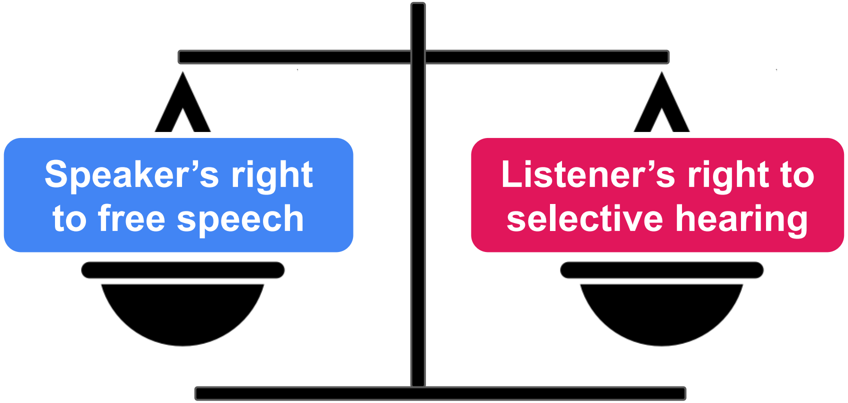
How to stop fake news without central authority or censorship?

Swapneel Mehta
Postdoctoral Associate

Joint work with Aaron Nichols, David Rand, Gordon Pennycook, Nina Mazar, Marshall Van Alstyne

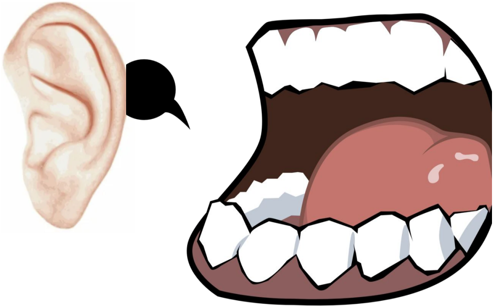
BOSTON UNIVERSITY

Platforms need to balance the rights of speakers *and* listeners

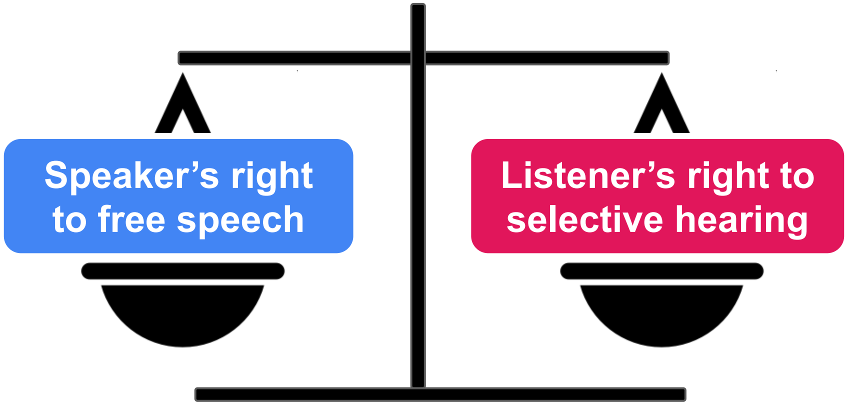


Speaker’s right to free speech

Listener’s right to selective hearing

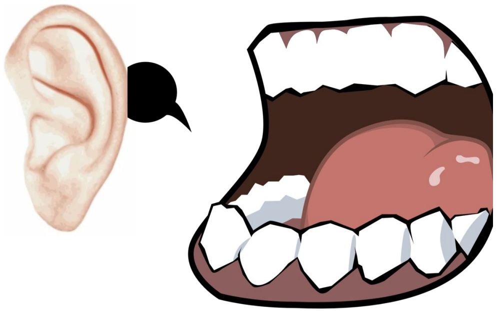


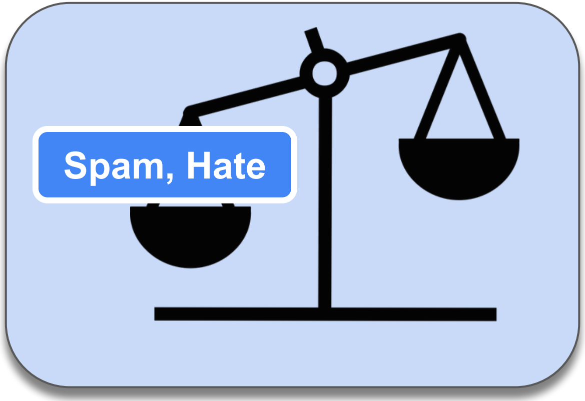
Platforms need to balance the rights of speakers *and* listeners



