Building User Trust in Recommendations via Fairness and Explanations

Keynote at HAAPIE 2020 in conjunction with ACM UMAP 2020

Dimitris Sacharidis
TU Wien
dimitris@ec.tuwien.ac.at

Trust in AI Systems

- COMPAS estimates the likelihood of a criminal to **reoffend**
  - used by **judges** in the US to guide their decisions

- **can we trust such a system?**

- ProPublica analyzed COMPAS predictions and found **bias against blacks:**
  - blacks are almost **twice as likely** as whites to be labeled a **higher risk** but not actually reoffend
  - whites are **much more likely** than blacks to be labeled lower risk, but actually reoffend

How do people feel about AI systems?

- survey of 5,000 consumers by Pegasystems (34% say they interact with AI systems)
- AI not trustworthy
  - only 9% very comfortable interacting with AI
- AI is biased
  - 53% say it's possible for AI to show bias in its decisions
- AI cannot utilize morality
  - 56% don't believe it is possible to develop AI that behaves morally

How comfortable are you/would you be with a business using Artificial Intelligence to interact with you?

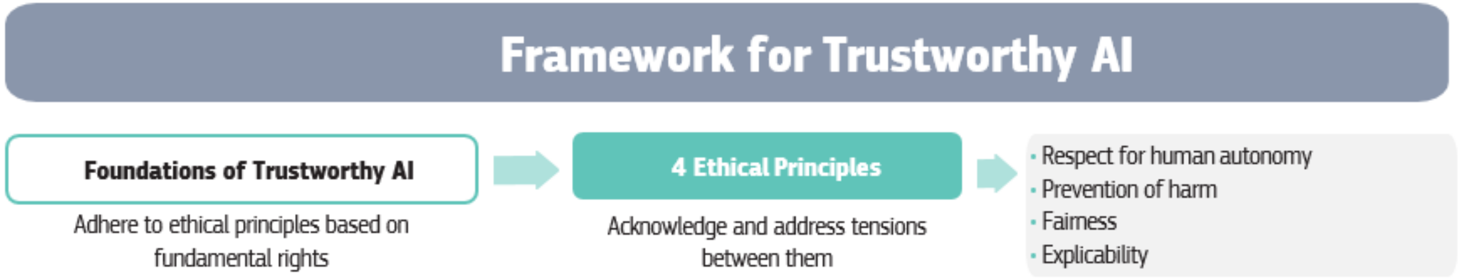
Ethical principles in AI systems

- review of 84 AI ethics guidelines [2019 Nature Machine Intelligence, A. Jobin et al.]
- transparency + fairness are the two most popular and important principles
- trust is the "end goal"

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice & fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom & autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

[2019 Nature Machine Intelligence A. Jobin et al.] *Artificial Intelligence: the global landscape of ethics guidelines*

Ethical principles in AI systems



- 4 ethical principles
  - “Ethics guidelines for trustworthy AI” [2019 European Commission]

- Respect for human autonomy
  - “leave meaningful opportunity for human choice”
- Prevention of harm
  - “protect human dignity as well as mental and physical integrity“
- Fairness
  - “ensure equal and just distribution of both benefits and costs”
- Explicability
  - “decisions explainable to those directly and indirectly affected”

Agenda

- Fairness
- Explanations
- Explanations + Fairness

Fairness in Recommendations

What is Fairness?

- fairness is a highly overloaded term
  - it can mean different things to different people in different contexts

- Dictionary: "***the state of being free from bias or injustice***"

- Political Science: "***distributive justice discusses fair allocation of resources among diverse members of a community***"
  - "A Theory of Justice" by J. Rawls (American philosopher)
    - "justice as fairness"; "social cooperation should be fair to all citizens regarded as free and as equals"
  - but what is a *fair allocation*?
    - *equality of outcome*: each person gets the same amount
    - *equality of opportunity*: equal grounds for competing for resources
    - *social welfare*: what benefits the society the most

What is Fairness?

