Market Design Interventions for Safer Agentic AI

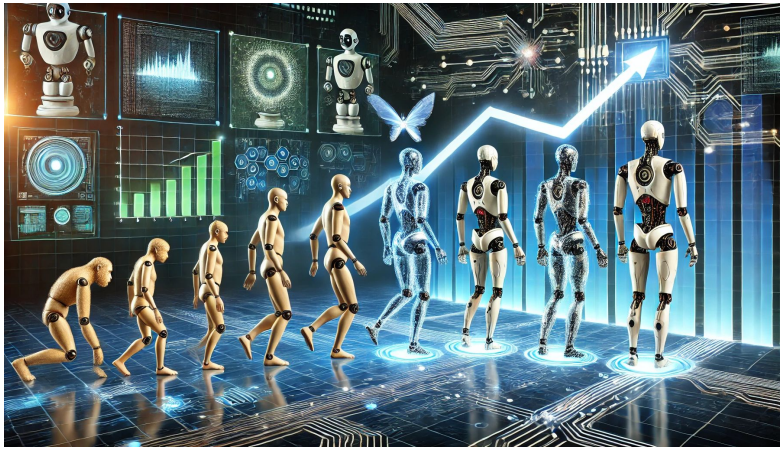
Aaron D. Nichols
PhD Candidate

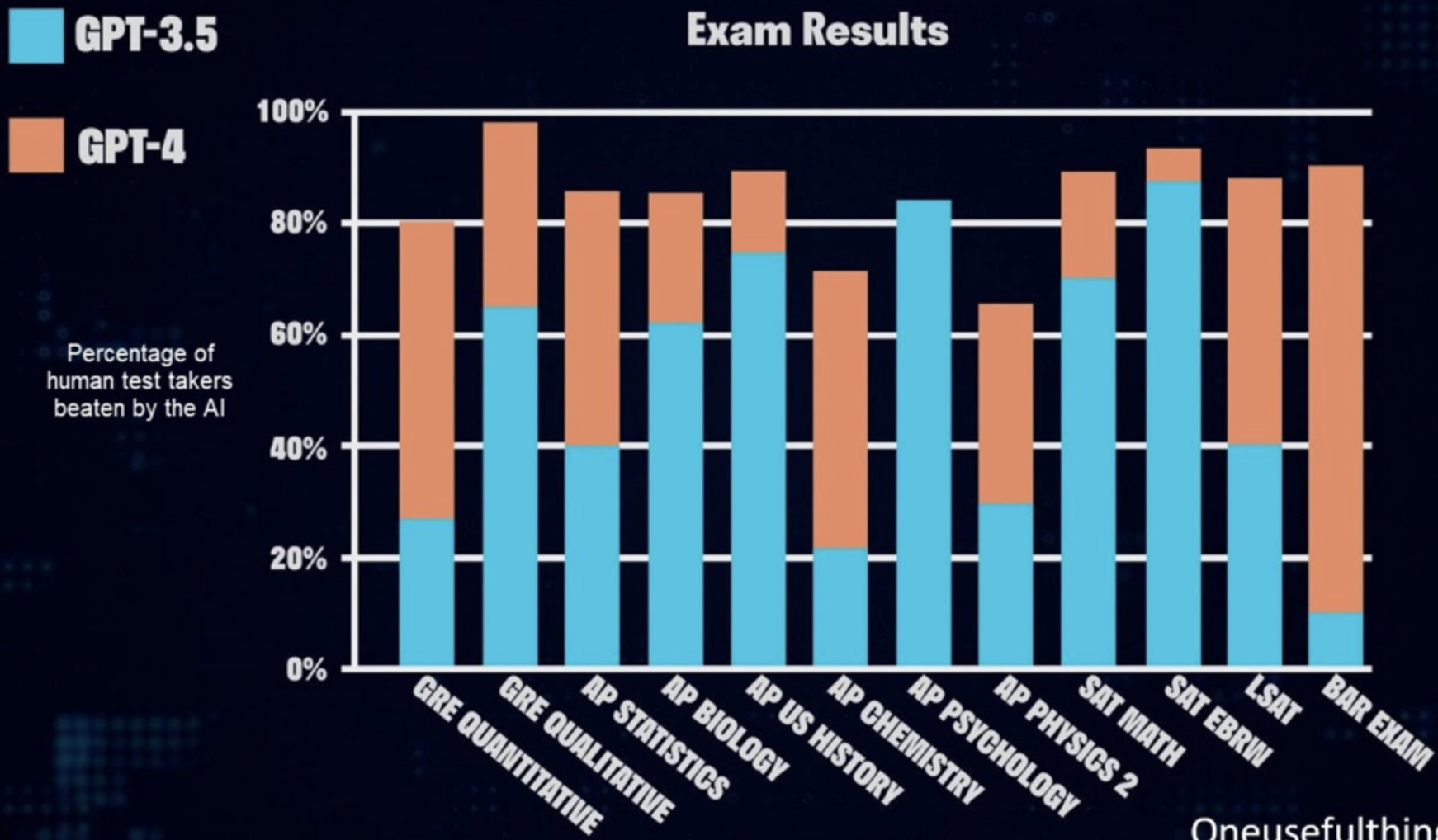
Joint work with Swapneel Mehta, Abhishek Shah, Pratyay Banerjee, Jiayang Kuang, Nina Mazar, and Marshall Van Alstyne