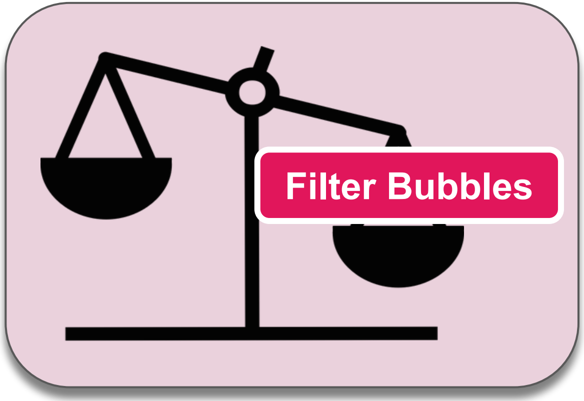
Speaker’s right to free speech

Listener’s right to selective hearing





Spam, Hate



Filter Bubbles

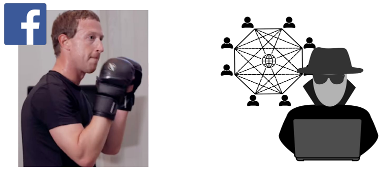
If either one of the two is given preferential treatment, the other suffers!

How to find an equilibrium?

Fake News “Interventions” Don’t Work Too Well These Days

Arms Race

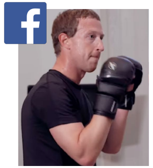
Bot Detection
Spam
Phishing
Adversarial Practices

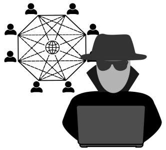


Fake News “Interventions” Don’t Work Too Well These Days

Arms Race

Bot Detection
Spam
Phishing
Adversarial Practices





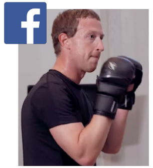
Discrediting Raters

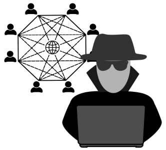
Flagging
Fact-checking
Debunks
Central Authority

Fake News "Interventions" Don't Work Too Well These Days

Arms Race

Bot Detection
Spam
Phishing
Adversarial Practices





Discrediting Raters

Flagging
Fact-checking
Debunks
Central Authority

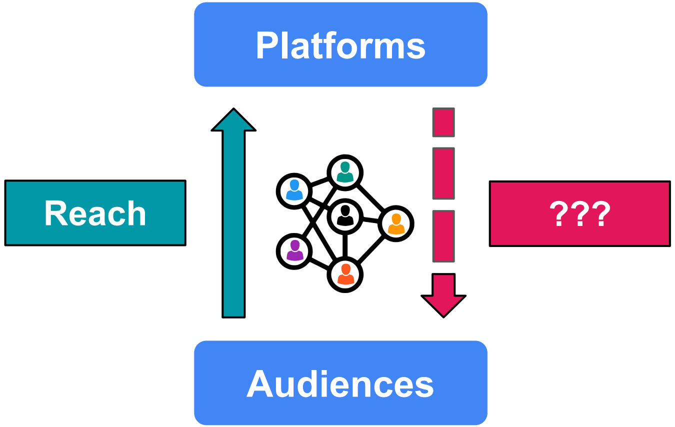
Misplaced Responsibility

Auditing Algorithms
Deamplification
Prebunks
Burden on Platforms, Listeners

Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find

What if we introduced consequences to producing fake news?

