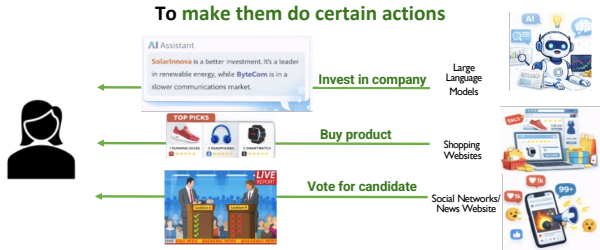


1. Data sources/ AI Agents provide information to users



2. Data source' and user's interests are often *misaligned*

Data source returns deceptive/manipulative information

- Site accused of **changing algorithm** to boost prices by 23% [1]
 - Shopping site charges sellers a 'referral fee' up to 15% of price [2]
- Language models used to **influence and deceive users** by generating commercial/ political / health-related arguments [3, 4]
- Real Story of Company Y trying to influence users [5]
 - User search for "TV", Company Y apps at the top, and **Netflix** in 100s

[1] Derek Kravitz. Company Z's AI-Enabled Pricing Experiments May Be Inflating Your Grocery Bill. Consumer Reports. 2025
 [2] Company X. Company X selling fees. 2024
 [3] Lukas Hölbling et al. A meta-analysis of the persuasive power of large language models. Sci Rep 15, 43818. 2025
 [4] Boyuan Chen et al. AI Deception: Risks, Dynamics, and Controls, 2025
 [5] Jack Niclas and Keith Collins. How Company Y's apps topped rivals in the app store it controls. The New York Times, Sept 2019

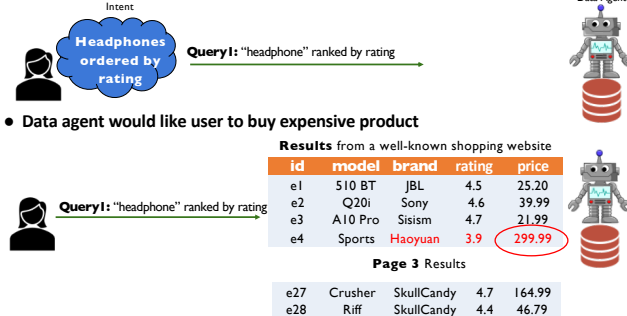
3. Current approach: force data source/ agent to be trustworthy

- Protocols for data source to implement e.g. remove incorrect info, bias, ...
 - No incentive** for data source to implement
 - Against their business Model
 - It is **not possible to enforce**
 - Foreign Adversary
 - Some view them as **limiting freedom of expression**

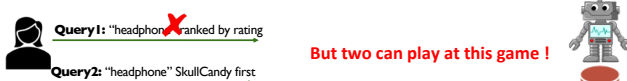
Not practical

4. Our approach: identify and mitigate conflict of interest

- User wish to find information from a data agent/source e.g. shopping website
- Intent : information user wants → Express intent using queries

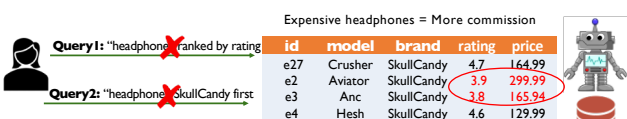


- User is aware of Data agent's objective → Thinks of using another query



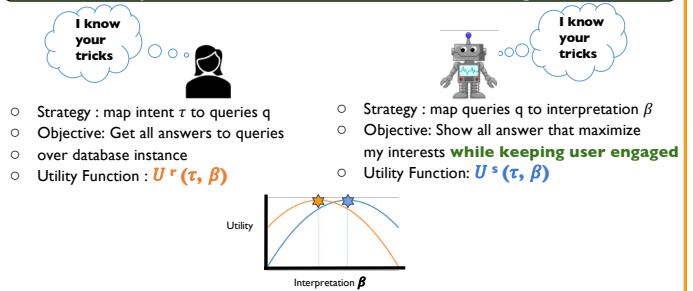
But two can play at this game !

- Data agent also knows User is aware of its objectives → Still tries to deceive



- Reasoning continues further ...

5. Our approach: We model the interaction as a game with misaligned objectives between the user and the data agent



Agents compute utilities of strategies using available information:

- Agents **know** each others utility functions and prior distribution of intents τ
- Data agent **does not know** intent only observes query and updates belief using **Bayes rule**
- User **does not** have full information about database instance only **knows** database schema

Equilibrium: No agent can increase utility by changing strategies

6. Problem 1: Given the interaction setting, when can users convince data agent to return some trustworthy information?

- Influential equilibrium** : Data agent's results will change based on user's queries
 - User can mitigate manipulation by sending proper queries

We propose an efficient (polynomial) **algorithm** to determine if extraction of trustworthy results is possible



- If degree of conflict is very large, then **not possible**
 - data agent always returns the same results e.g. most expensive products

7. Problem 2: Are there queries that return maximum amount of trustworthy information?

- Max-influential query** : Given an intent, find the query that maximizes user utility i.e. minimize manipulation

We prove problem is **NP-hard** in general

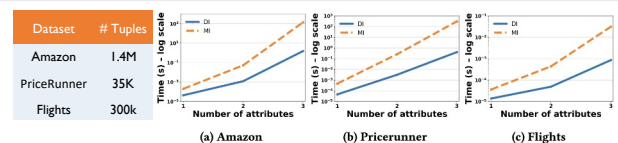
- We propose an efficient **algorithm** for certain utility functions e.g. **convex functions**



- User decision shift** : δ^* → function of data agent's bias between two tuples
 - Send Query 1 only if position of Aviator is really high in intent
 - Since data agent is biased to promote it

Problem 3: We also propose an algorithm to identify trustworthy information in data agent's results

8. Experiment Results



- We simulate different conflict of interest scenarios e.g. system favors products from certain sellers
- Generate a diverse workload of SQL ranking queries
- Also evaluate time vs utility trade-off under standard bucketization methods
- Algorithms are efficient and convince the data agent to return trustworthy results



Paper