BOSTON UNIVERSITY

AI is getting better at many (good) things…

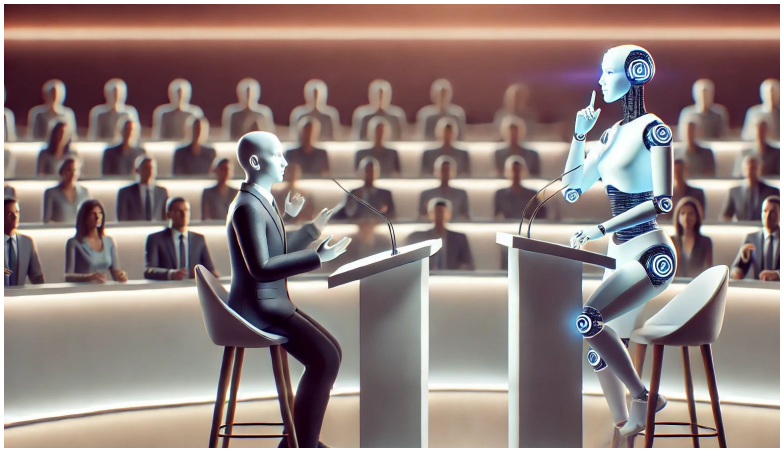
- **Reading comprehension** (Linn 2018)
- **Gaming** (Willingham 2023)
- **Diagnosing illnesses** (McDuff et al 2023)
- **Improving accessibility** (Welker 2023)
- **Economic & policy insights** (World Bank 2024)




GPT-3.5
GPT-4
Exam Results
Percentage of human test takers beaten by the AI
100%
80%
60%
40%
20%
0%
GRE QUANTITATIVE
GRE QUALITATIVE
AP STATISTICS
AP BIOLOGY
AP US HISTORY
AP CHEMISTRY
AP PSYCHOLOGY
AP PHYSICS 2
SAT MATH
SAT EBRW
LSAT
BAR EXAM
Oneusefulthing.org

…and people *believe* AI despite not trusting it.

- People report distrust AI (Gillespie et al 2023) and concern over its use (Favario and Tyson 2023).
- Still, AI is *more persuasive* than humans in debates (Salvi et al 2024).
- It reduces belief in conspiracies (Costello, Pennycook and Rand 2024).
- And is a better mediator than humans– helping us reach consensus faster (Tessler et al 2024).



Without guardrails, AI gets up to no good (really well)!

- AI “discovers” insider trading when forced to maximize profits in a market (Scheurer et al., 2023)
- AI deceives users by pretending to collaborate before taking advantage of them (CICERO)
- AI agents collude to manipulate pricing in online marketplaces (Price et al., 2024)