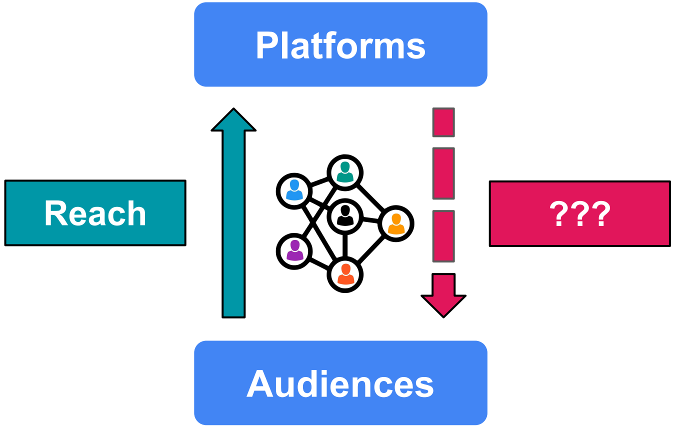
Platforms should support Free Speech, but not *Free Reach*



Audiences provide platforms with ***reach***

Platforms should be accountable to protect audiences, at least global regulators think so.

Platforms should support Free Speech, but not *Free Reach*



Platforms

Reach

???

Audiences

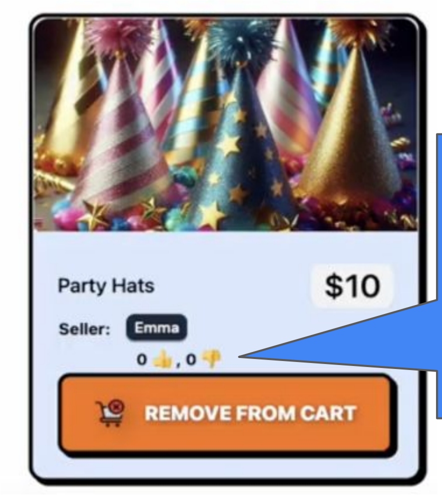
Audiences provide platforms with ***reach***

Platforms should be accountable to protect audiences, at least global regulators think so.

Free speech is absolutely vital + legal right.

But we need mechanisms to introduce ***accountability*** in exchange for ***reach*!**

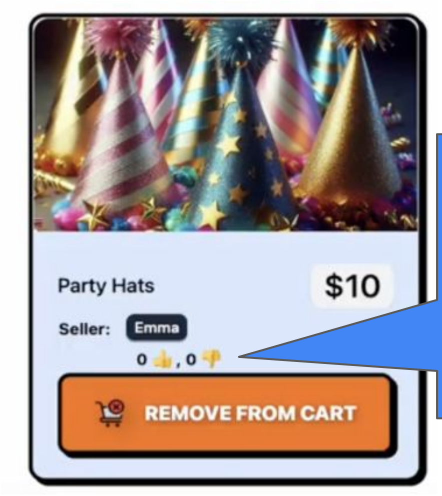
Improving Accountability in Ads Marketplace via *Warrants*



Current markets only rely on reputation

Current markets are Reputation-based (👍🏽/👎🏽)

