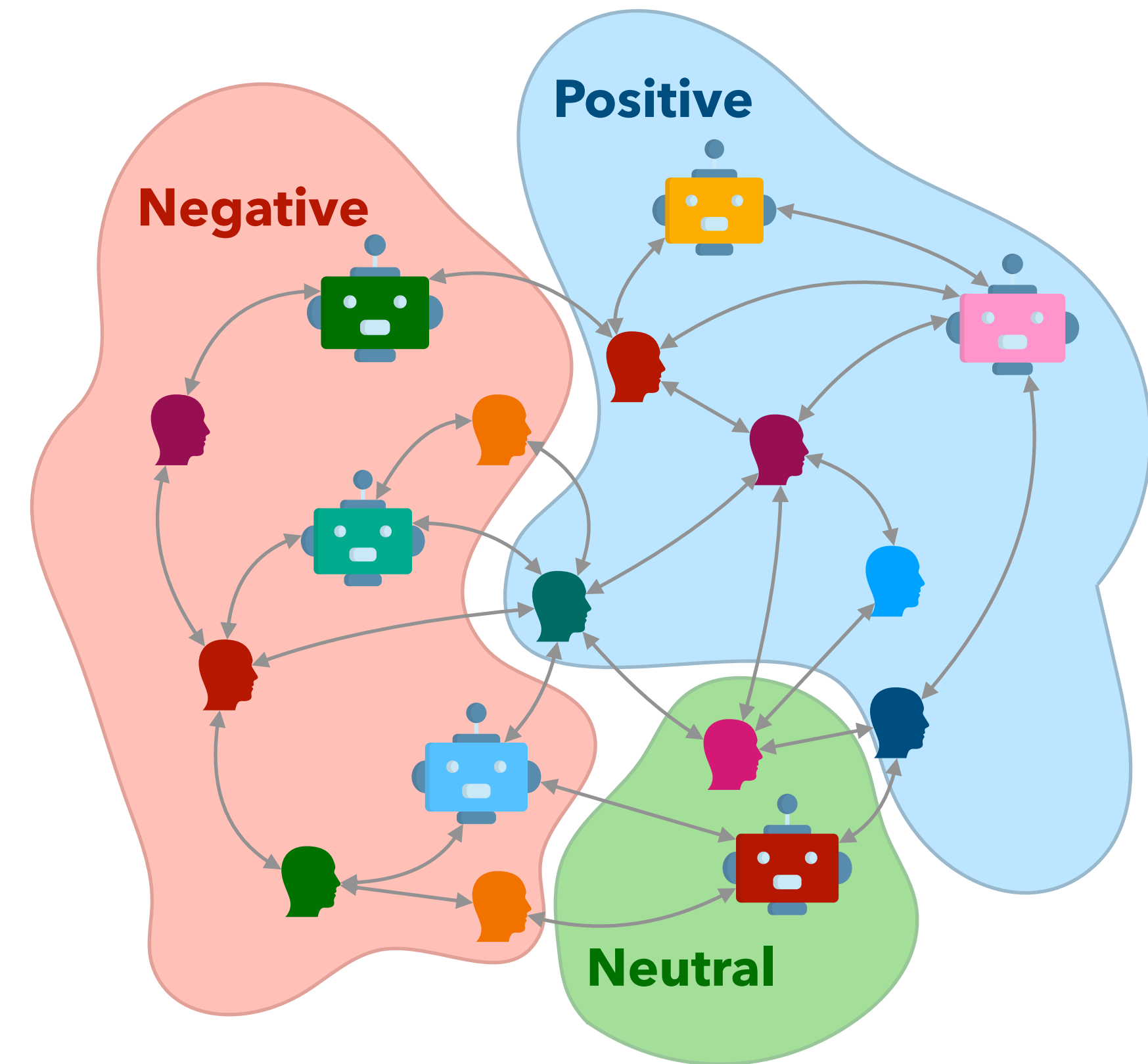
# Appendix – Quantifying Polarization

Neighbor Correlation Index  
(NCI, Cinus et al., 2022)