- Legal Systems: ***"fairness as non-discrimination"***
  - *disparate treatment*: intentional discrimination on *protected groups* (defined on race, color, gender, etc.); not "*color-blind*"
    - e.g., only black applicants are required to take a pre-employment assessment test
  - *disparate impact*: a procedure that has disproportionate impact on protected groups
    - e.g., all applicants are tested but only blacks are eliminated based on the results of the assessment.
  - *affirmative action*: promote non-discrimination and support historically disadvantaged groups; *quota systems*
    - e.g., to address gender imbalance in STEM

What is Fairness?

- many definitions even in the context of AI/ML
  - separate notions in **classification**, **ranking**, **recommendation**
  - e.g., 21 fairness definitions by [2018 FAT* A. Narayanan]

- one abstract definition to rule them all
- **fairness** is the absense of **harmful discrimination** (or **bias**)

- sidenote: not all forms of discrimination and bias are harmful
  - recommenders earn their living by personalization (=discrimination)
  - perhaps **differentiation** is a better term for non-harming discrimination/bias

[2018 FAT* A. Narayanan] *21 fairness definitions and their politics*

Let’s break it down

- **fairness** is the absense of **harmful discrimination**

- **from whom?**
  - the **AI/ML system**, and by extension from the **system owner**

- **why?**
  - could manifest due to decisions made by the system owner (intentional or not), due to the training data, etc. (**sources** of discrimination)

Let’s break it down

- **fairness** is the absense of **harmful discrimination**

- **what is discrimination?**
  - difference in the treatment and/or impact of people
  - this means that two or more **individuals** or **groups** of people are **compared**

- **what is harm?**
  - categorization by [2017 NIPS K. Crawford]; boundaries not always clear
  - **representational harms**: e.g., stereotyping, racial/gender miscategorization
  - **distributional harms**: unfair distribution of a resource

[2017 NIPS K. Crawford] *The Trouble with Bias*

Fairness in Recommendations

- claim: almost all fairness concerns raised in the context of recommenders are about distributional harms

- fairness is the **fair distribution** of a **resource**
  - fair **for whom?**
  - **when** is the distribution fair?
  - distribution **of what** resource?

For Whom?

- recommender systems are multi-sided, with multiple stakeholders

- in the context of fairness, two important sides [2017 FATML R. Burke]
- **consumers**/end-users/receivers of recommendations
- **providers**/owners/producers of items being recommended

- harms can be for consumers, or providers, or both (e.g., reciprocal recommendations)
  - harms can be financial, ethical, legal, depending on the type of resource

[2017 FATML R. Burke] *Multisided Fairness for Recommendation*

For Whom?

- how do you compare? (to see if discrimination exists)

- **individual fairness**: compare two or more individuals that are similar (e.g., in terms of demographics, qualifications)
  - e.g., do two similar-qualified people get the same job offers?
  - **within-groups**

- **group fairness**: compare different groups of people; grouping attributes (e.g., demographics) often called **sensitive** or **protected**
  - e.g., are blacks receive similar recommendations as whites?
  - **across-groups**

When is the distribution fair?

- it defines when (distributional) harm occurs
- it depends on the resource

- fair typically means:
  - **equal** or uniform distribution
  - **proportional** to some given target (fixed, dynamic, etc.)
    - e.g., affirmative action

- one typically defines what perfect fairness means
- and measures unfairness by how far from being perfect you are.
  - e.g., Gini coefficient, measures of statistical divergence

Of What?

- what are the resources that can be distributed by a recommender?
- two main resource types
- **utility**: how relevant/accurate are the recommendations
  - requires feedback
- **exposure**: how much the recommender exposes/promotes items to users

Fairness of Utility

- **utility** is typically measured from the consumer viewpoint
  - hence, often (but not always!) associated with consumer fairness

- consumer viewpoint [2017 NIPS S. Yao et al.]
  - **rating prediction accuracy** should be **balanced** between protected and non-protected **consumer groups**

- provider viewpoint [2019 KDD A. Beutel et al.]
  - **ranking accuracy** should be **balanced** between protected and non-protected **providers groups**

[2017 NIPS S. Yao et al.] *Beyond Parity: Fairness Objectives for Collaborative Filtering*

[2019 KDD A. Beutel et al.] *Fairness in Recommendation Ranking through Pairwise Comparisons*

Fairness of Exposure

- **exposure** is typically measured from the provider viewpoint
  - hence, often (but not always!) associated with provide fairness

- consumer viewpoint [2018 FAT* R. Burke et al.]
  - **number of recommendations** of desired items should be **balanced** between protected and non-protected **consumer groups**

- provider viewpoint [2019 RecSys W. Liu et al.]
  - **number of recommendations** from each **provider** should be **equal**

[2018 FAT* R. Burke et al.] *Balanced Neighborhoods for Multi-sided Fairness in Recommendation*

[2019 RecSys W. Liu et al.] *Personalizing Fairness-aware Re-ranking for Microlending*

Taxonomy of Fairness in Recommendations

of What?

for Whom?

