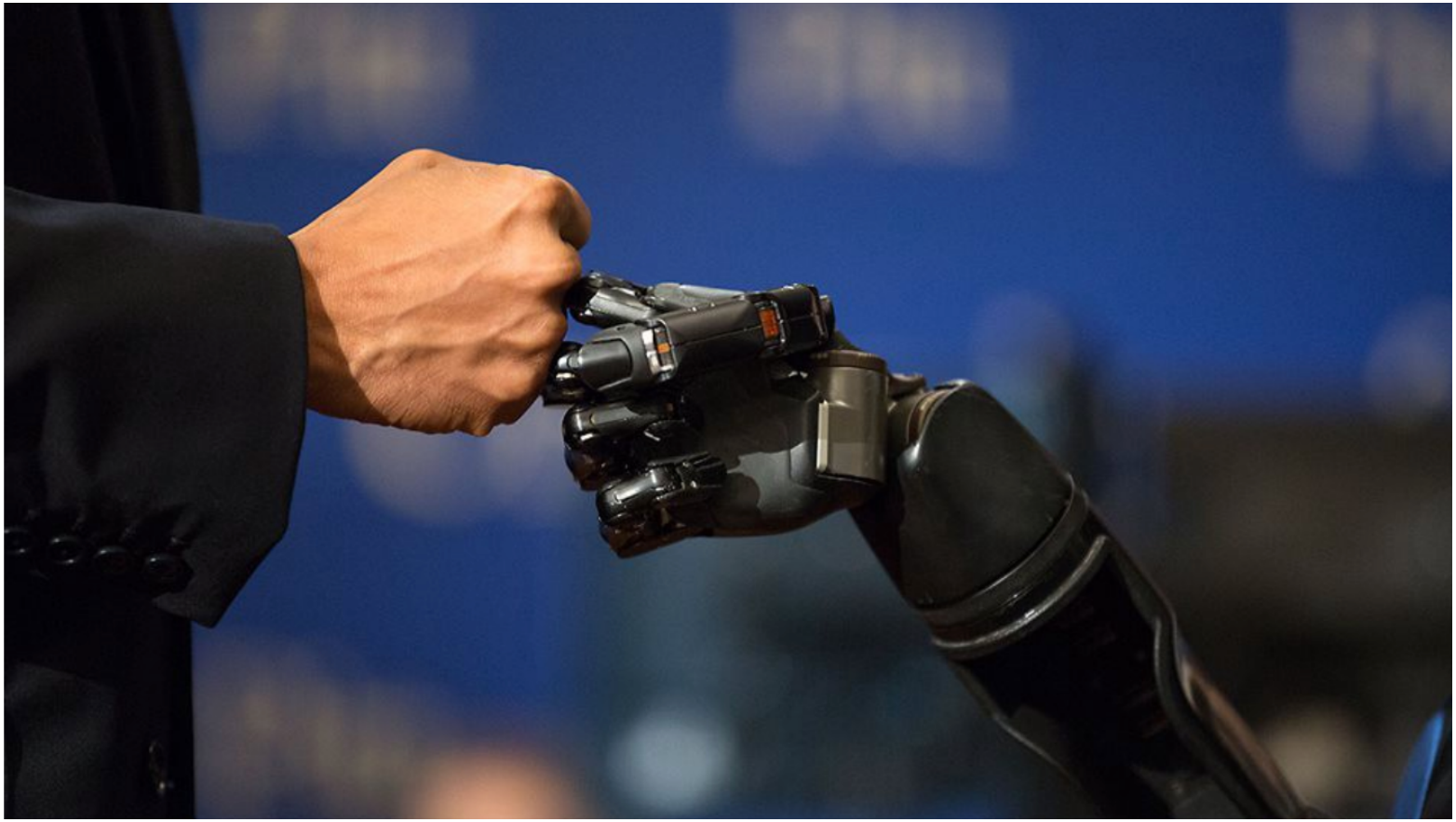
Trust and Explanation

Applied Machine Learning Days EPFL 2019

School of Computer and
Communication Sciences
EPFL
Pearl Pu