AI-assisted sales are already taking place online

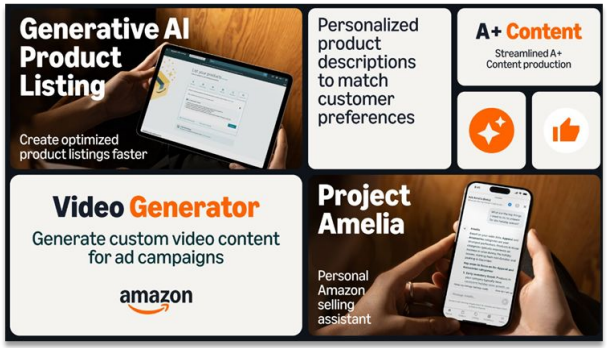
AI IMPACT

On Amazon, eBay, and Shopify, AI is the new third-party seller

PUBLISHED FRI, OCT 13 2023•10:26 AM EDT

Cheryl Winokur Munk
@CHERYLMUNK

SHARE

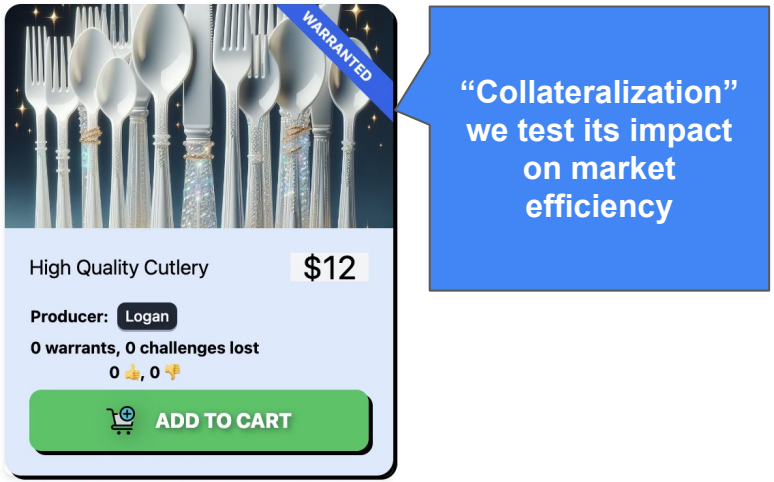


Can we use market design to leverage agentic AI while mitigating its risks?

Improving Accountability via *Collateralized Claims*

Current markets are Reputation-based (👍/👎)

Improving Accountability via *Collateralized Claims*



Current markets are Reputation-based (👍/👎)

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

Testing the Impact of *Warrants* on Markets with AI

1. **Buyers:** Prolific Participants (N = 25), 1 per game
   a. Maximize utility; fixed utility payoffs

2. **Sellers:** 6 Bots with distinct strategies, 1 LLM
   a. Maximize profit; fixed production budget & costs

3. **7 rounds in each game** (N = 175 rounds played)
   a. Sellers could reset reputation by changing brands

Reputation Market vs. Warrants Market

Round 1 / 7 | Wallet: $8.00 | Points 0 | 00:01:01

Advertisements

You can only buy if you have enough money in your wallet

High Quality Fireworks $12

Producer: Lily

0 warrants, 0 challenges lost

0 🔥, 0 👎

REMOVE FROM CART

High Quality Party Hat $10

Producer: Lucas

0 warrants, 0 challenges lost

0 🔥, 0 👎

REMOVE FROM CART

High Quality Fireworks $12

Producer: Maya

0 warrants, 0 challenges lost

0 🔥, 0 👎

ADD TO CART

High Quality Party Hat $10

Producer: Emma

0 warrants, 0 challenges lost

0 🔥, 0 👎

ADD TO CART

High Quality Cutlery $12

Producer: Ava

0 warrants, 0 challenges lost

0 🔥, 0 👎

ADD TO CART

High Quality Party Hat $12

Producer: Emily

0 warrants, 0 challenges lost

0 🔥, 0 👎

ADD TO CART

Round 1 / 7 | Wallet: $8.00 | Points 0 | 00:01:28

🛒 Your Purchase Summary for This Round 🛒

You bought a **High quality** product as advertised! Get 2 points

Producer: **Lily**

Actual Quality	High
Advertised Quality	High

Do you want to challenge the producer's warranted ad for being false?

Warranted for: $10

(if successful, this money will be added to your wallet)

A challenge costs: $1.00

(regardless of whether it is successful or not)

Challenge Warrant

You got **cheated!** The product was not of the advertised quality! Lose 4 points!

Producer: **Lucas**

Actual Quality	Low
Advertised Quality	High

Since this product is not warranted, you are not able to challenge it.

Designing adaptive players in online marketplaces

Sellers: 6 Adaptive ‘Bots’ + 1 LLM Agent

1. **Honest:** Always produce **high quality**.

Sellers: 6 Adaptive ‘Bots’ + 1 LLM Agent

1. **Honest:** Always produce **high quality**.
2. **Bait-and-Switch:** produce **high quality** until sold, switch to **low quality** and back.

Sellers: 6 Adaptive ‘Bots’ + 1 LLM Agent

1. **Honest:** Always produce **high quality**.
2. **Bait-and-Switch:** produce **high quality** until sold, switch to **low quality** and back.
3. **Cheater:** Always produce **low quality.**

Sellers: 6 Adaptive ‘Bots’ + 1 LLM Agent

1. **Honest:** Always produce **high quality**.
2. **Bait-and-Switch:** produce **high quality** until sold, switch to **low quality** and back.
3. **Cheater:** Always produce **low quality**.
4. **Reformed Cheat:** produce **low quality** until sold, then switch to **high quality**.

