Transparency and Accountability of Explanations for Algorithmic Systems

Krishna P. Gummadi

Max Planck Institute for Software Systems

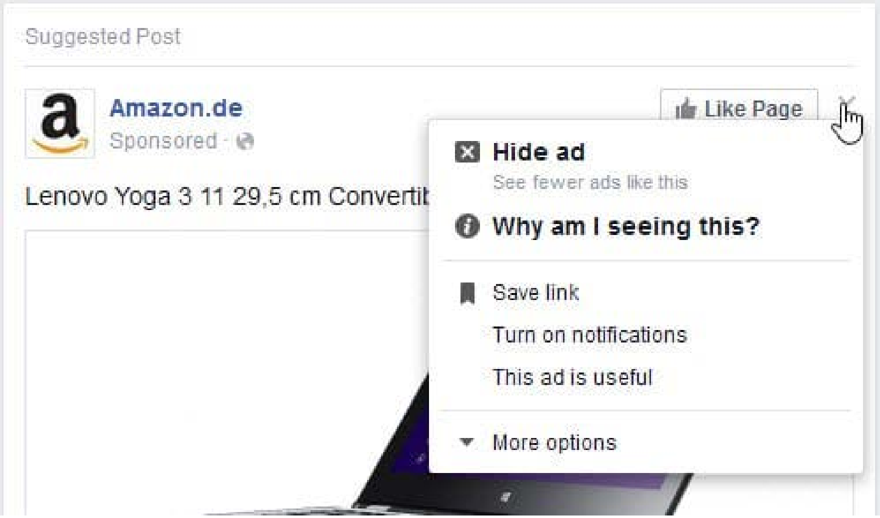
Bringing transparency to web services

- Traditional perspective: Adversarial service provider
  - Need to reverse-engineer black-box web services

  - Analyze inputs & outputs, to learn how the black-box works
    - Classic datamining / learning problem

Transparency through explanations

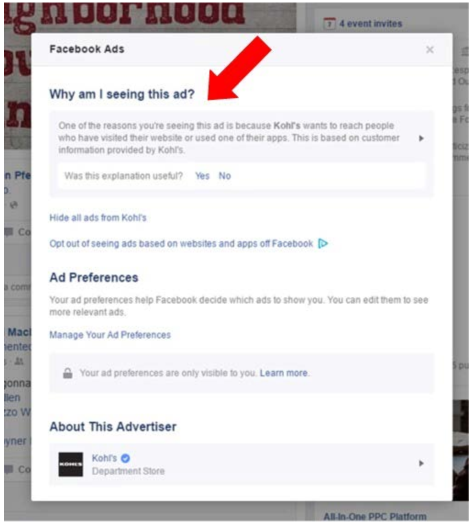
- Provided by service operators themselves



Suggested Post

Amazon.de
Sponsored

Lenovo Yoga 3 11 29,5 cm Convertib



Facebook Ads

Why am I seeing this ad?

One of the reasons you're seeing this ad is because **Kohl's** wants to reach people who have visited their website or used one of their apps. This is based on customer information provided by Kohl's.

Was this explanation useful? Yes No

Hide all ads from Kohl's

Opt out of seeing ads based on websites and apps off Facebook

Ad Preferences

Your ad preferences help Facebook decide which ads to show you. You can edit them to see more relevant ads.

Manage Your Ad Preferences

Your ad preferences are only visible to you. Learn more.

About This Advertiser

Kohl's
Department Store

- Voluntary explanations
  - To enhance user trust & cooperation
- Or required by law – right to explanation

We need standards for explanations

Need to guard against adversarial explanations:

- Insufficient / unsatisfactory explanations
  - That offer no insightful / actionable information to consumers
- Misleading / fake explanations:
  - Designed to influence consumers to behave a certain way
  - Designed to gain consumer acceptance for a service

A case study: Facebook ads

- Facebook gathers lots of data (features) on users
  - Demographical
    - Relationship:
      - Interested In: Men and Women, Men, Unspecified, Women
      - Status: Separated, Widowed, Open Relationship, Divorced, In a relationship, Married, Engaged, Unspecified, Single, Complicated Civil Union, Domestic Partnership
  - Behavioral
  - Interests

- Each user feature is a boolean variable

Background: Facebook ad targeting

- To target users, advertisers specify a boolean formula over the features

- Typically, in a restricted CNF form
  - (F1 v F2 v F3….) ^ (F’1 v F’2 v F’3….) ^ ….. ^ -FK ^ -F’K

- Users are targeted, when their feature values inferred by Facebook satisfy the targeting formula

- Most formulas tend to specify location, gender, age

Explanations provided by FB

- Beyond location, gender, age: picks exactly one of the several features used in targeting formula

  - *“One reason you're seeing this ad is that Peek & Cloppenburg wants to reach people interested in Shopping and fashion, based on activity such as liking Pages or clicking on ads.”*

  - *“There may be other reasons why you're seeing this advert, including that Acer wants to reach people aged 18 to 45 who live or have recently been in Germany. This is information based on your Facebook profile and where you've connected to the Internet.”*

Are the explained features…

- Complete?
- Useful?
  - Necessary? Sufficient? Most important?
- Correct?
- Personalized?
- Deterministic?

Vague explanations: Example

- Explanation to consumers:
  - *“One reason you're seeing this ad is that Peugeot wants to reach people who are part of an audience created based on data provided by Acxiom. Facebook works with data providers to help businesses find the right audiences for their ads. Learn more about data providers.”*

- Information provided to advertisers:
  - **Demographics** > Financial > Income > Geschätztes monatliches Nettoeinkommen 2.600 bis 3.600 EURO
  - **Description:** Dieser Haushalt hat wahrscheinlich ein monatliches Nettoeinkommen von 2.600 bis 3.600 EURO.
  - **Source:** Partner Category provided by Acxiom….

Summary

- Lots of focus on how to explain algorithmic systems
  - But, why should we trust explanations?

- Case study of Facebook targeted ad explanations
  - Not clear what properties they satisfy

- Need to have standards for explanations
  - Constructing satisfactory explanations is non-trivial!