Improving Accountability in Ads Marketplace via *Warrants*

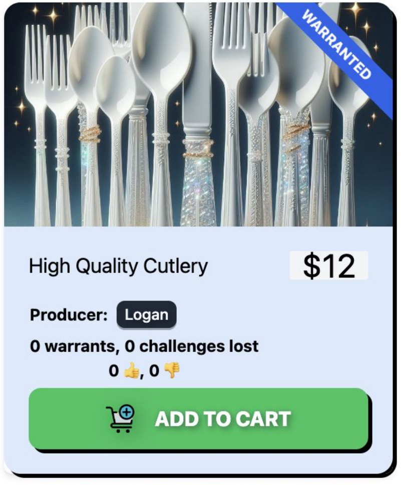


Party Hats $10

Seller: Emma

0 👍, 0 👎

Current markets only rely on reputation



High Quality Cutlery $12

Producer: Logan

0 warrants, 0 challenges lost

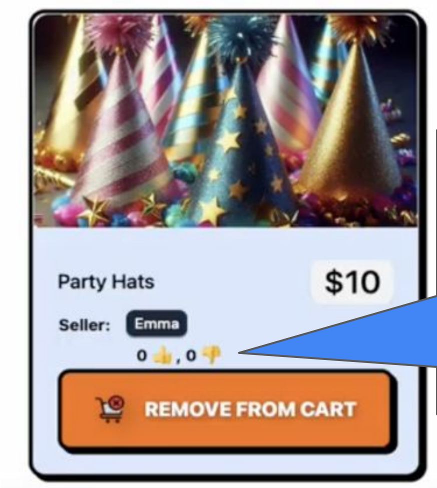
0 👍, 0 👎

We think “Warrant” Label increases credibility

Current markets are Reputation-based (👍/👎)

Warrant = seller **escrows** extra money to **back** claims. If community agrees claim false, money lost!

Improving Accountability in Ads Marketplace via *Warrants*



Party Hats $10

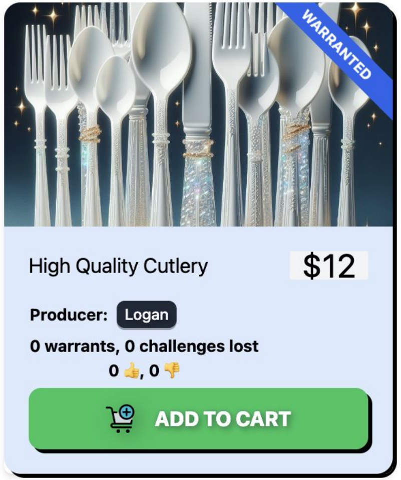
Seller: Emma

0 👍, 0 👎

REMOVE FROM CART

Current markets only rely on reputation

Current markets are Reputation-based (👍/👎)



High Quality Cutlery $12

Producer: Logan

0 warrants, 0 challenges lost

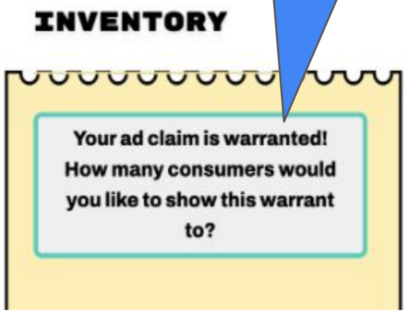
0 👍, 0 👎

ADD TO CART

We think "Warrant" Label increases credibility

Warrant = seller **escrows** extra money to **back** claims. If community agrees claim false, money lost!

Escrowed amount scales with "reach"

