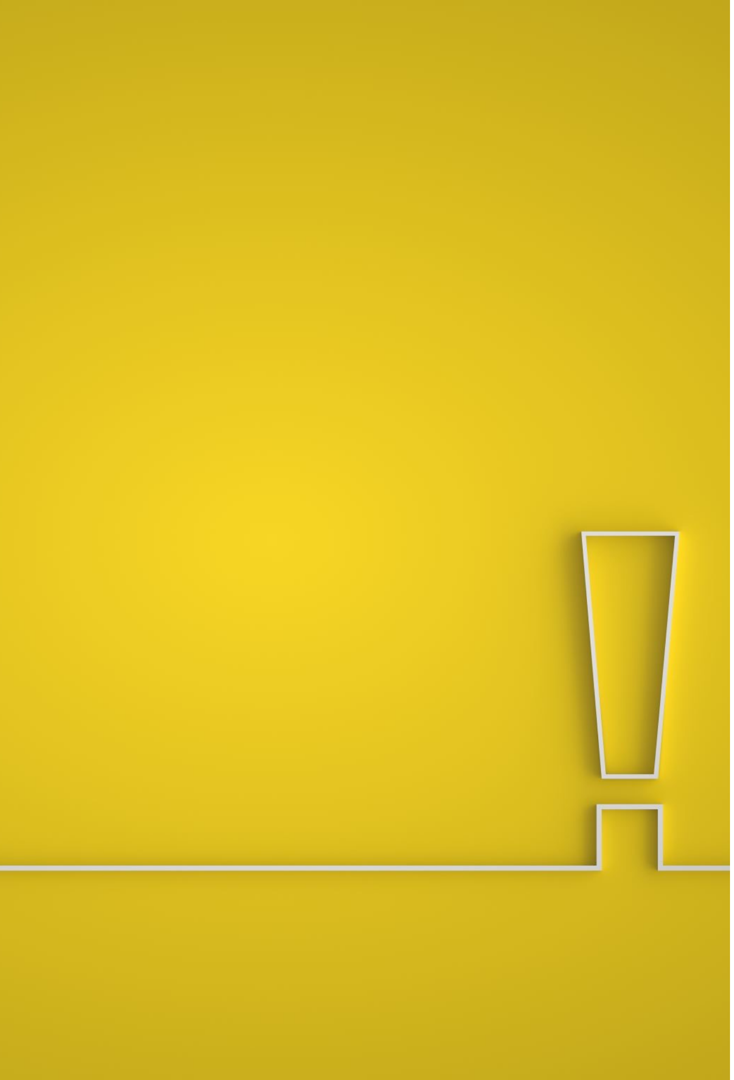
Whom are you Explaining to, and Why?

Professor Nava Tintarev
Department of Advanced Computing Sciences



Disclaimers:

- No ChatGPT/LLM was used (or hurt) during the course of developing this presentation.
- All errors herein are my own.
- Focus on artificial advice givers, and explanations for end-users.

High stakes, poor transparency!

2,75 miljoen euro boete voor Belastingdienst om toeslagenaffaire

De Belastingdienst krijgt een boete van 2,75 miljoen euro voor de toeslagenaffaire. Dat meldt de Autoriteit Persoonsgegevens, de privacywaakhond van de overheid.

19th Oct. 2021, het Parool

Welfare surveillance system violates human rights, Dutch court rules

Government told to halt use of AI to detect fraud in decision hailed by privacy campaigners

5th Feb. 2020, The Guardian

Top 400-lijst

Ook jongeren die niets hebben uitgehaald komen op Amsterdamse risicolijst Top 400

In de documentaire 'Moeders' komen moeders aan het woord van jongeren die in de Top 400 staan. De vrouw op de foto is een actrice, niet een echte moeder van een Top-400 kind. Beeld VPRO, BALDR Film

22nd Nov. 2022, Trouw

High stakes, poor transparency!

2,75 miljoen euro boete

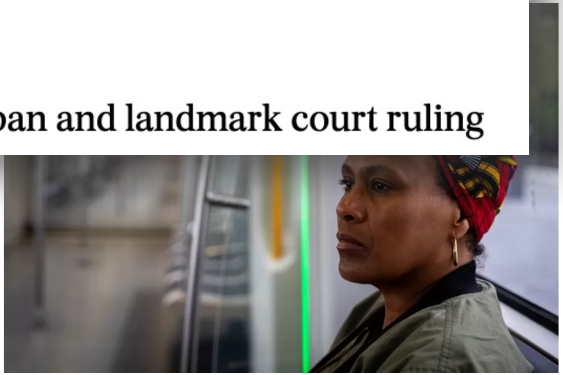
Top 400-lijst

Ook jongeren die niets hebben uitgehaald komen

THE ALGORITHM ADDICTION

DECEMBER 20, 2022

Mass profiling system SyRI resurfaces in the Netherlands despite ban and landmark court ruling



In de documentaire 'Moeders' komen moeders aan het woord van jongeren die in de Top 400 staan. De vrouw op de foto is een actrice, niet een echte moeder van een Top-400 kind. Beeld VPRO, BALDR Film

19th Oct. 2021, het Parool

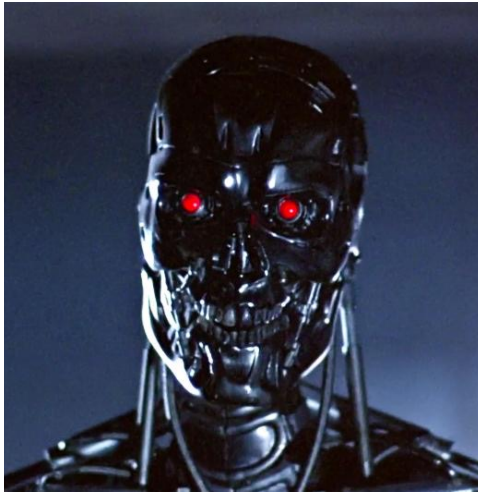
Welfare surveillance system violates human rights, Dutch court rules

Government told to halt use of AI to detect fraud in decision hailed by privacy campaigners

5th Feb. 2020, The Guardian

22nd Nov. 2022, Trouw

I for one, welcome our AI ~~overlords~~ collaborators

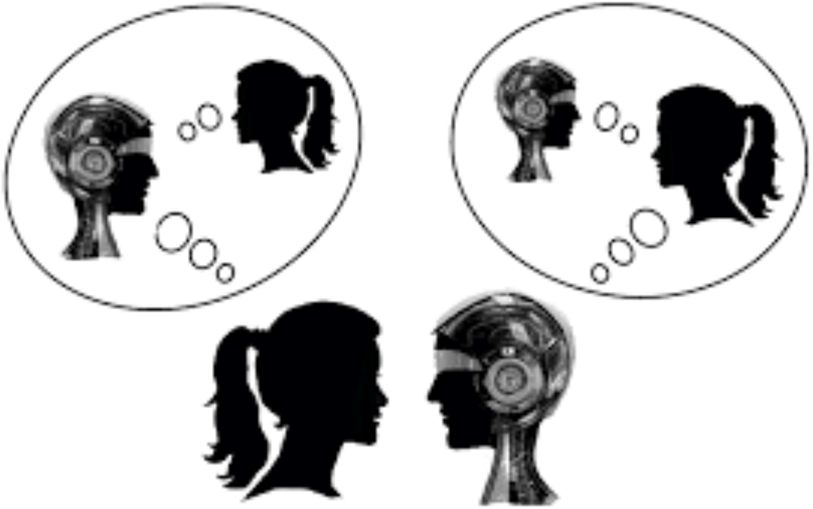


Credit: Alamy

- AI is not evil
- Human-like appearance leads to assumptions.
- AI systems make mistakes.
- AI can be used as a tool.
- Do you know who else makes mistakes….

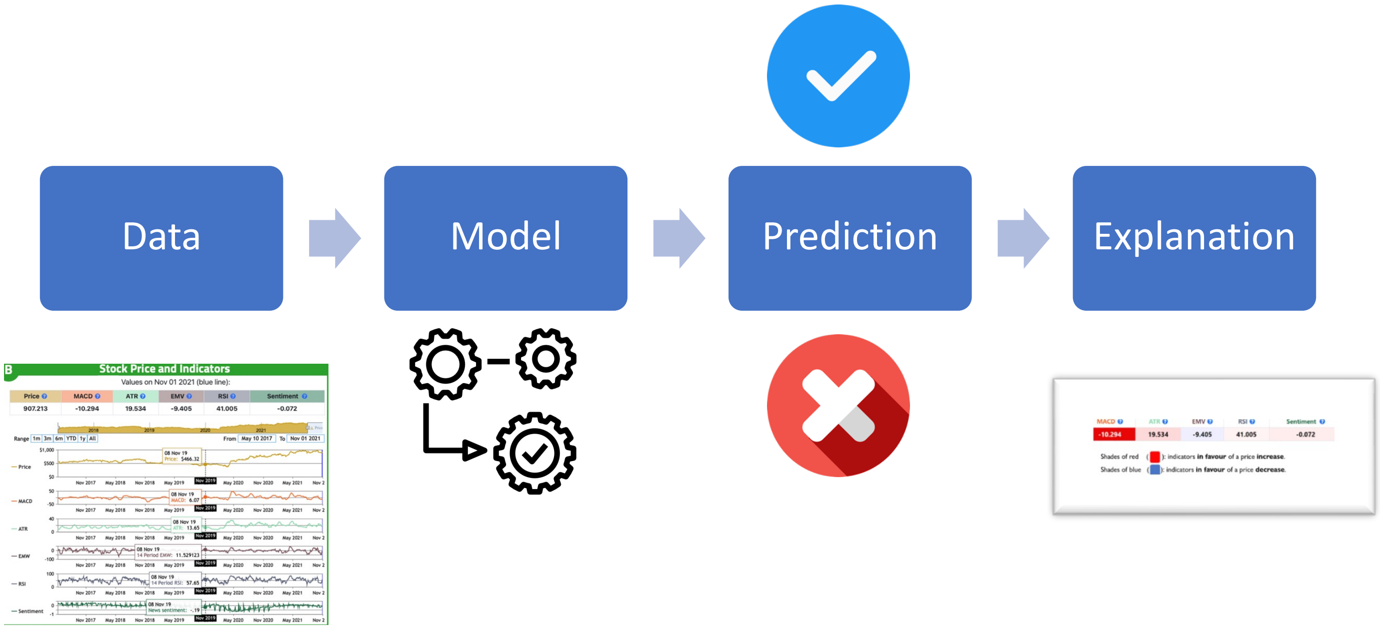
“Hybrid-intelligence”