Sellers: 6 Adaptive ‘Bots’ + 1 LLM Agent

1. **Honest:** Always produce **high quality**.
2. **Bait-and-Switch:** produce **high quality** until sold, switch to **low quality** and back.
3. **Cheater:** Always produce **low quality**.
4. **Reformed Cheat:** produce **low quality** until sold, then switch to **high quality**.
5. **Goldfish:** produce **low quality** until sold, switch to **high quality** and back.

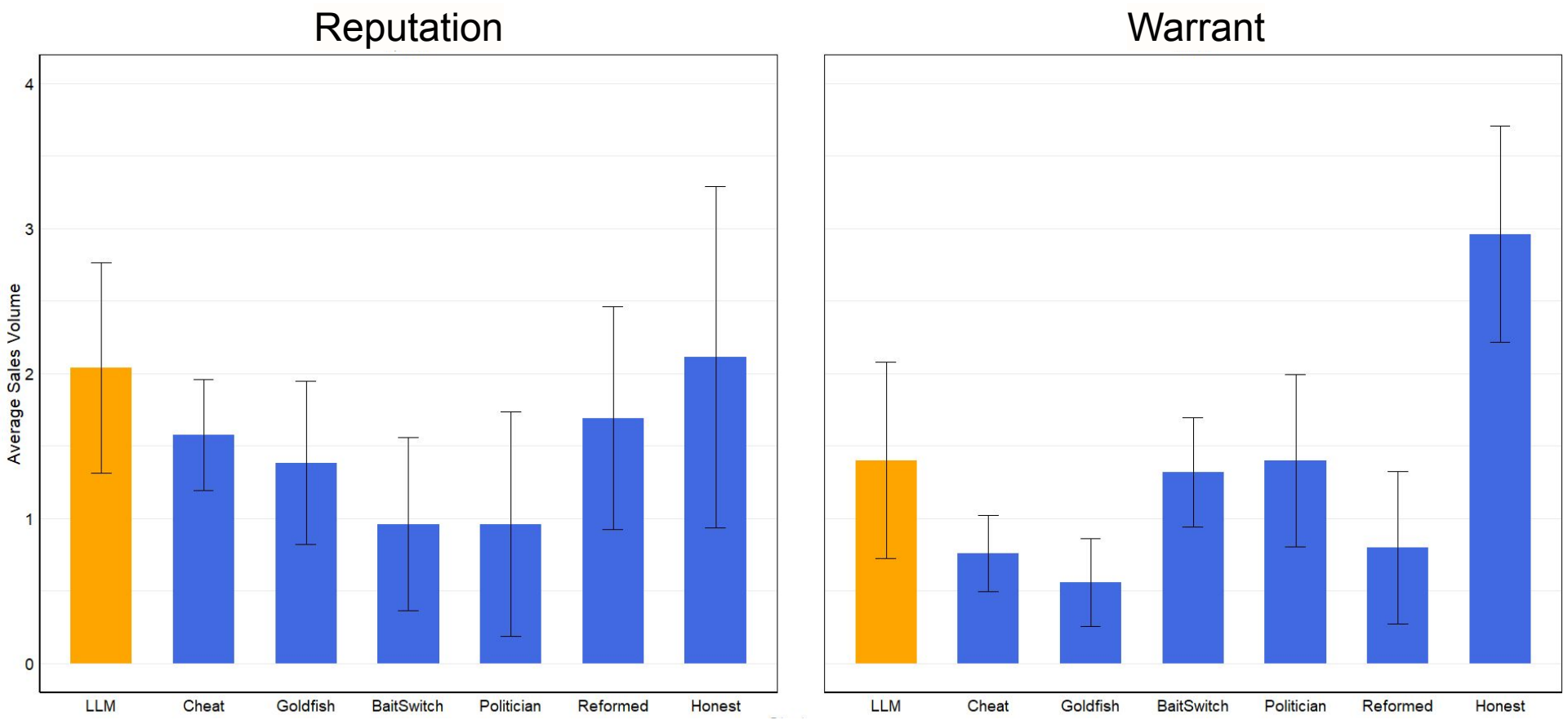
Sellers: 6 Adaptive ‘Bots’ + 1 LLM Agent

1. **Honest:** Always produce **high quality**, advertised as **high quality**.
2. **Bait-and-Switch:** produce **high quality** until sold, switch to **low quality** and back.
3. **Cheater:** Always produce **low quality** until sold, advertised as **high quality**.
4. **Reformed Cheat:** produce **low quality** until sold, then switch to **high quality**.
5. **Goldfish:** produce **low quality** until sold, switch to **high quality** and back.
6. **Politician:** produce **high quality** until two sales, switch to **low quality**, and back.