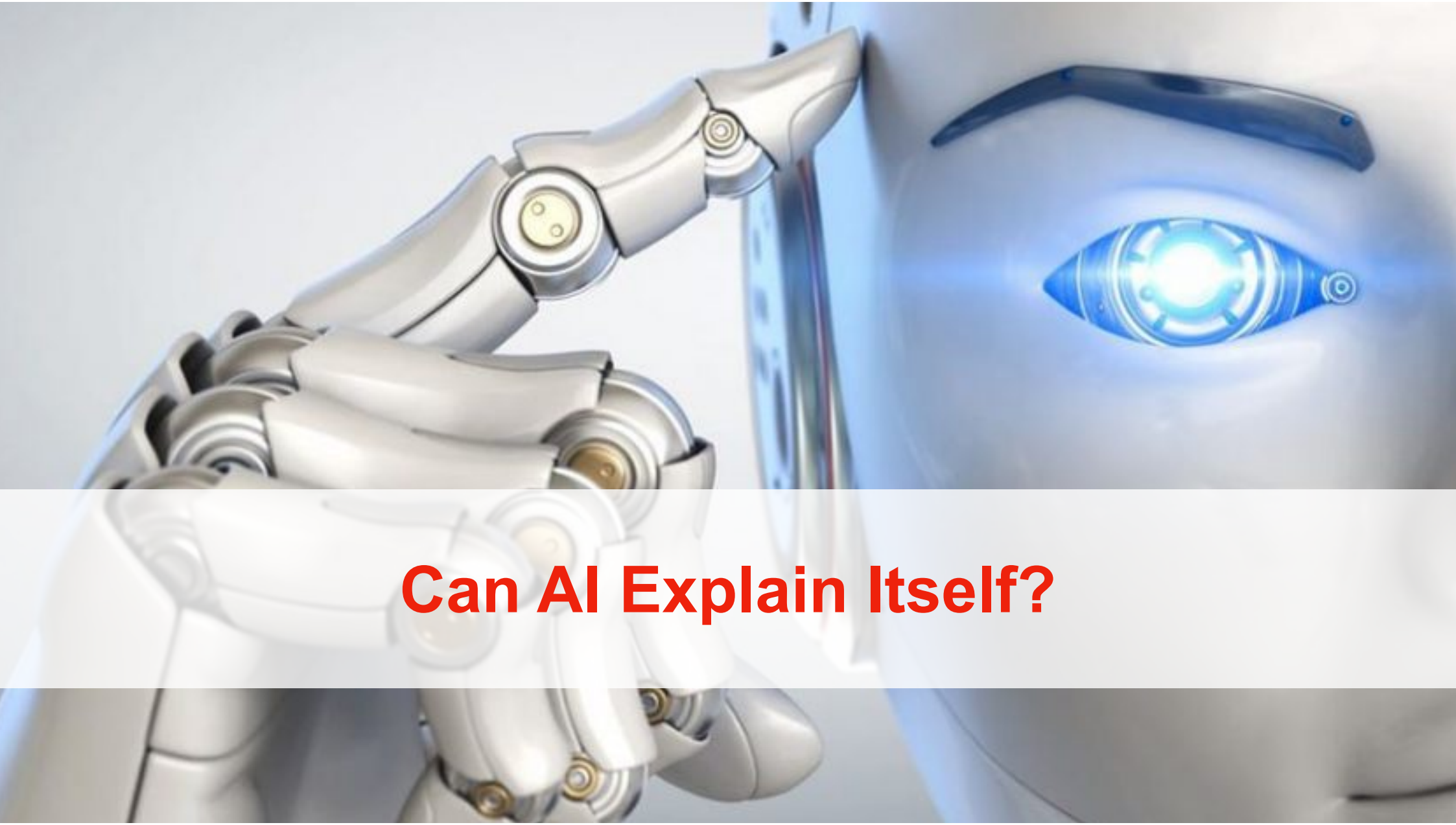
What is trust?

Behavior?
Belief?
Attitude?
Intention?
Emotion?



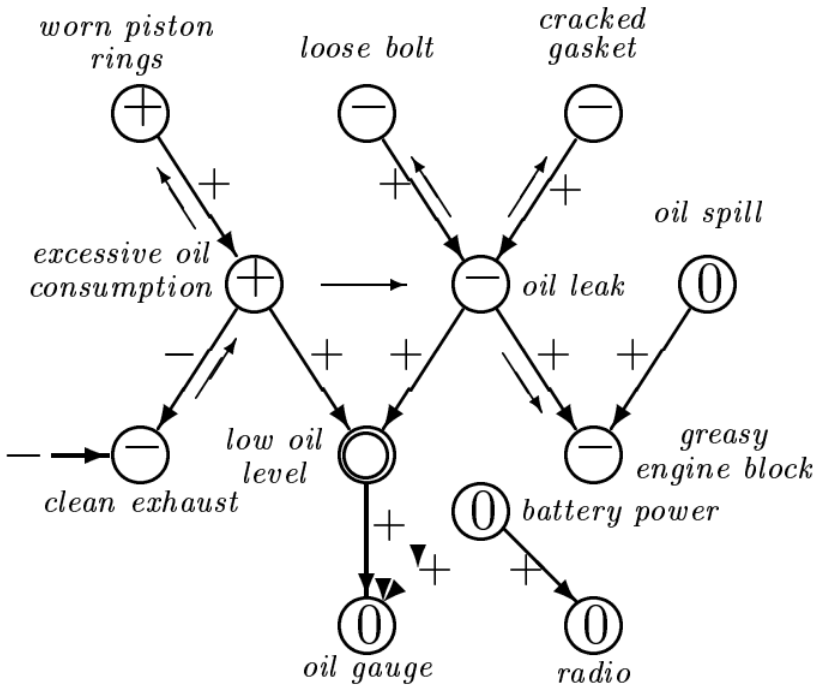
Trust Building

How do we do it?