Credit: Web Vectors by Vecteezy



Credit: Hybrid-Intelligence Centre

Before we can diagnose the causes



What Prof. Tintarev thinks “went wrong”

Overemphasis on model “performance”

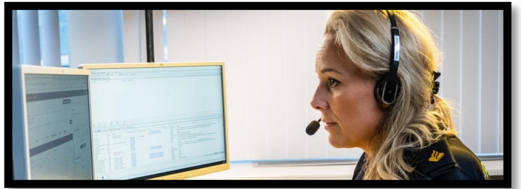
- insufficiently grounded in application and consequences

Explanations considered late

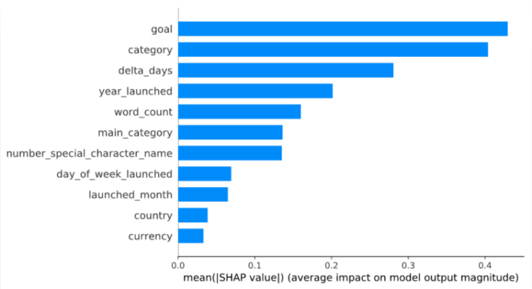
- If at all
- Not human understandable

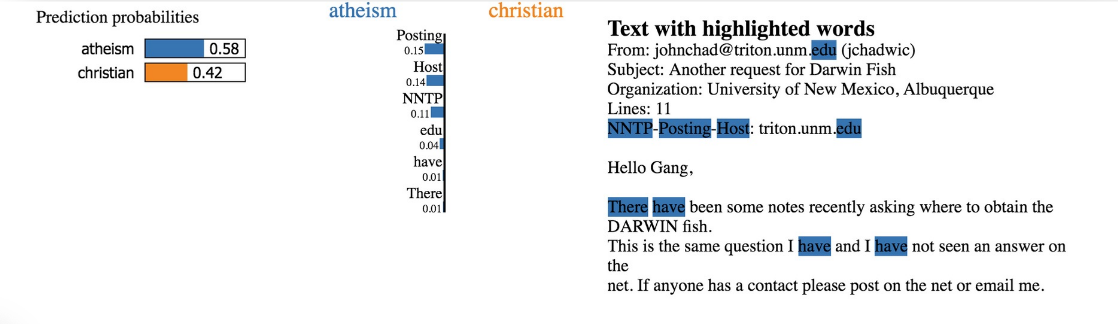
Credit: [xkcd]()

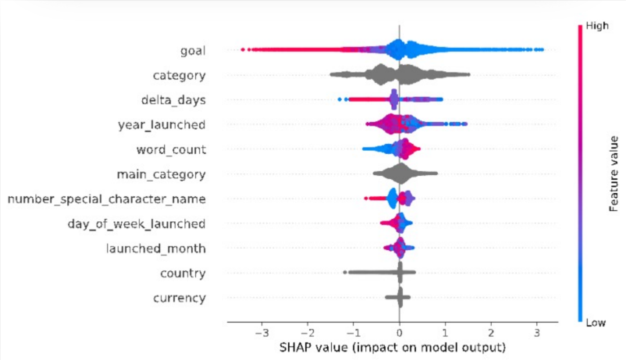
- **Predictive policing by the Dutch police.**
- We discovered that the interpretation and filtering of the AI **outputs was too difficult to leave to the police officers** themselves.
- To solve this problem, the police set up an intelligence unit which translates the AI outputs into what police officers must actually do.



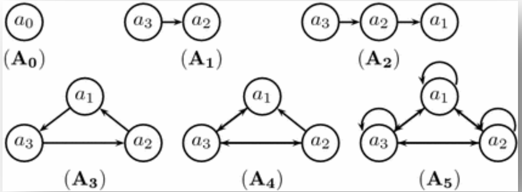
Credit: politie.nl



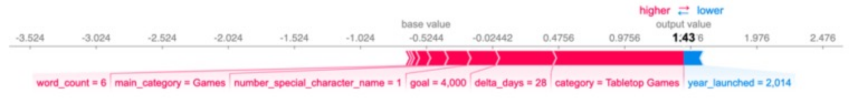




Progress on generating explanations! How understandable are they?

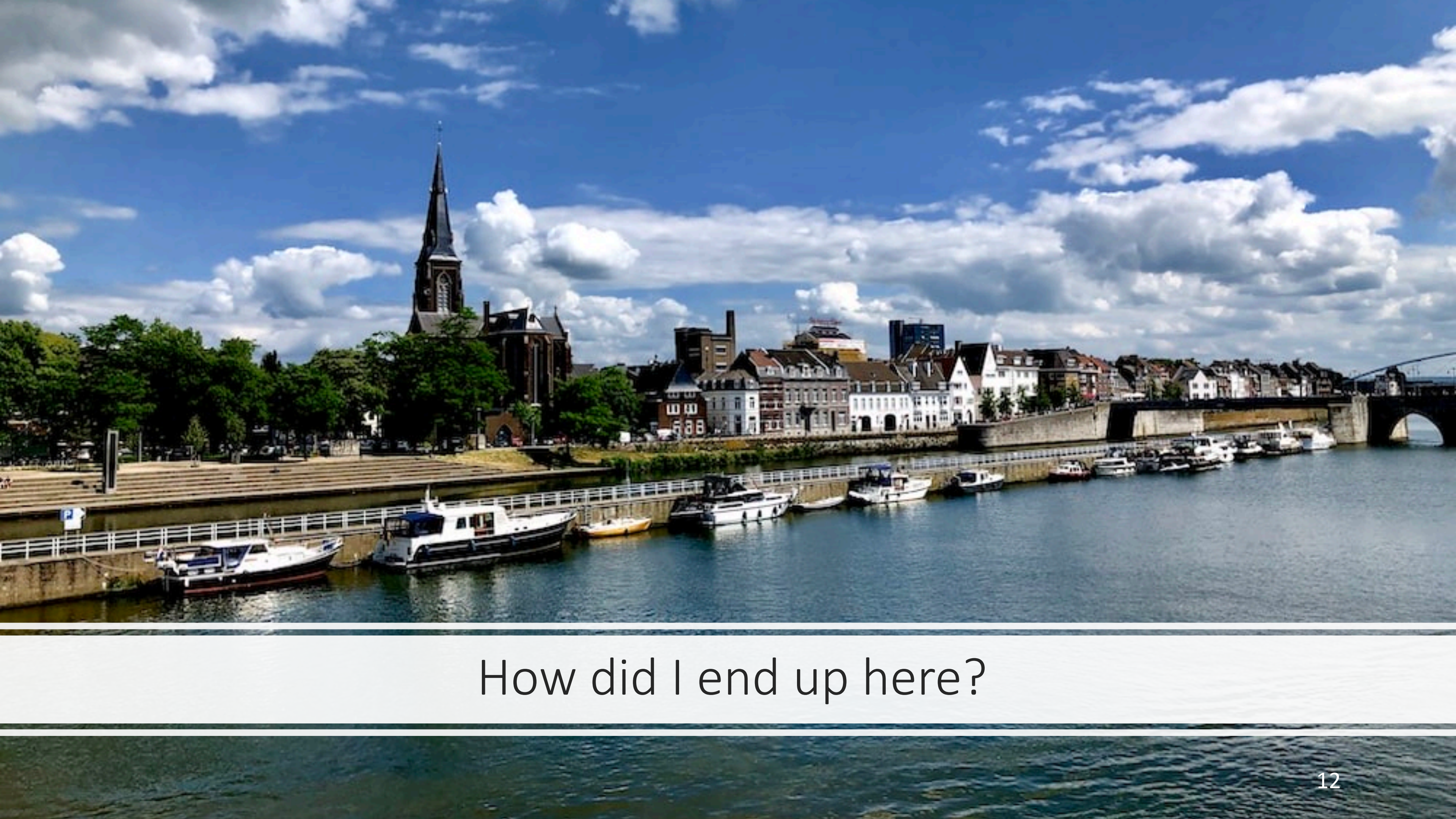


Tumor is diagnosed as malignant if, $[((\text{smoothness} \geq 0.089) + (\text{standard error of area} \geq 53.78) + (\text{largest radius} \geq 18.225)) \geq 2] + [((98.76 \leq \text{perimeter} < 114.8) + (\text{largest smoothness} \geq 0.136) + (105.95 \leq \text{largest perimeter} < 117.45)) \geq 2] \geq 1$



~~Explainability~~ Interpretability is

- the degree to which a human can understand the cause of a decision [Miller 2017; Biran and Cotton]
- the degree to which a human can consistently predict the model's result [Kim et al 2016]
- To which extent the model and/or the prediction are human-understandable [Guidotti et al 2018, Amparore et al 2021]

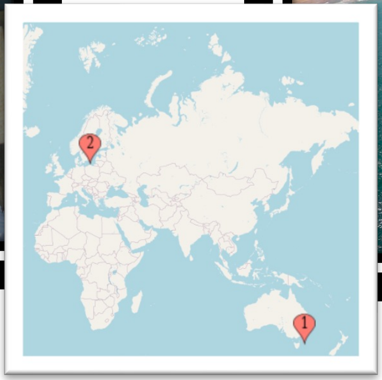


How did I end up here?