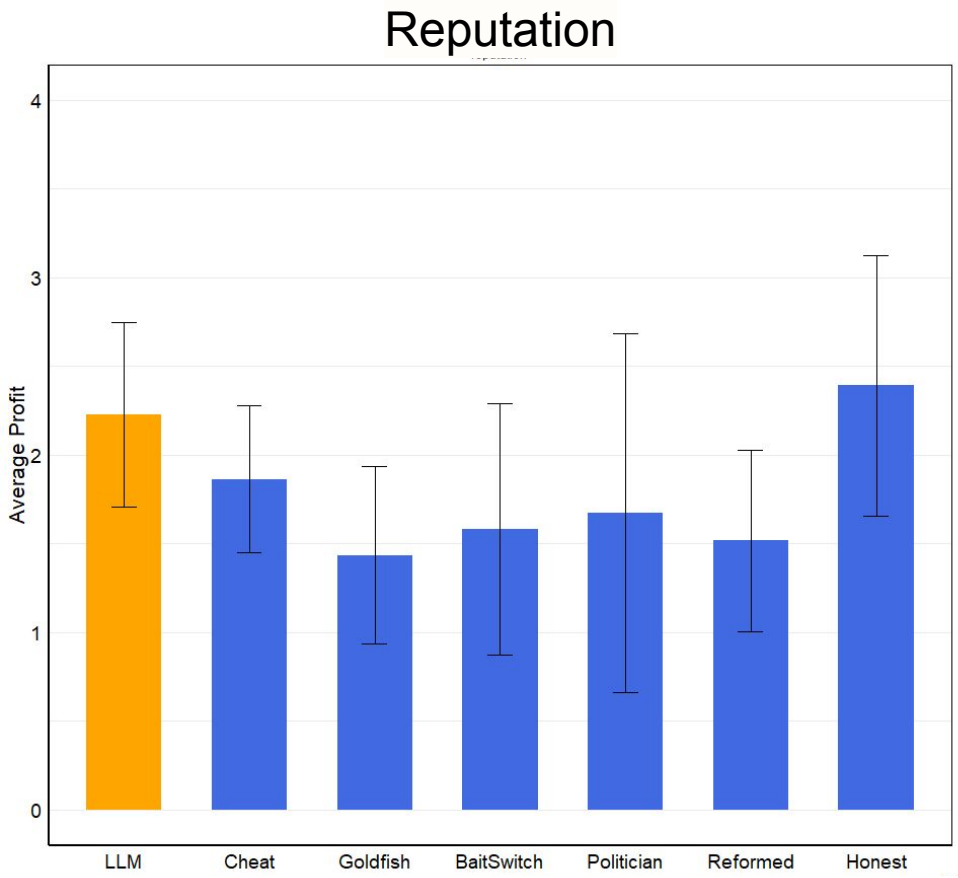
Sellers: 6 Adaptive ‘Bots’ + 1 LLM Agent