	Accuracy	**Exposure**
Consumer	prediction accuracy [2017 NeurIPS] ranking accuracy [2018 FAT * M. Ekstrand et al.]	equal exposure [2018 FAT* R. Burke et al.]
Provider	ranking accuracy [2019 KDD]	exposure coverage [2018 CIKM, 2019 RecSys] equal exposure [2018 FAT * R. Burke et al.] calibrated exposure [2018 RecSys]

some representative work:

[2017 NeurIPS S. Yao et al.] *Beyond Parity: Fairness Objectives for Collaborative Filtering*

[2019 KDD A. Beutel et al.] *Fairness in Recommendation Ranking through Pairwise Comparisons*

[2018 FAT * M. Ekstrand et al.] *All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness*

[2018 CIKM R. Mehrotra et al.] *Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems*

[2018 FAT * R. Burke et al.] *Balanced Neighborhoods for Multi-sided Fairness in Recommendation*

[2019 RecSys W. Liu et al] *Personalizing Fairness-aware Re-ranking for Microlending*

[2018 RecSys H. Steck] *Calibrated Recommendations*

Explanations of Recommendations

Explaining Recommendations

- recommendations are everywhere – explanations are there as well!

Related to items you've viewed

Frequently Bought Together

Customers Who Bought This Item Also Bought

Recommended for You Based on

Popular on Netflix

Trending Now

Watch It Again

Because you watched Dark

Why Explain?

Purpose	Description
Transparency	Explain how the system works
Effectiveness	Help users make good decisions
Trust	Increase users' confidence in the system
Persuasiveness	Convince users to try or buy
Satisfaction	Increase the ease of use or enjoyment
Education	Allow users to learn something from the system
Scrutability	Allow users to tell the system it is wrong
Efficiency	Help users make decisions faster
Debugging	Allows users to identify that there are defects in the system

[1984 B. G. Buchanan et al.] *Explanations as a Topic of AI Research, in Rule-based Systems*

[2017 UMUAI I. Nunes et al.] *A systematic review and taxonomy of explanations in decision support and recommender system*

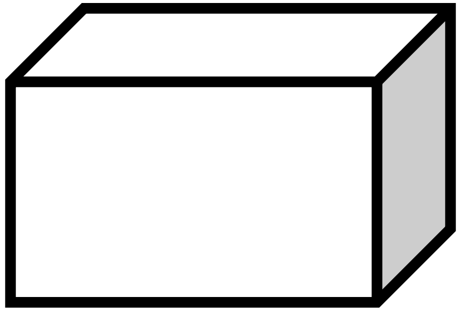
[2007 ICDE_w N. Tintarev et al.] *A survey of explanations in recommender systems.*

White-Box vs Black-Box Models

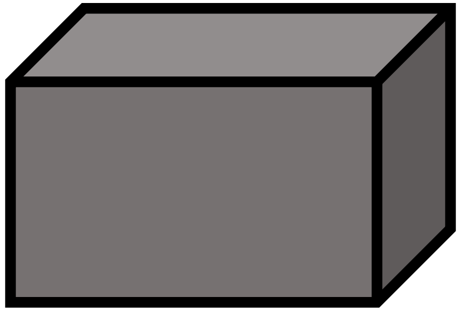
- an **explanation** describes how the system reaches a decision
  - requires access to the inner workings of the system
  - a **white box model**

- but often the recommender is a **black box model**
  - no knowledge of inner workings
  - we can only try to **interpret** how it reaches a decision