Worlds apart?

Sept 2000 – March 2005
Computer science: Uppsala University, Sweden

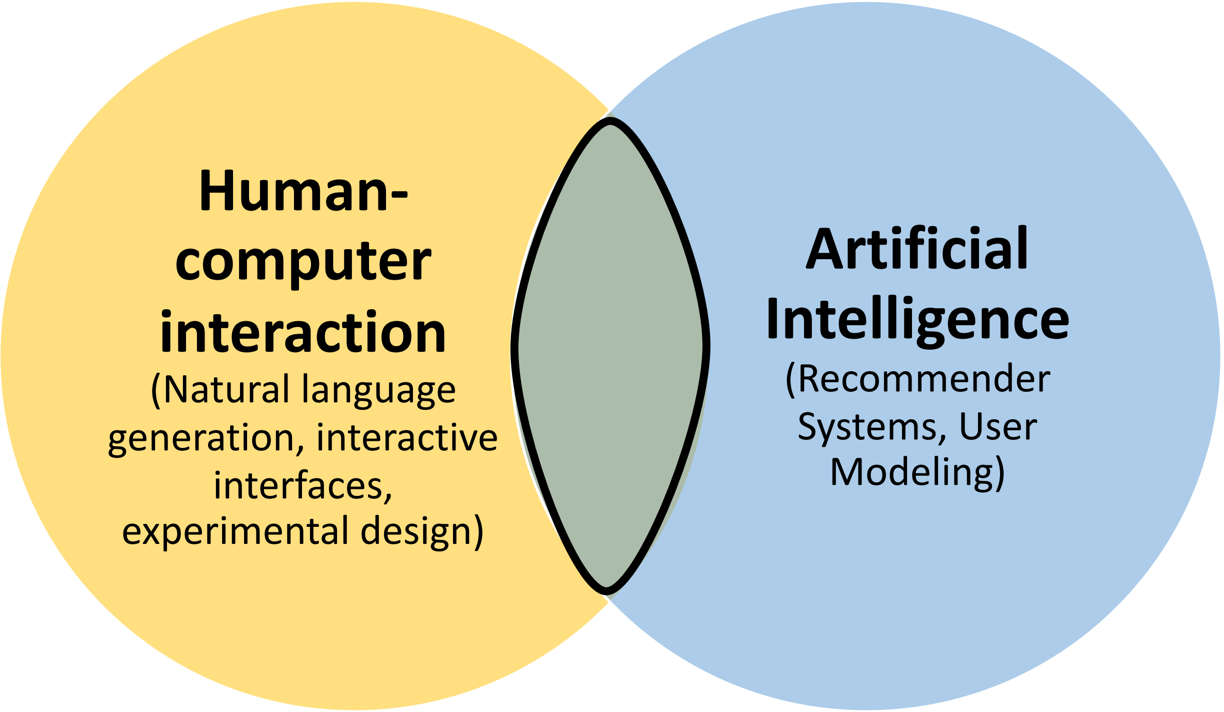
Feb 2003 – Nov 2003
Psychology: Uni. of Wollongong, Australia



Intelligent User Interfaces

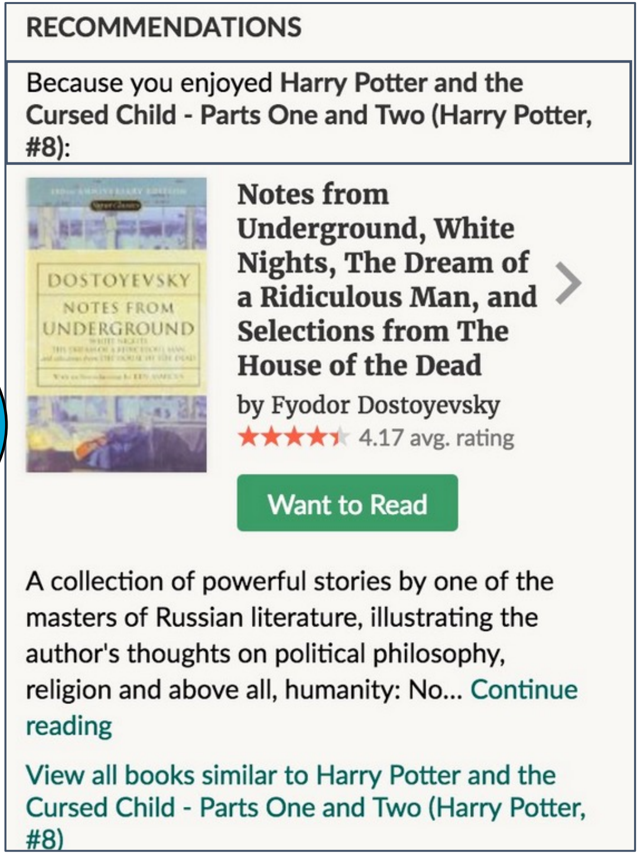
Program chair:
UMAP’21, IUI’20

Senior reviewer:
ECAI, IUI, Recsys, UMAP, TiiS



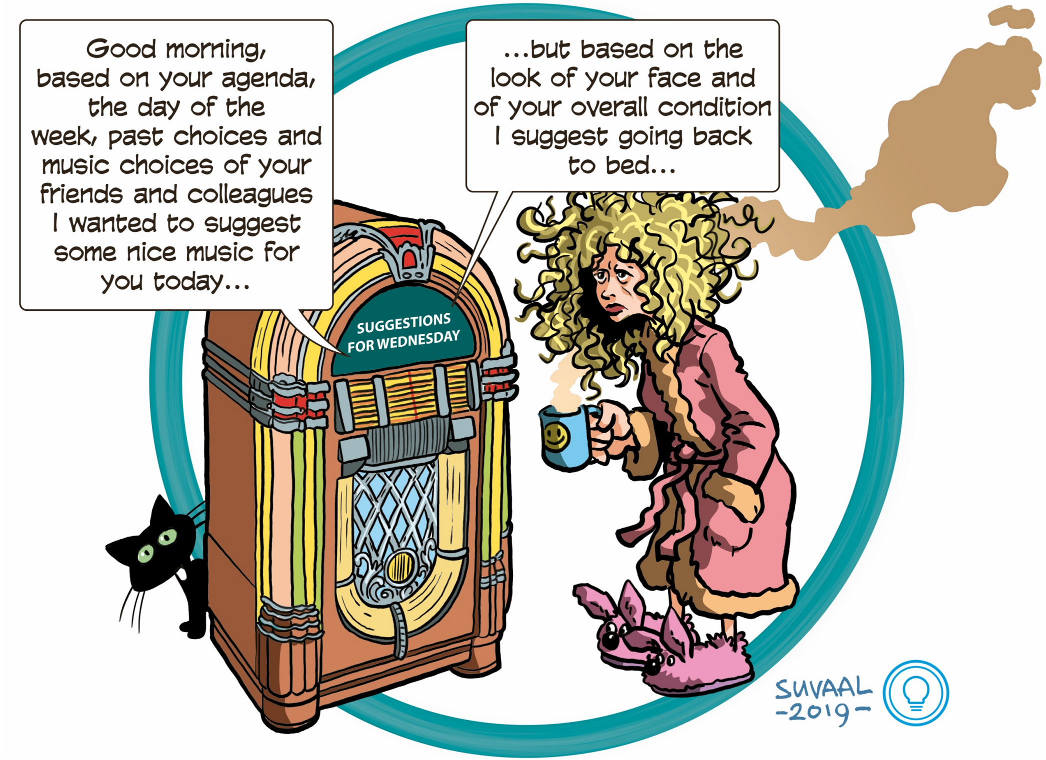
PhD: Decision support in Recommender systems

Users wonder why this book.
What makes for a **good** explanation?




Judith Masthoff


Ehud Reiter



Credit: Erwin Suvaal from CVIII ontwerpers

- The explanation here is not fully transparent.
- It recommends <u>no</u> music.
- It is rather useful.
- Sometimes this is refered to as Justification.