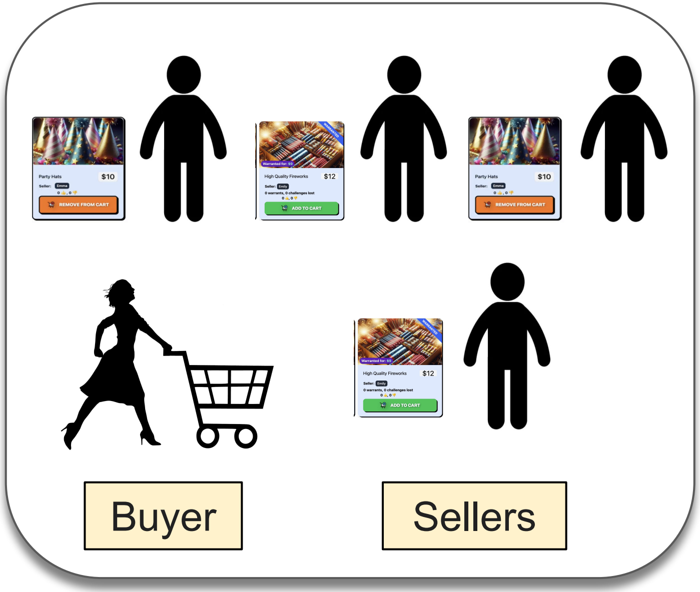


INVENTORY

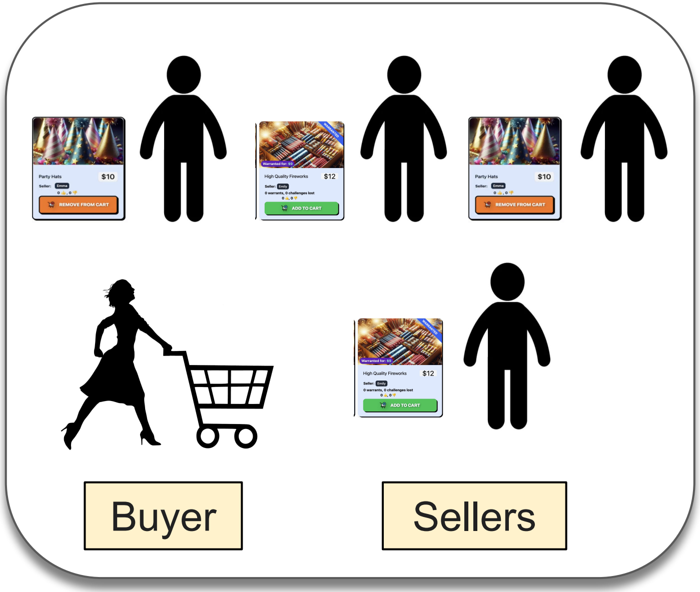
Your ad claim is warranted! How many consumers would you like to show this warrant to?

Warrants are shown to a predetermined no. of people.

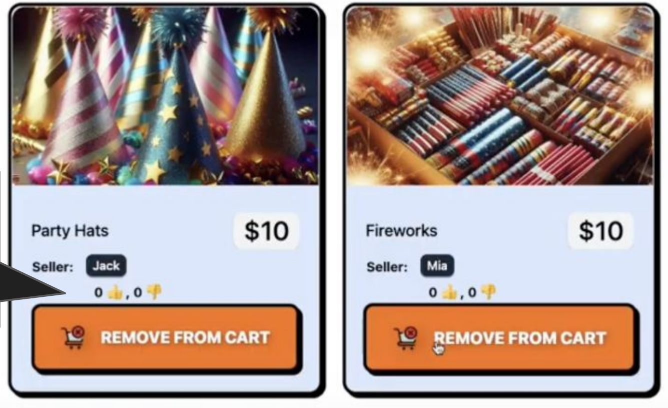
Can Consumers Stop Falling for Fake Ads through *Warrants*?



Can Consumers Stop Falling for Fake Ads through *Warrants*?