- often distinction between **explanation** and **interpretation** of a model



white box model

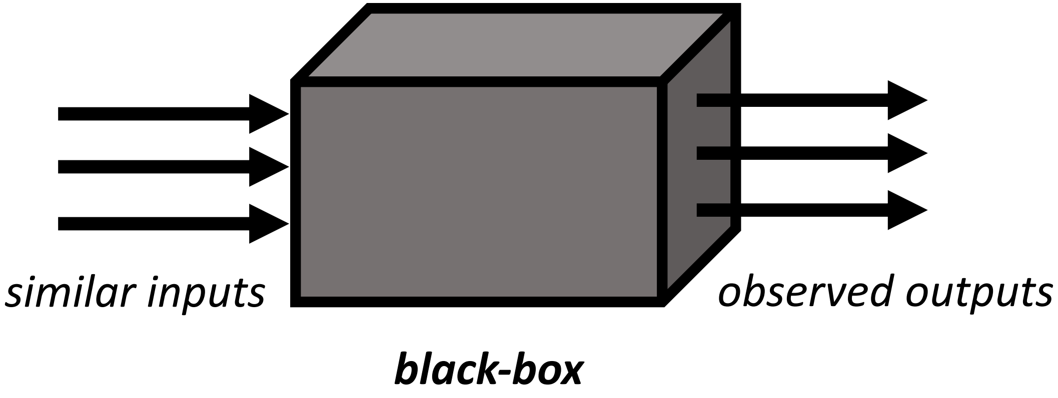


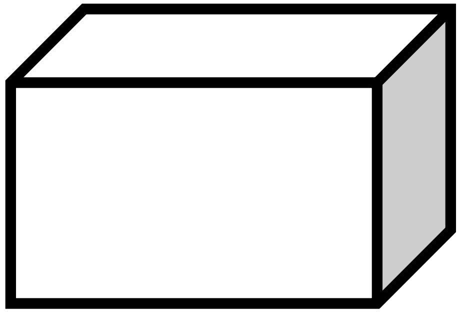
black box model

Local Proxy Models

- cannot see inside the black box, but can **observe** its **inputs** and **outputs**
- to explain a **given input-output** pair (preferences-recommendations)

1. push similar inputs and **observe** outputs
2. fit a transparent **local proxy** model to the observed input-output pairs
3. use the local proxy model to **generate explanations** for the given input-output pair