Why are you explaining?

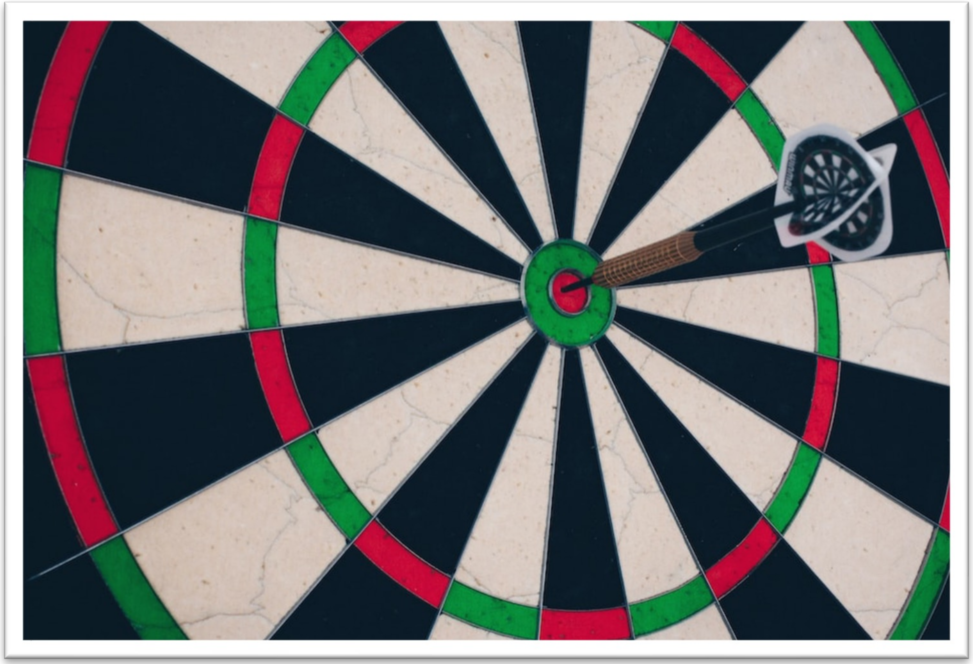
Why explain (recsys)

Purpose	Description
Transparency	**How** was this recommendation made?
Effectiveness	**Why** (not) this item?
Persuasiveness	Why **must** you buy this item?
Trust	Increase users' confidence in the system
Scrutability	Allow users to tell the system it is wrong
Satisfaction	Increase the ease of use or enjoyment
Efficiency	Help users make decisions faster

Recommender Systems Handbook Chapter (3rd Ed) [Tintarev and Masthoff 2022]

But what is your goal?

- Transparent, not Scrutable.
- Effective, not Transparent.
- **Satisfying, but not Effective.**



Satisfying, but not Effective

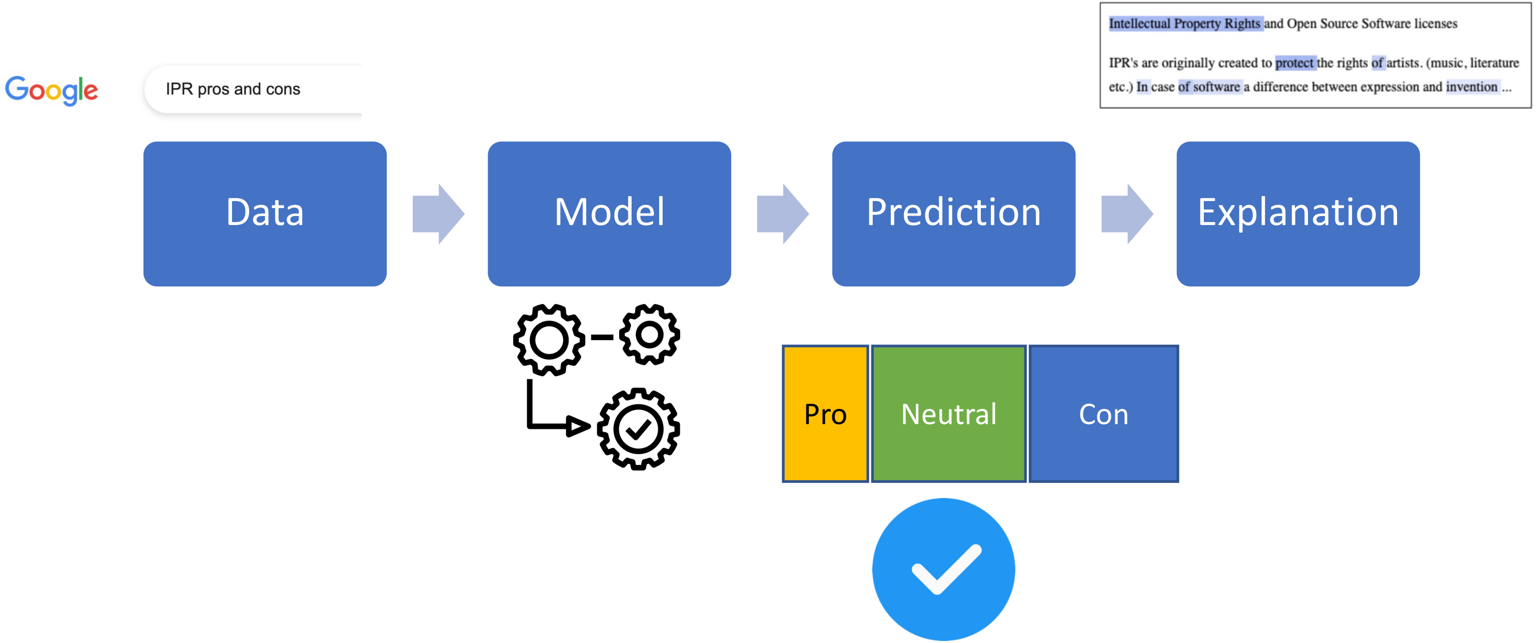
- Benefit of personalized explanations?
- Personalized explanations worse for decisions
- but people were more satisfied! [Tintarev & Masthoff, 2012]
- What you measure matters!

Non-personalized:
"This movie belongs to the genre(s): Action & Adventure and Comedy. On average other users rated this movie 4/5.0"

Personalized:
"Unfortunately, this movie belongs to at least one genre you do not want to see: Action & Adventure. It also belongs to the genre(s): Comedy. This movie stars Jo Marr and Robert Redford."

Baseline:
"This movie is not one of the top 250 movies in the Internet Movie Database (IMDB). "

Transparency as a purpose: disputed topics



Transparency as a purpose: disputed topics

[Draws et al. CHIIR’23]

- Predict stance for debated topics
- Equally correct and incorrect
- State-of-the-art prediction models and explanation models
- **Task:** ”guess” the prediction
    - E.g., what do you think the system predicted this search result as? NEUTRAL about IPR.

Intellectual Property Rights and Open Source Software licenses

IPR's are originally created to protect the rights of artists. (music, literature etc.) In case of software a difference between expression and invention ...

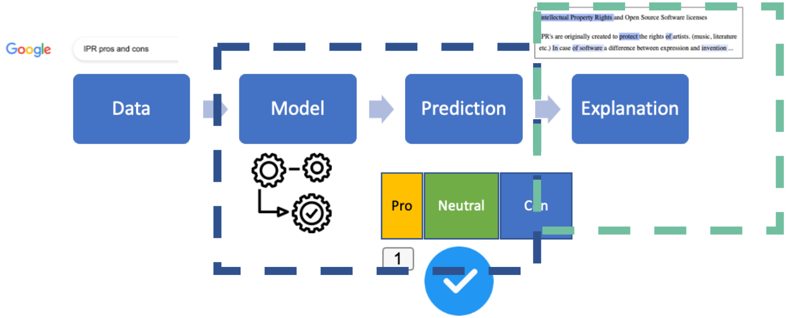
Pro

Neutral

Con

Tim Draws, PhD, TU Delft

Transparency as a purpose: disputed topics

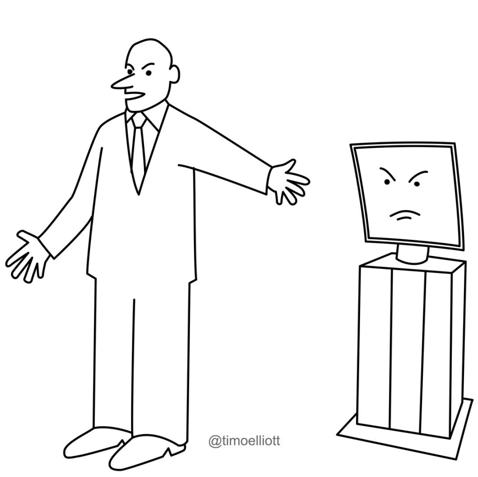


Depending on the classification models:

- We influenced how correct predictions are
- **Combinations** of model and explanation worked better.
- <u>I.e., how understandable the explanation is</u> – for a person

- Focus on transparency – but no scrutability.

Trust as a purpose



His decisions aren't any better than yours — but they're WAY faster...