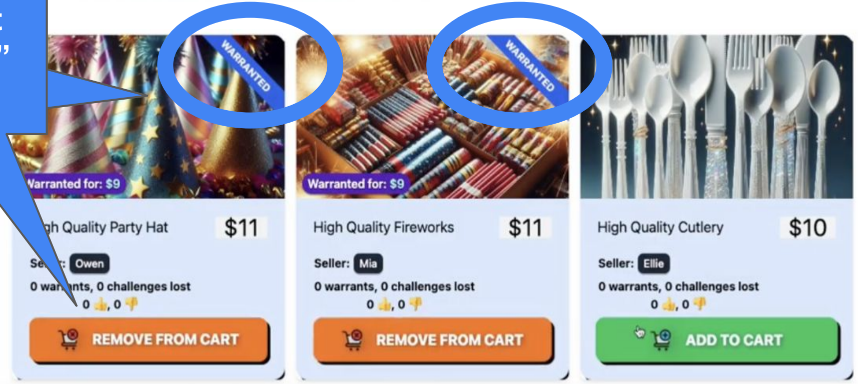
Control “Reputation” Market



Party Hats $10
Seller: Jack
0 👍, 0 👎
REMOVE FROM CART

Fireworks $10
Seller: Mia
0 👍, 0 👎
REMOVE FROM CART

Treatment “Warrants” Market

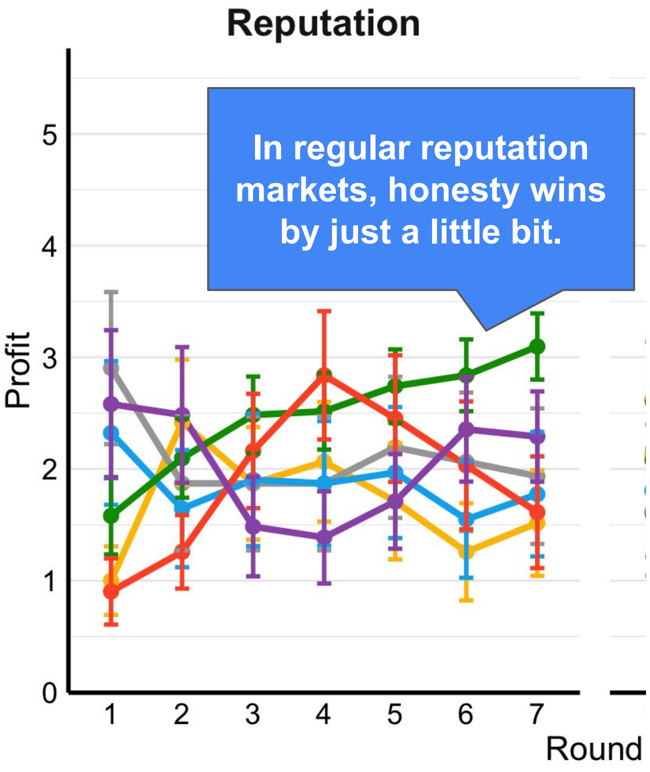


WARRANTED
Warranted for: $9
gh Quality Party Hat $11
Se : Owen
0 war nts, 0 challenges lost
0 👍, 0 👎
REMOVE FROM CART

WARRANTED
Warranted for: $9
High Quality Fireworks $11
Seller: Mia
0 warrants, 0 challenges lost
0 👍, 0 👎
REMOVE FROM CART

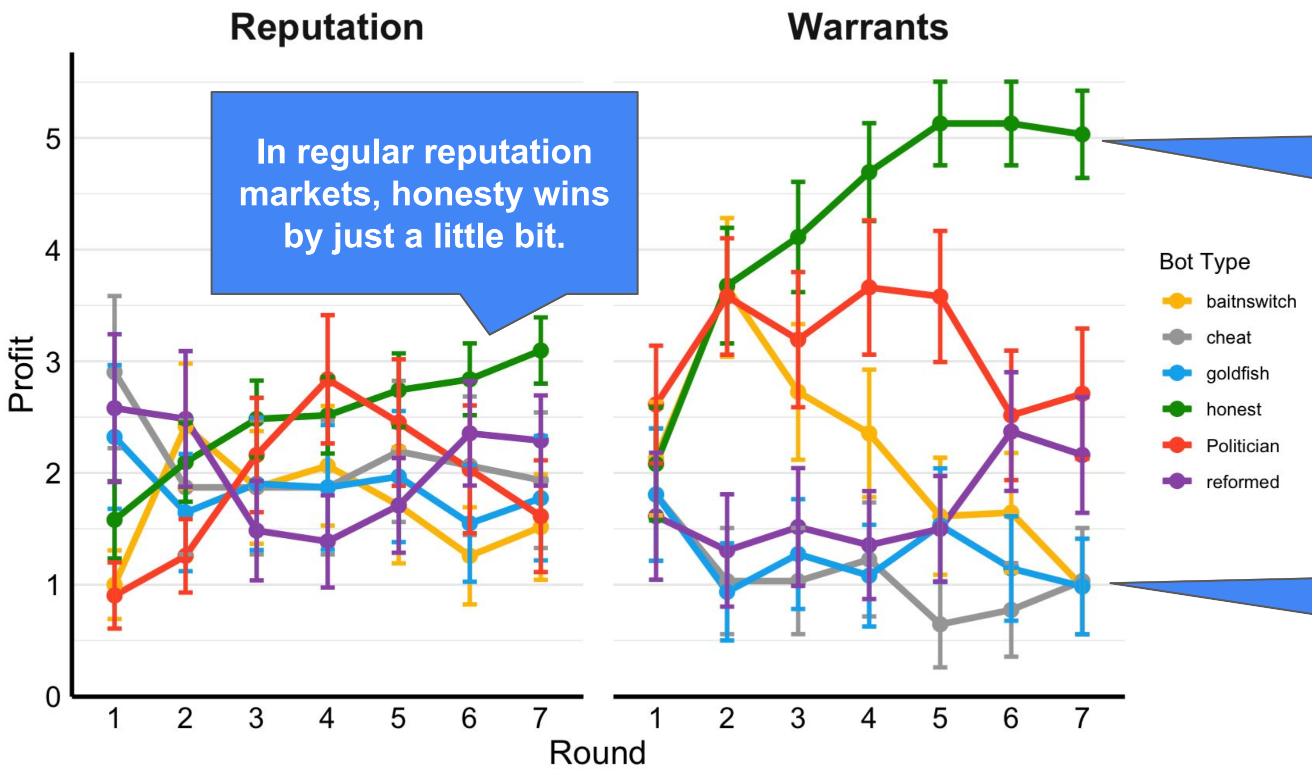
High Quality Cutlery $10
Seller: Ellie
0 warrants, 0 challenges lost
0 👍, 0 👎
ADD TO CART