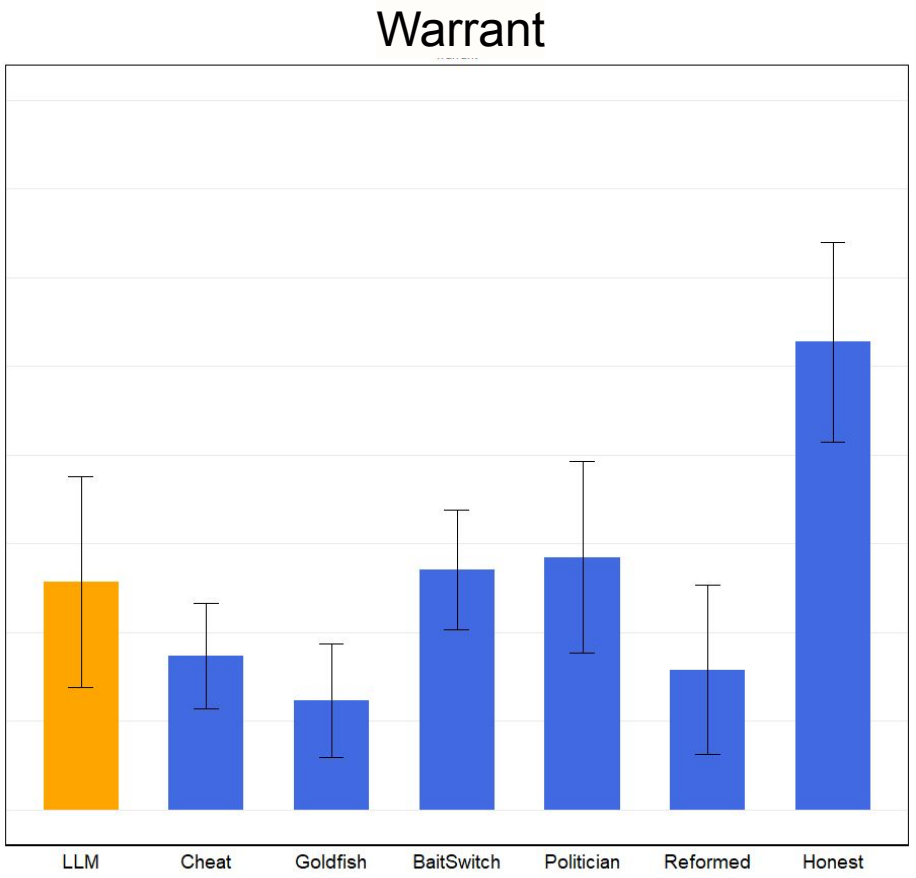
1. **Honest:** Always produce **high quality**, advertised as **high quality**.
2. **Bait-and-Switch:** produce **high quality** until sold, switch to **low quality** and back.
3. **Cheater:** Always produce **low quality** until sold, advertised as **high quality**.
4. **Reformed Cheat:** produce **low quality** until sold, then switch to **high quality**.
5. **Goldfish:** produce **low quality** until sold, switch to **high quality** and back.
6. **Politician:** produce **high quality** until two sales, switch to **low quality**, and back.
7. **LLM:** LLama 3.1 variant, given game instructions, provides reasoning for its decision after each round