Credit: Timo Elliott

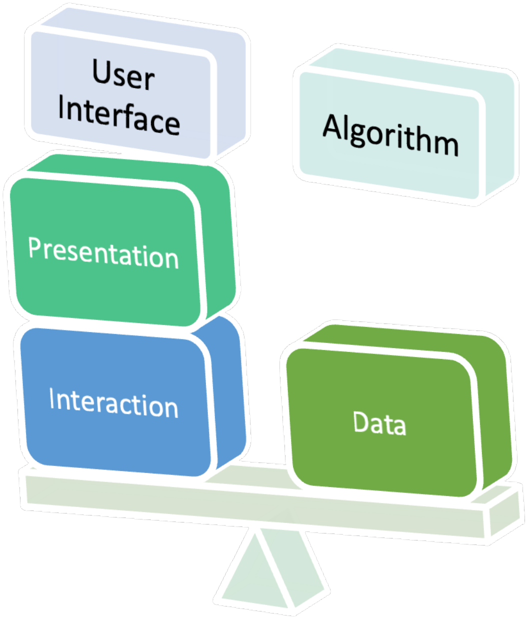
Trust as a purpose

- Is there such a thing as “too much” trust?
- What is “appropriate trust”?
- Is there a difference b/w explanation styles?

How you explain matters

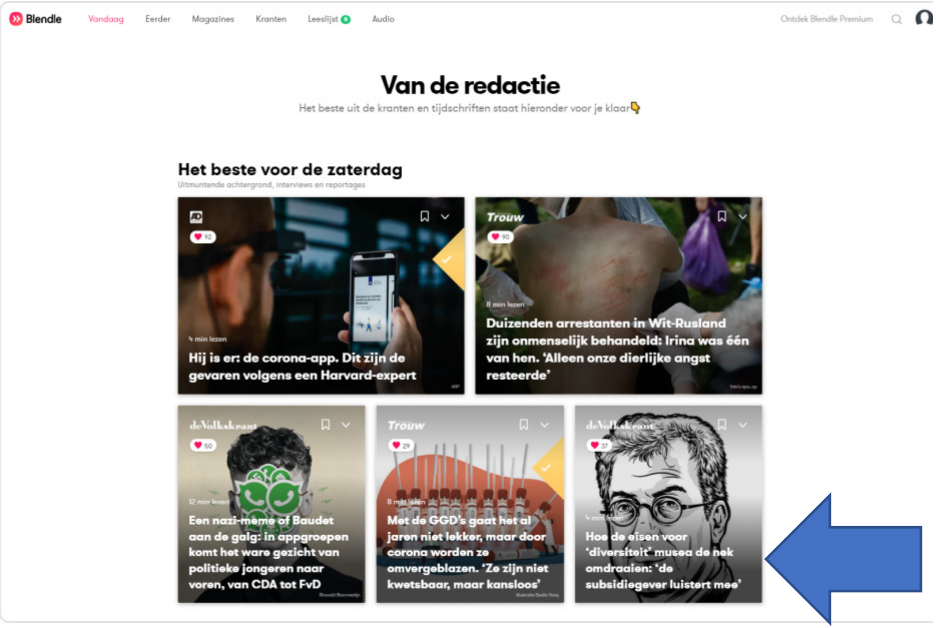
Top 10 Lessons Learned Developing Deploying and Operating Real-world Recommender Systems - *Francisco J. Martin, Strands (Recsys'10)*

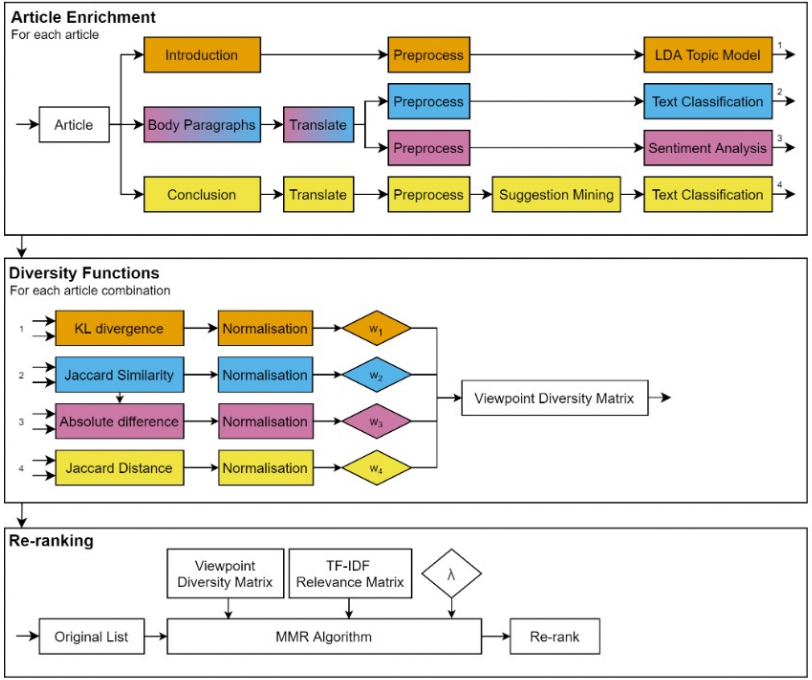
Focus on interface as much as algs; … the User Interface needs to get the lion's share of the effort (50%) compared to algorithms (5%), knowledge (20%), analytics (25%)



Mats Mulder, Oana Inel, Jasper Oostman, FACCT'20

Presenting diverse articles (Blendle)



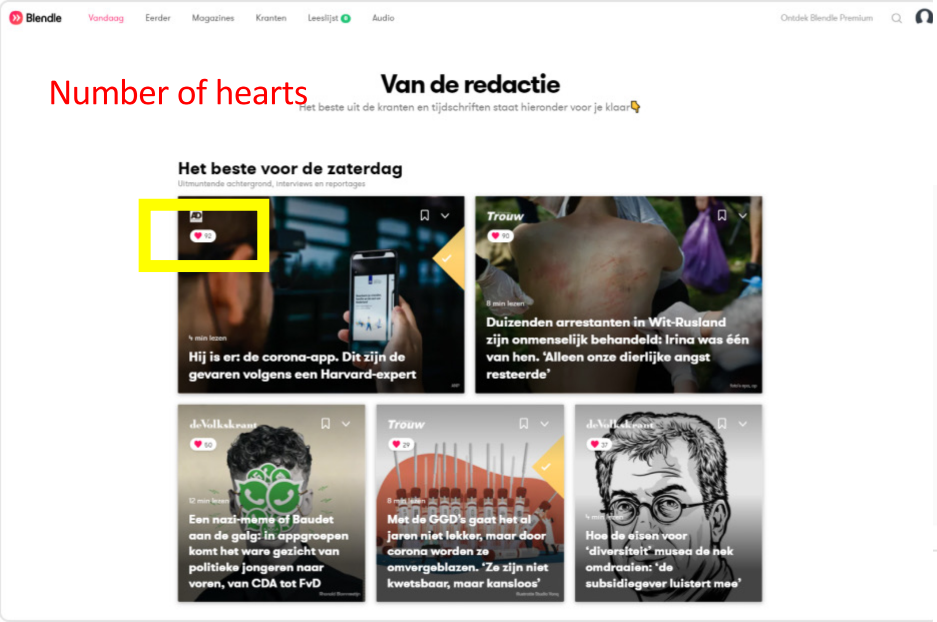


Mats Mulder, Oana Inel, Jasper Oostman, FACCT'20

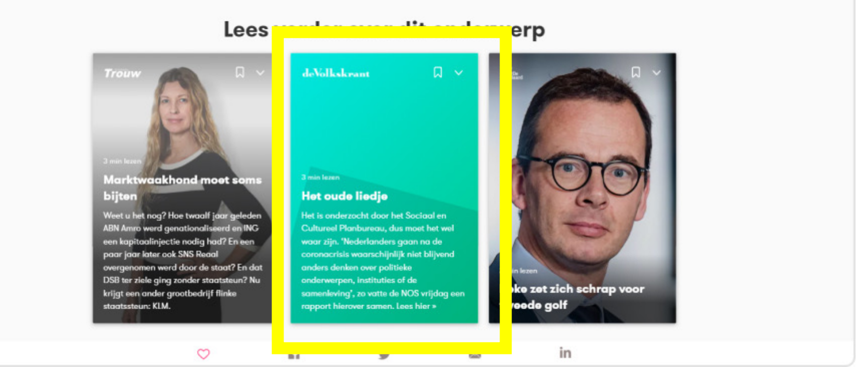
What do people click on? Blendle

Interface matters?