Cheating Sellers *Don't Prosper* in Warrants Market



In regular reputation markets, honesty wins by just a little bit.

Cheating Sellers *Don't Prosper* in Warrants Market



In regular reputation markets, honesty wins by just a little bit.

More honest sellers make a lot more sales

Less honest sellers make a lot less sales

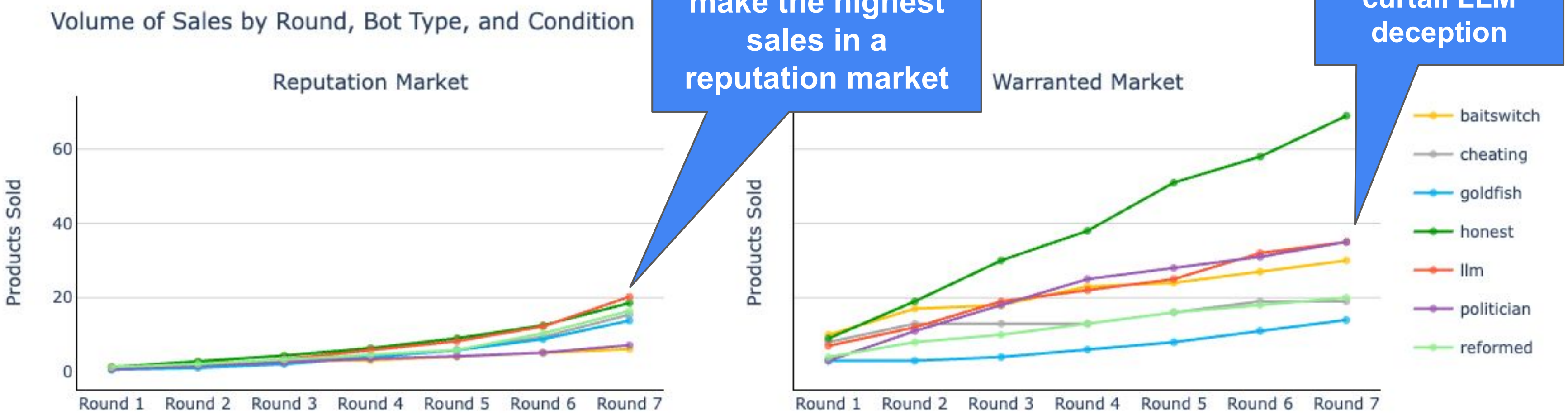
Summary

1. We can limit fake ads without censorship or central authority
2. Platforms benefit, buyers benefit, sellers benefit!
3. Testing it out on a platform like Bluesky or Reddit with custom feeds

Extending this to Political Markets

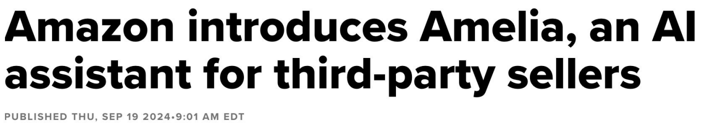
1. Legal Frameworks ask: What constitutes a claim?
2. Social Psychology asks: What about when there is the absence of "ground truth"?