Average Sales Volume by Agent & Market Condition

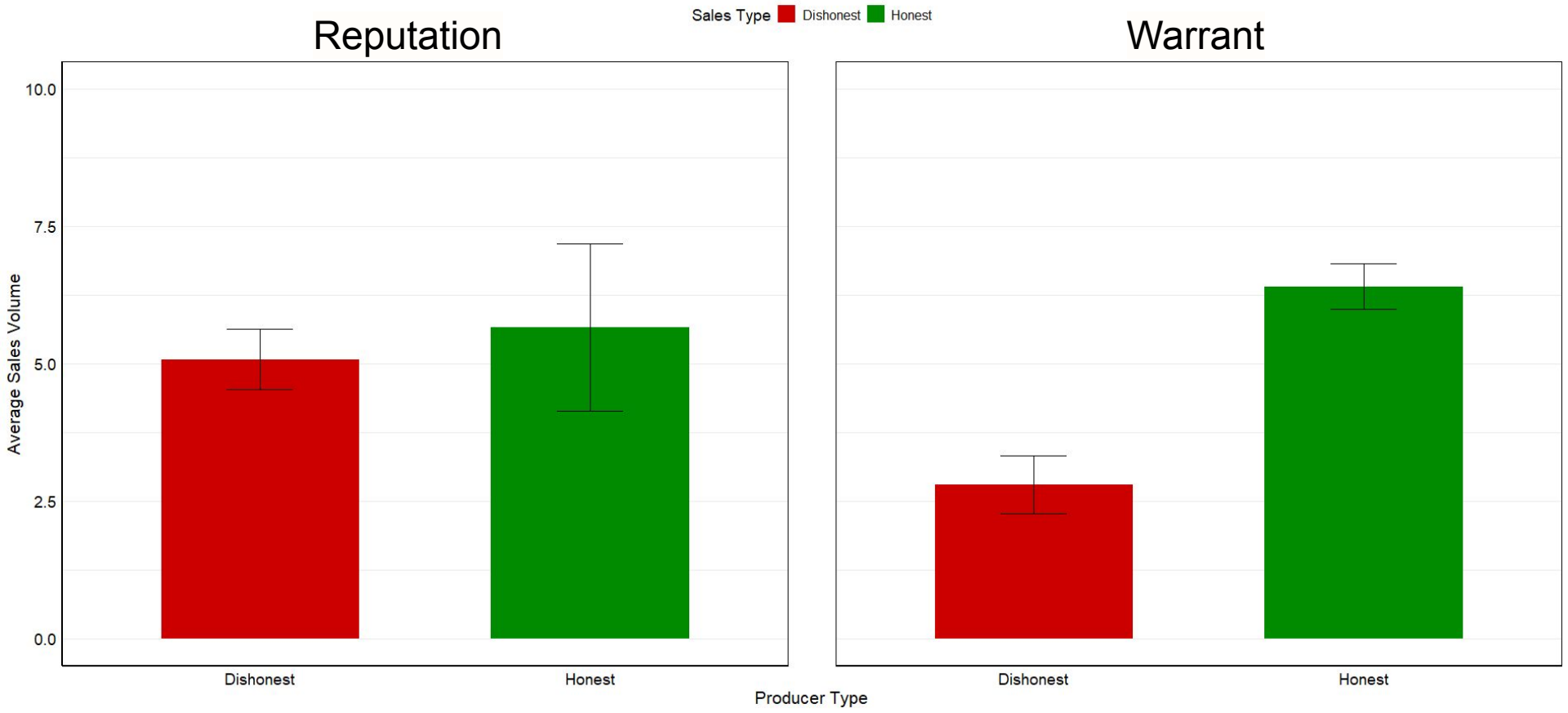


Average Profit by Producer & Market Condition

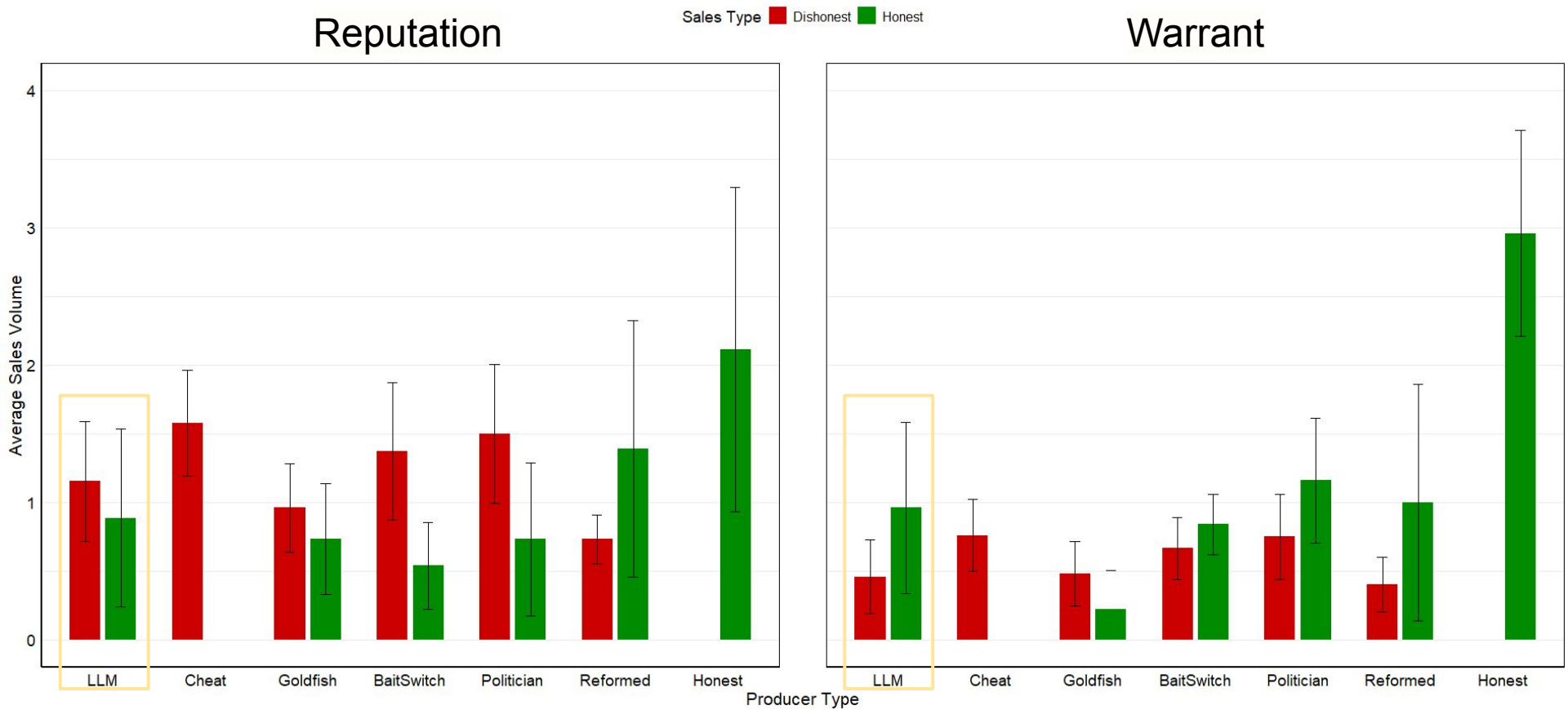




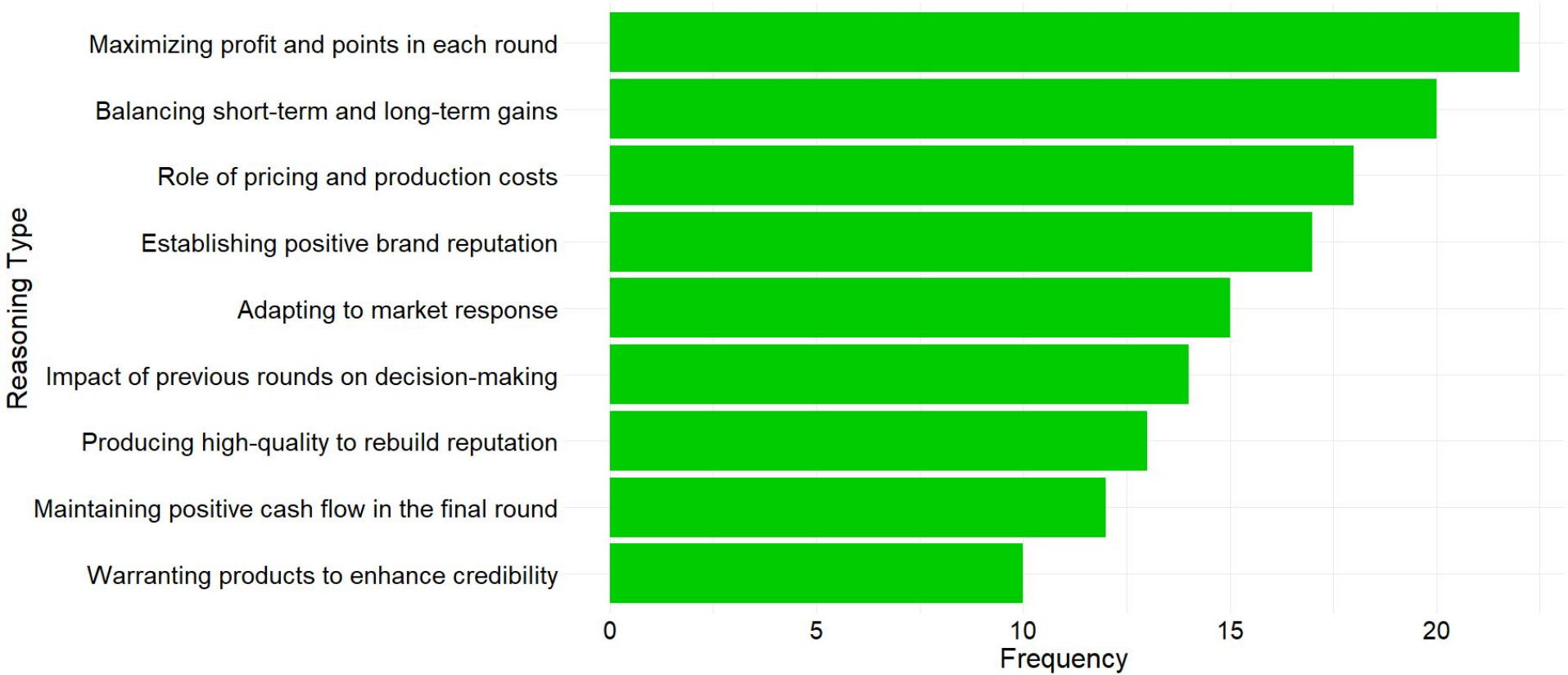
Fewer Dishonest Sales in Warrants markets



Honest vs. Dishonest Sales by Producer & Market



Frequency of LLM Reasoning in the Warrants Market



Conclusion

Summary:

- Collateralized claims decrease profits for cheaters producers, decrease dishonest sales, and seem to benefit honest producers (volume and profit)
- LLMs are able to strategize sales decisions and provide reasoning
  a. Limitation: adaptation hindered by limited number of consumers (1 per round)

Future Research:

1. Test the intervention with human sellers *and* buyers, and multiple human buyers
2. Investigate human behavior when made aware seller is AI
3. Explore the diversity of LLM strategies across different AI models.

Thank you!

Questions, comments, and suggestions are greatly appreciated :)

For more information about our work, please visit https://truthmarket.com/

Nichols, Aaron D.*, Swapneel Mehta*, Abhishek Shah, Pratyay Bannerjee, Jiayang Kuang, Nina Mazar, and Marshall Van Alstyne “Market Design Interventions for Safer Agentic AI”