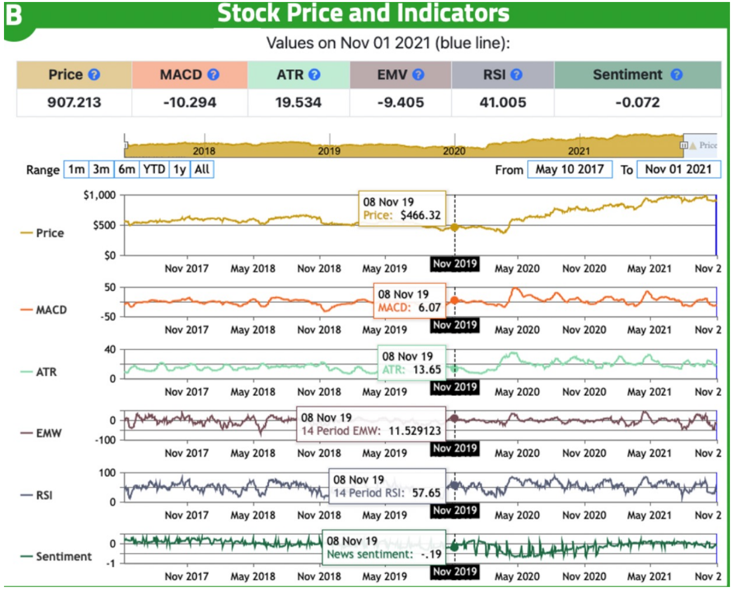
Number of hearts

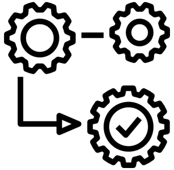


Thumbnail 3.1% more



Back to Trust as a purpose





SELL

HOLD

You can explain in many different ways!

Cau et al. IUI'23; TiiS'23

Inductive Explanation

The AI generated this prediction learning from these examples:

Example	Date	Price	MACD	ATR	EMV	RSI	Sentiment	Next week price
1	2018-06-07	539.507	-11.286	17.761	-7.281	42.927	-0.164	increase
2	2018-10-29	625.385	-8.183	18.369	-8.525	43.288	0.009	decrease
3	2019-10-14	438.583	-13.277	17.910	-8.187	44.041	-0.201	increase

Abductive Explanation

The AI focused on these indicators for making the prediction:

MACD	ATR	EMV	RSI	Sentiment
-10.294	19.534	-9.405	41.005	-0.072

Shades of red ( ): indicators **in favour** of a price **increase**.

Shades of blue ( ): indicators **in favour** of a price **decrease**.

The brighter an indicator value is, the higher its importance in the AI's decision.

Deductive Explanation

The AI generated this prediction following this rule learned from the data:

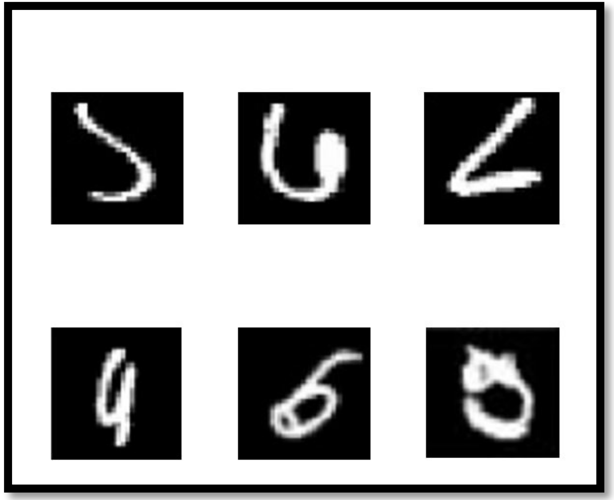
	Indicator	Value	Check	Threshold value	Contrast	Decision
if	RSI	41.005	>	32.392	22.3%	
and	MACD	-10.294	<=	-0.655	17.8%	increase
and	ATR	19.534	>	3.459	5.9%	

Federico Cau (UM)

Davide Spano (Cagliari)

Hanna Hauptmann (Utrecht)

“Too much” trust



Agree more with the system

- System right - human right 94%.
- System wrong - human wrong 75%.
- Also agree when system is wrong
  → more mistakes for the person!

D **Inductive Explanation**

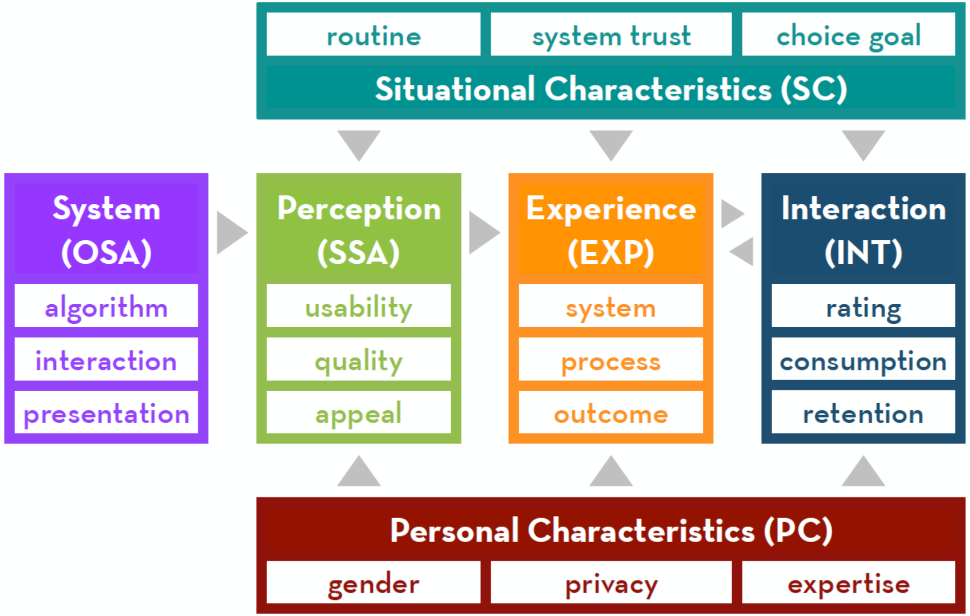
The AI generated this prediction learning from these examples:

Example	Date	Price	MACD	ATR	EMV	RSI	Sentiment	Next week price
1	2018-06-07	539.507	-11.286	17.761	-7.281	42.927	-0.164	increase
2	2018-10-29	625.385	-8.183	18.369	-8.525	43.288	0.009	decrease
3	2019-10-14	438.583	-13.277	17.910	-8.187	44.041	-0.201	increase

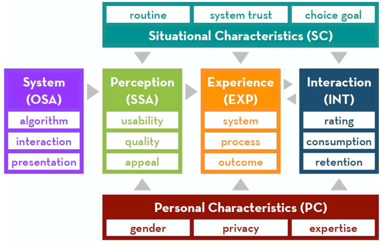
Explaining, to whom?

Factors influencing explanation effectiveness

[Knijnenburg2012]



Situational Characteristics?

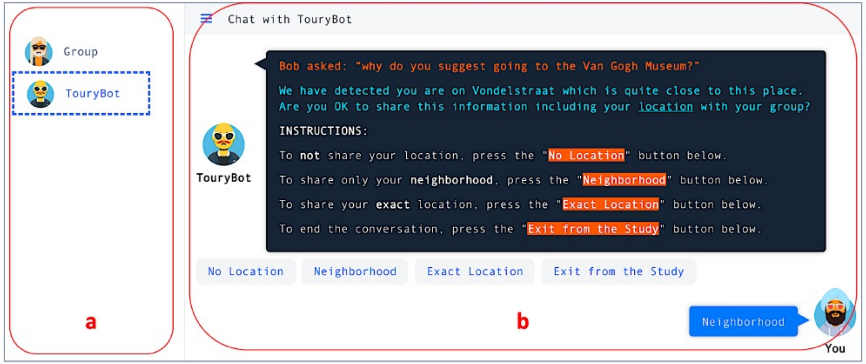


- Explaining the``Unexpected” [Rieger2021, Draws2020, Draws2022, Draws2023]
- Group explanations [Najafian2023, Barile2021]
- Multi-stakeholder explanations e.g., jobs [Schellingerhout’22]

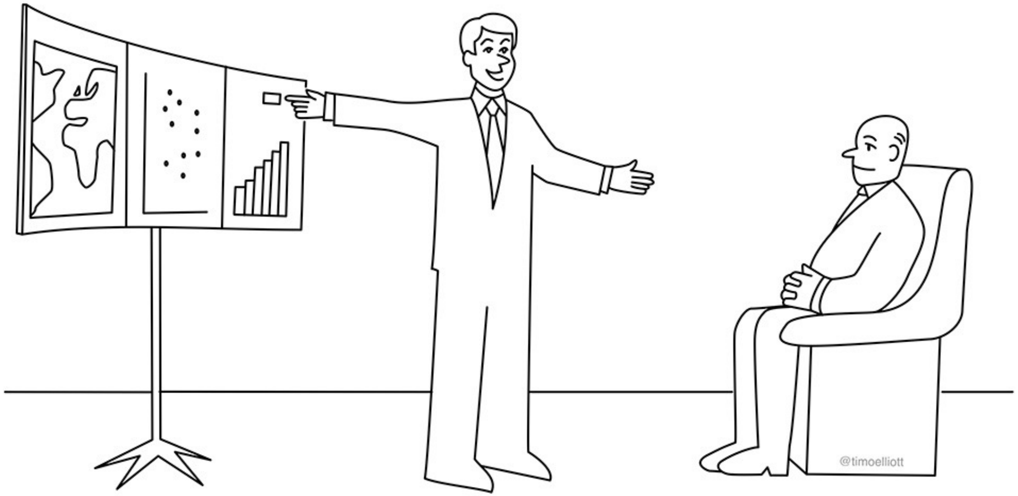

Dr Shabnam Najafian


Francesco Barile, Assistant Professor, University of Maastricht


Roan Schellingerhout, PhD, University of Maastricht



Explaining the “unexpected”



“And our unique JustifyIt™ feature uses deep learning to find data that agrees with your point of view!”