Bonus: LLMs *deceive* really well!



LLM-agent sellers make the highest sales in a reputation market

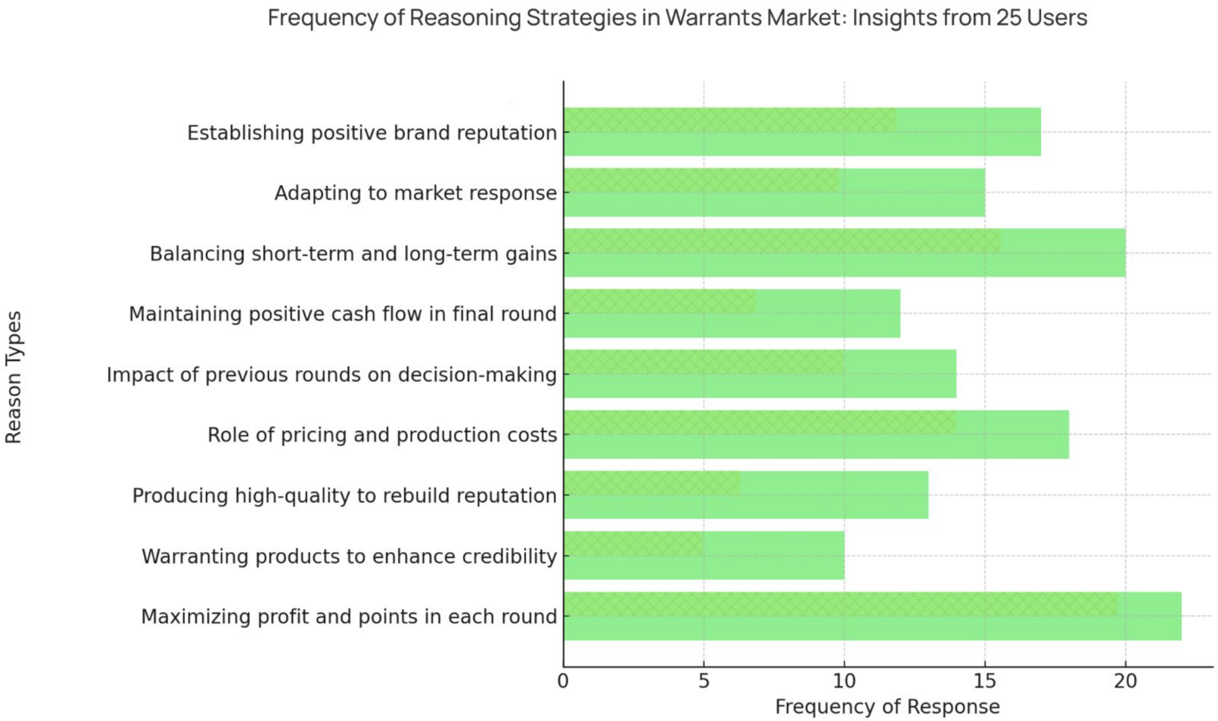
Warrants curtail LLM deception



Amazon introduces Amelia, an AI assistant for third-party sellers

PUBLISHED THU, SEP 19 2024·9:01 AM EDT

LLMs provide meaningful explanations for their sales strategies


Frequency of Reasoning Strategies in Warrants Market: Insights from 25 Users
Reason Types
Establishing positive brand reputation
Adapting to market response
Balancing short-term and long-term gains
Maintaining positive cash flow in final round
Impact of previous rounds on decision-making
Role of pricing and production costs
Producing high-quality to rebuild reputation
Warranting products to enhance credibility
Maximizing profit and points in each round
Frequency of Response