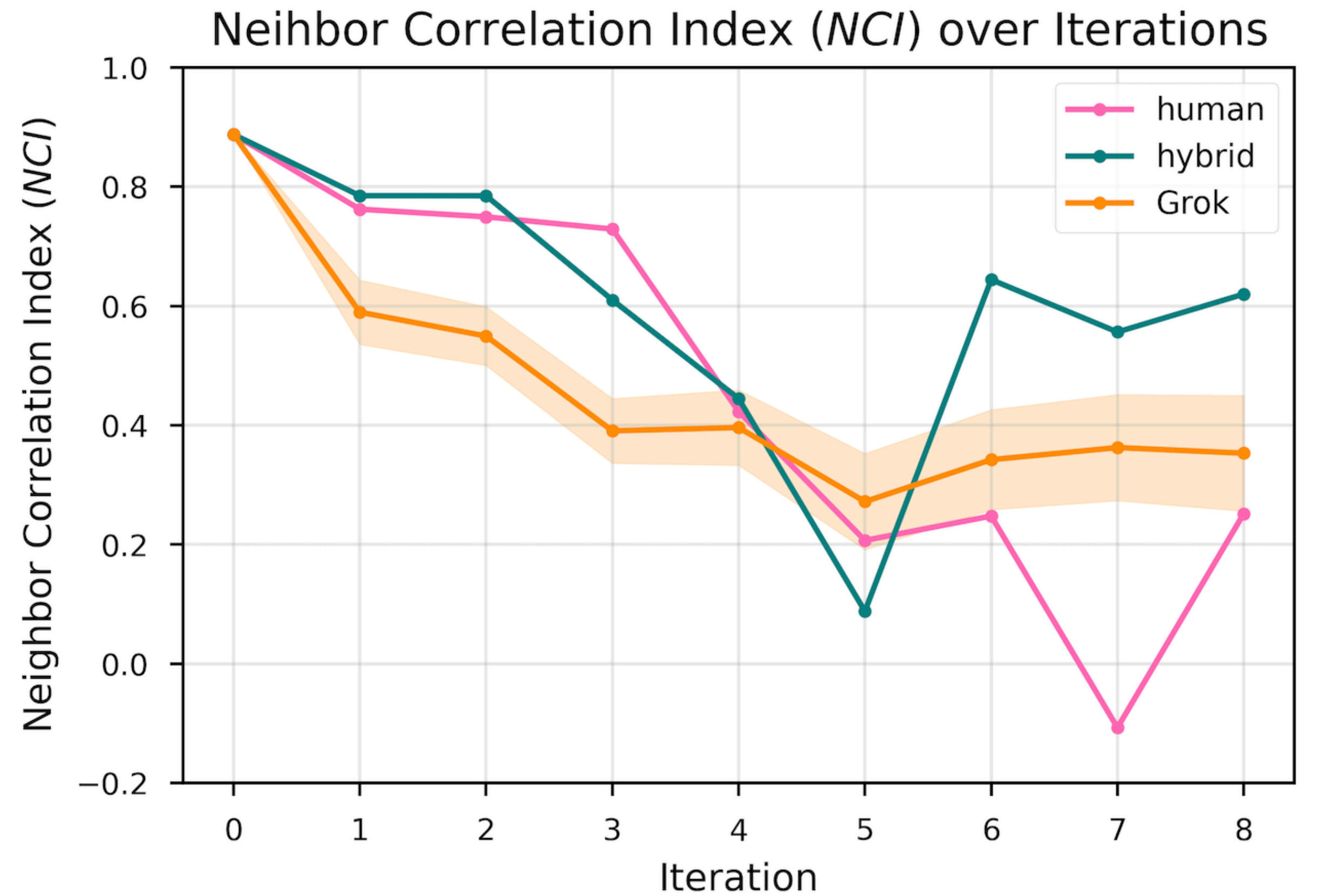
Measure of the similarity between  
a node and its neighbors.

$$\text{NCI} = \rho(\mathbf{z}, \mathbf{n})$$

$$n_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} z_j.$$



# Appendix – Quantifying Polarization





# Future projects

## Opinion diffusion

**Towards more robust and complex hybrid social network simulations:**

Exploring more topics, complex topologies, to study emerging collective hybrid behaviors.

**SHARD** (ASDTRID): applications on misinformation and fake news propagation.

## Test bed

**Emerging need for (safe) controlled simulation environments to benchmark AI models regarding their safety.**

Red-teaming, prompt-injection, malware detection, security breaches...

## Human-in-the-loop

**Internal use, industrial use-cases:**

Enhancing human/analysts reasoning via agentic/CoT threads. Human-AI collaboration, hybrid reasoning.

**NOESIS** (Pionniers de l'IA)

## Simulated hyper graphs benchmarks

**Issues with real world datasets:**

No control, no ground truth, subjective evaluation.

Emerging need for more complex datasets (connectivity, time...) in the TDL landscape.