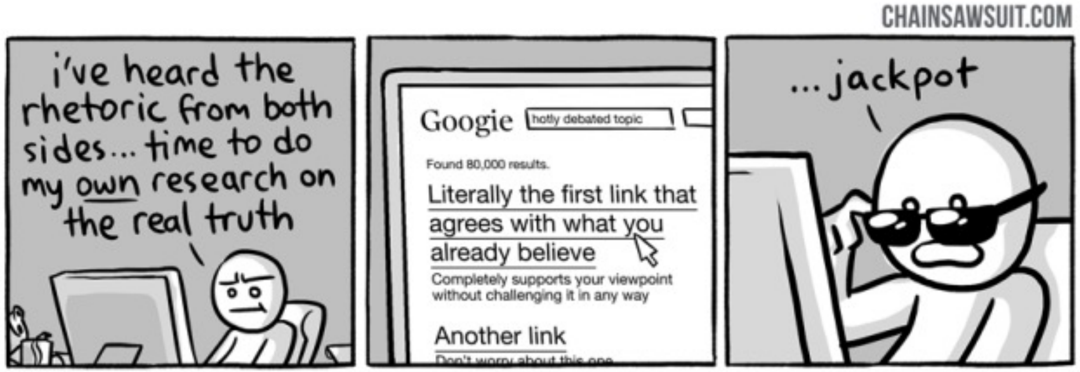
Credit: Timo Elliott

Explanations for "unexpected" content

- Humans also make mistakes!

- During online information search, users tend to select search results that confirm pre-existing beliefs or values and ignore competing possibilities **(Confirmation bias)** [Azzopardi2021]

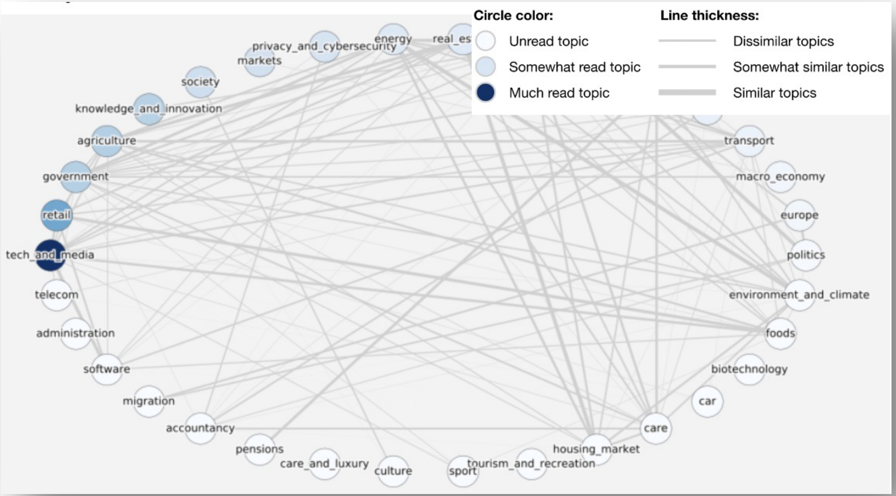
- **Results representing alternative viewpoints are in that sense "unexpected"**



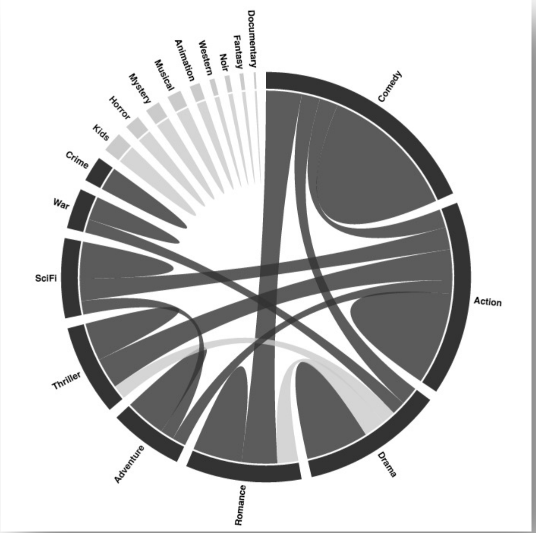
Credit: Chainsawsuit

Explanations for "unexpected” content

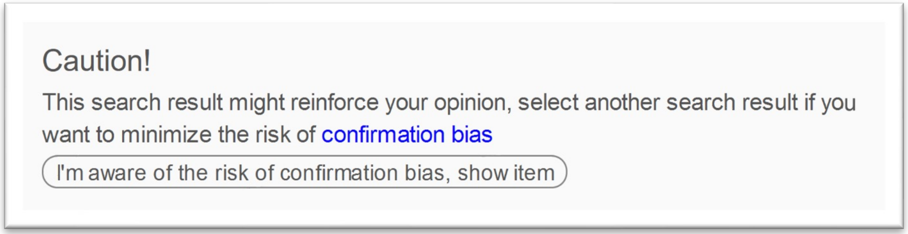
Sullivan et al. Collaboration with FD Media @ICTwIndustry, *ExUm Workshop, UMAP*, 2019



Tintarev et al. *ACM Symposium On Applied Computing,* **2018**



Alisa Rieger,
PhD, TU Delft

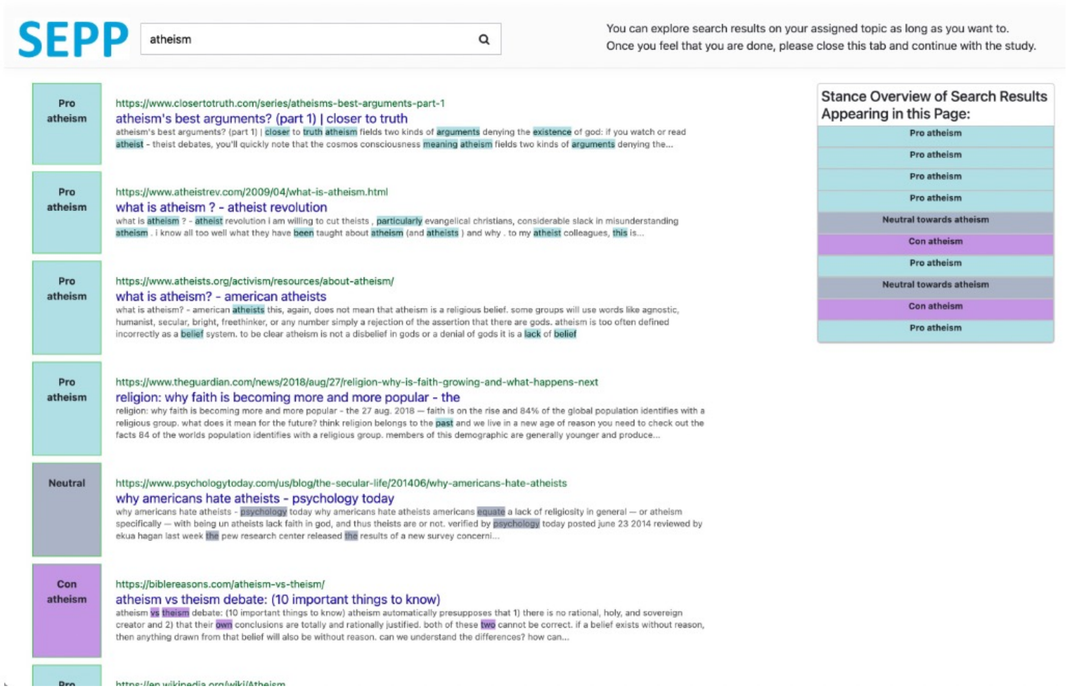


Rieger et al. *ACM HyperText'2021*

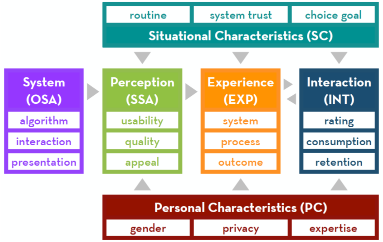
Cognitive bias mitigation in search

- Aim reduce clicks on results that confirm user's held opinion
- **Targeted warning:** effective
- **Extra step to view:** reduces clicks, but…
- Works well on random search results.
- Potential of misuse…
- Boosting! Preserve user autonomy, and enduring effects. [Lorenz-Spreen et al., 2021]

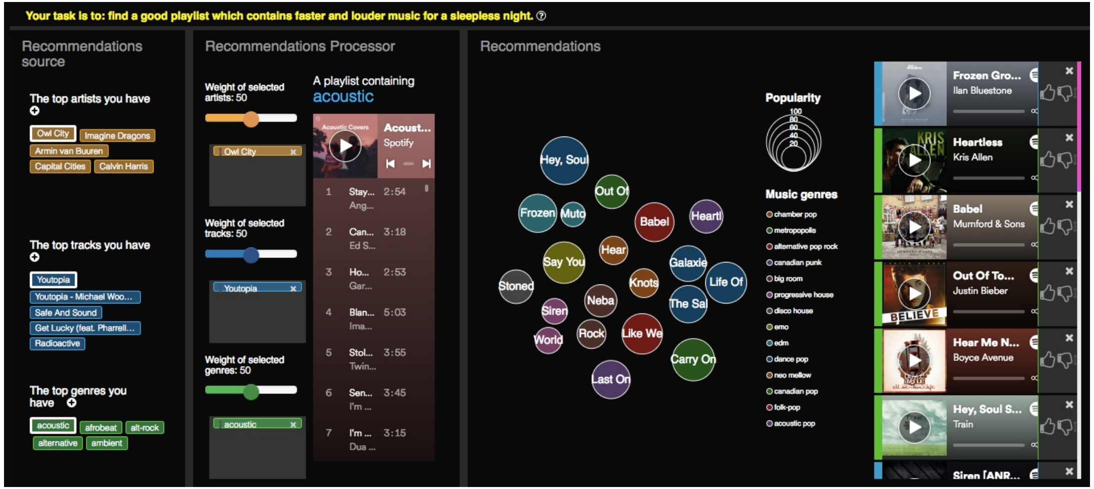
Interactions in search interface



Personal Characteristics?