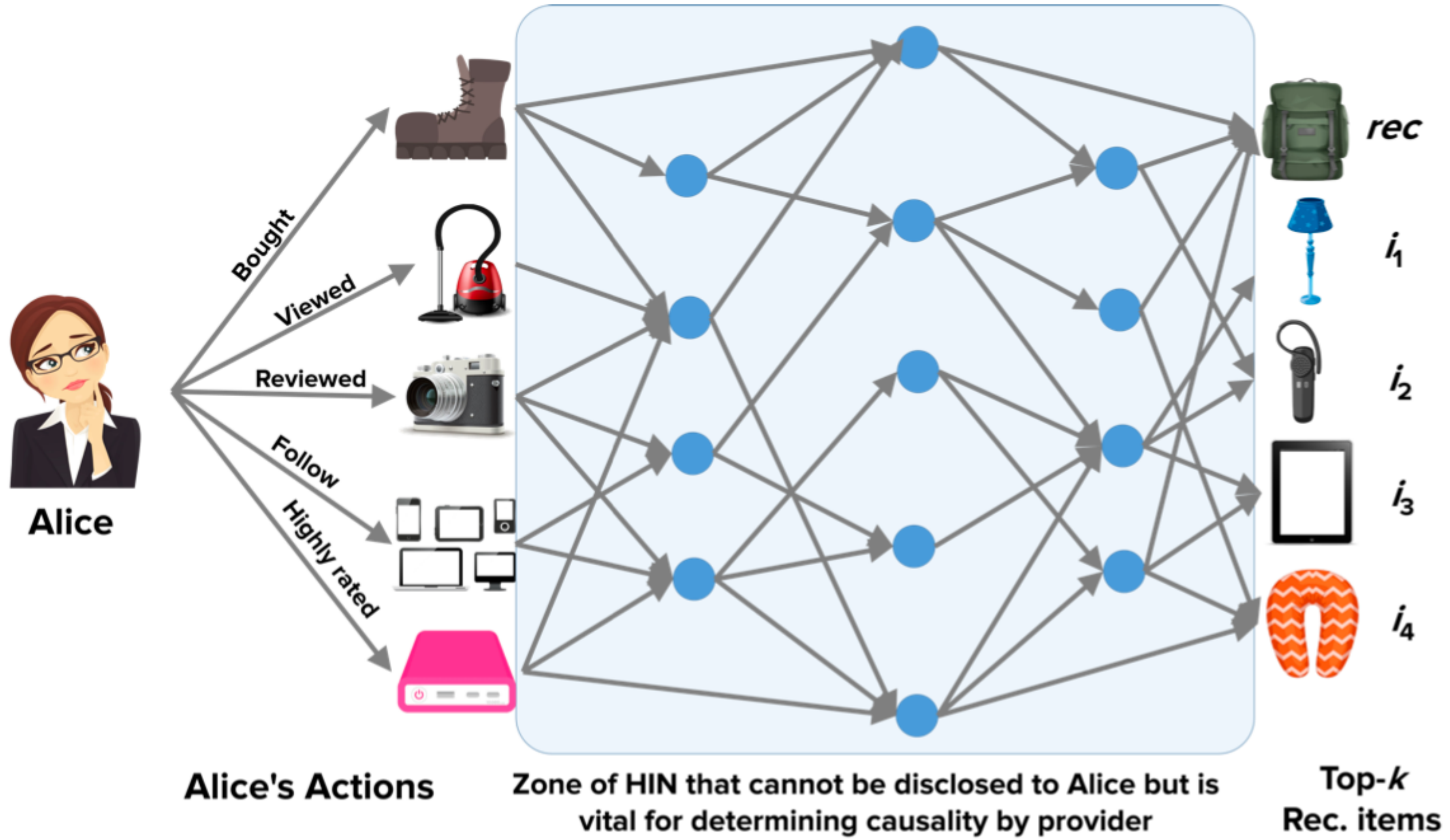
[2019 SAC C. Nóbrega et al.] *Towards Explaining Recommendations Through Local Surrogate Models*

Counterfactual Explanations

- consider a **causal relationship**: *"If X had not occurred, Y would not have occurred"*
- it explains why Y occurred: it's because X occurs

- a **counterfactual explanation** of a specific **recommendation** describes the smallest **change** to **preferences** that results in **not seeing** that recommendation
  - Example: to explain "Why was I recommended item Y?" look for smallest changes in preferences so that item Y no longer appears in the recommendations

- preferences = **factual**
- change to preferences = **counterfactual**



Alice: Why did I receive this recommendation “Jack Wolfskin backpack”?

PRINCE: You **bought** “Adidas Hiking Shoes”;
You **reviewed** “Nikon Coolpix Camera” with “Sleek! Handy on hikes!”;
You **rated** “Intenso Travel Power Bank” highly.

If you **had not** done these actions:
“iPad Air” **would have replaced** “Jack Wolfskin backpack”.

[2020 WSDM A. Ghazimatin et al.] *PRINCE: Provider-side Interpretability with Counterfactual Explanations in Recommender Systems*

Explanations + Fairness

Towards Fairness-Aware Explanations

- the user (consumer or provider) may wonder if they are treated fairly

- if they are treated **fairly**, how can the system **assure** the user?
  - provide **fairness assurances**

- if they are treated **unfairly**, how can the system **explain** itself?
  - provide **unfairness explanations**
  - further investigate whether the unfairness violation can be justified (was it intentional?)

- **fairness-aware explanations**
  - fairness assurances and unfairness explanations

Towards Fairness-Aware Explanations

- the paradigm of counterfactual explanations may be useful but there are some key **differences**

- **desired output is not unique**
  - in conventional explanations, the desired output is the recommendation list without a specific item
  - in fairness-aware explanations, the desired output is a more fair output, which can be achieved in many ways

- **there are multiple instances**
  - in conventional explanations, there is a single instance to explain
  - fairness definitions are often based on aggregating multiple instances

conclusion

Take Away

- to build user **trust** in recommender systems

- you need to be able to **offer explanations to** and **ensure fairness of** multiple stakeholders

- interesting **research directions** to explore, also at the **intersection** of explainability and fairness.

thank you