Working memory & Expertise [Jin et al 2018: UMAP and Recsys, UMUAI 2019]
Need for Cognition [Rieger et al. *ACM HyperText'2021* 🏆]

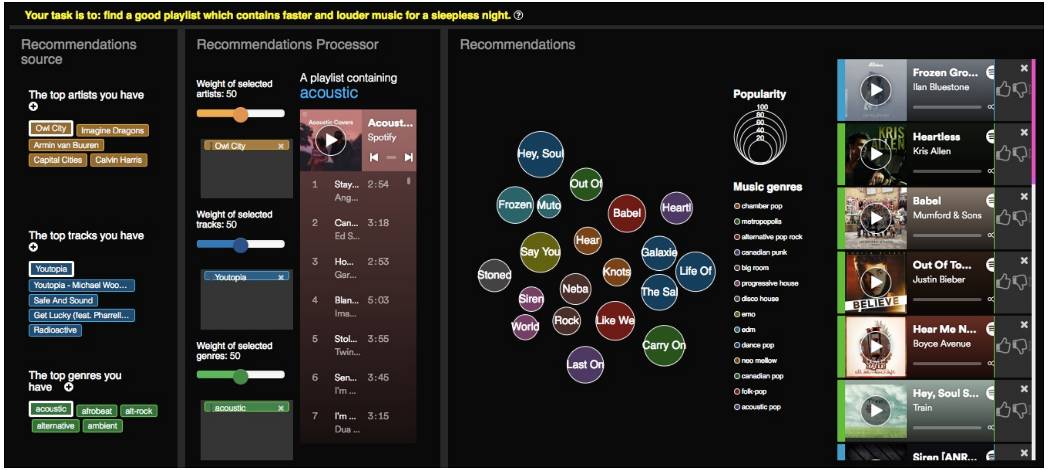


Katrien Verbert
Professor, KU Leuven

Yucheng Jin
Assistant Professor, HKBU

Expertise

- These are complex interfaces
- More effective for people who are more expert
- Still! Higher acceptance and equal cognitive load even for high complexity [Jin et al UMUAI'18]
  - Increases interaction
  - And accuracy



Take home messages

- Understanding is only <u>a first step.</u>
- Distrust of AI is healthy. But AI systems are not evil!
- Human-centered (X)AI benefits from questioning assumptions.
  - Long Term Program (10Y) (Co-I). ROBUST: Trustworthy AI-based Systems for Sustainable Growth.

Thank you

In a world where you can be anything, be kind.

Web references

- **Current algorithms.** Lighthouse reports, 20th December 2022.
- **Welfare scandal.** The Guardian, 5th February 2020.
- **Toeslagenaffaire.** Het Parool, 19th October 2021.
- **Top 400-Lijst.** Trouw, 22nd November 2022.
- **High-risk domains.** CACM, 28th July 2020.

References (1/3)

Amparore, E., Perotti, A., & Bajardi, P. (2021). To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods. *PeerJ Computer Science*, *7*, e479.

Azzopardi, L. (2021). Cognitive biases in search: a review and reflection of cognitive biases in Information Retrieval. *Conference on Human Information Interaction and Retrieval* (CHIIR) (pp. 27-37).

Barile, F., Najafian, S., Draws, T., Inel, O., Rieger, A., Hada, R., & Tintarev, N. (2021). Toward Benchmarking Group Explanations: Evaluating the Effect of Aggregation Strategies versus Explanation. *Perspectives workshop. ACM RecSys*.

Biran, O. and Courtenay, C. (2017). "Explanation and justification in machine learning: A survey." *IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8. No. 1. 2017.

Guidotti, R. et al. "A survey of methods for explaining black box models." *ACM computing surveys (CSUR)* 51.5 (2018) : 1-42 .

Bountouridis, D., Harambam, J., Makhortykh, M., Marrero, M., Tintarev, N., & Hauff, C. (2019). SIREN: A Simulation Framework for Understanding the Effects of Recommender Systems in Online News Environments. *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*.

Cau, F. M., Hauptmann, H., Spano, L. D., & Tintarev, N. (2023). Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. *IUI*.

Cau, F., Hauptmann, H., Spano, D., & Tintarev, N. Effects of AI and Logic-Style Explanations on Users’ Decisions under Different Levels of Uncertainty. *TiiS (*2023, to appear).

Draws, T., Liu, J., & Tintarev, N. (2020). Helping users discover perspectives: Enhancing opinion mining with joint topic models. *2020 International Conference on Data Mining Workshops (ICDMW)*, 23–30.

🏆 Draws, T., Inel, O., Tintarev, N., Baden, C., & Timmermans, B. (2022). Comprehensive viewpoint representations for a deeper understanding of user interactions with debated topics. *ACM SIGIR Conference on Human Information Interaction and Retrieval*, 135–145.

References (2/3)

Draws, T., Ramamurthy, K. N., Soares, I. B., Dhurandhar, A., Padhi, I., Timmermans, B., & Tintarev, N. (2023). Explainable Cross-Topic Stance Detection for Search Results. *CHIIR*.

Jin, Y., Tintarev, N. & Verbert, K. (2018). "Effects of individual traits on diversity-aware music recommender user interfaces." *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*.

Jin, Y., Tintarev, N., & Verbert, K. (2018). Effects of personal characteristics on music recommender systems with different levels of controllability. *Proceedings of the 12th ACM Conference on Recommender Systems*, 13–21.

Jin, Y., Tintarev, N., Htun, N. N., & Verbert, K. (2020). Effects of personal characteristics in control-oriented user interfaces for music recommender systems. *User Modeling and User-Adapted Interaction*, *30*, 199–249.

Been, K., Khanna, R. & Koyejo, O.O. (2016). “Examples are not enough, learn to criticize! Criticism for interpretability.” Advances in Neural Information Processing Systems.

Knijnenburg, Bart P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. "Explaining the user experience of recommender systems." *User modeling and user-adapted interaction* 22 (2012): 441-504.

Martin, F.J. (2009). Top 10 Lessons Learned Developing Deploying and Operating Real-world Recommender Systems. ACM Conference on Recommender Systems, 1–2

Najafian, S., Draws, T., Barile, F., Tkalcic, M., Yang, J., & Tintarev, N. (2021). Exploring User Concerns about Disclosing Location and Emotion Information in Group Recommendations. *ACM Hypertext*.

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38.

Mulder, M., Inel, O., Oosterman, J., & Tintarev, N. (2021). Operationalizing framing to support multiperspective recommendations of opinion pieces. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 478–488.

References (3/3)

🏆 Rieger, A., Draws, T., Theune, M., & Tintarev, N. (2021). This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (pp. 189-199).

Schellingerhout, Roan, Volodymyr Medentsiy, & Maarten Marx (2022). "Explainable Career Path Predictions using Neural Models." (2022). RecSys-in-HR Workshop. Recsys'22.

Sullivan E., et al. (2019). Reading news with a purpose: explaining user profiles for self-actualization. *Adjunct Proceedings of the Conference on User Modeling, Adaptation and Personalization (UMAP)*.

Lorenz-Spreen, Philipp, et al. "Boosting people's ability to detect microtargeted advertising." Scientific Reports 11.1 (2021): 1-9.

Tintarev, N., Rostami, S., & Smyth, B. (2018). Knowing the Unknown: Visualising Consumption Blind-Spots in Recommender System. *ACM Symposium on Applied Computing (SAC)*.

Tintarev, N., & Masthoff, J. (2022). Beyond Explaining Single Item Recommendations. In: Ricci, F., Rokach, L., Shapira, B. (eds) *Recommender Systems Handbook*. Springer, New York, NY.

Tintarev, N., & Masthoff, J. Evaluating the effectiveness of explanations for recommender systems: Methodological issues and empirical studies on the impact of personalization. *User Modeling and User-Adapted Interaction*, *22 (2012)*, 399–439.

Tintarev, N., Kang, B., Höllerer, T., & O'Donovan, J. (2015). Inspection Mechanisms for Community-based Content Discovery in Microblogs. *Joint Workshop on Interfaces and Human Decision Making for Recommender Systems in Conjunction with Recsys.*, *1438*, 21–28.

Tintarev, N., O'Donovan, J., & Felfernig, A. (2016). Introduction to the special issue on human interaction with artificial advice givers. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *6*(4), 1-12.