Can AI Explain Itself?

Mismatch of research
methodology and ***outcome***



Qualitative influence of greasy engine block on worn piston rings: Greasy engine block is evidence for oil leak. Oil leak and excessive oil consumption can each cause low oil level. Oil leak explains low oil level and so is evidence against excessive oil consumption. Decreased likelihood of excessive oil consumption is evidence against worn piston rings. Therefore, greasy engine block is evidence against worn piston rings.

What's wrong?

- Users do not understand complex causal explanations
- Users prefer proximal causes over distal ones
- Users understand causality by "undo" simulation
- Sometimes causality doesn't exist

Correct method: User-centric research

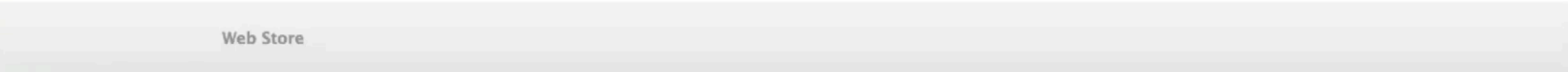
- Literature research - explanation in social science, philosophy, psychology
- Users' perception of AI - expectations, fears, mental models, attitudes, habits
- Build user models and requirements
- Design, prototype, and test

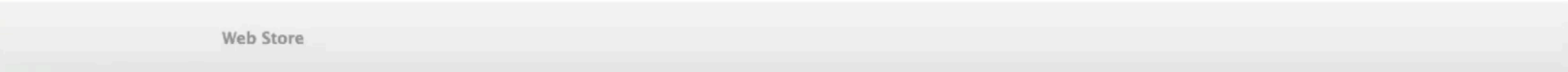
Explain ML & high dimensional space

Video of students' work


AI FOR
MUSIC
HKU
EPFL


AI FOR
MUSIC
HKU
EPFL


Web Store


Web Store

The Problem: Spam Detection

- **How to predict whether or not an email is spam based on the words contained in the email?**
- About 4.5 billion spam emails were sent each day in 2017
- Almost every major email provider (Gmail, Yahoo, Hotmail) uses a spam detection algorithm

https://www.statista.com/statistics/420391/spam-email-traffic-share/

The Problem: Spam Detection

- **How to predict whether or not an email is spam based on the words contained in the email?**
- About 4.5 billion spam emails were sent each day in 2017
- Almost every major email provider (Gmail, Yahoo, Hotmail) uses a spam detection algorithm

https://www.statista.com/statistics/420391/spam-email-traffic-share/

Trust-Inspiring Interface Design Principles

- Overview-detail techniques (infoviz)
- Transformation (AI)
- ***Contrasting*** as explanation (SS)
- Relatedness (User research)

Conclusion

- XAI is likely to make AI happen
- Follow the user-centric design techniques
- Provide explanation according to users' needs

Future topics for XAI

- Visualizing hidden spaces
- Explanation as conversation
- Chatbot that responds to users' emotions
- Privacy, security, ethics, etc.

Reference

Li Chen and Pearl Pu. **Trust Building in Recommender Agents**. In *Proceedings of the Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces at the 2nd International Conference on E-Business and Telecommunication Networks (ICETE'02)* , Reading, UK, October 3-7, 2005, pp.135-145.

Pearl Pu and Li Chen. **Trust Building with Explanation Interfaces.** In *Proceedings of the 11 th International Conference on Intelligent User Interface (IUI'06)* , Sydney, Australia, pages 93-100, January 29-February 1, 2006.

Pearl Pu and Li Chen. **Trust-Inspiring Explanation Interfaces for Recommender Systems**. Journal of Knowledge Based Systems, Elsevier Publishers, Volume 20, Issue 6, Pages 542-556, Auguest 2007.

Pearl Pu, Li Chen and Rong Hu. **A User-Centric Evaluation Framework for Recommender Systems**. In Proceedings of the 5th ACM Conference on Recommender Systems (RecSys'11), pages 157 – 164, Chicago, IL, USA, October 23 – 27, 2011.

Marek J Druzdzel. Qualitiative verbal explanations in bayesian belief networks. AISB QUARTERLY, pages 43–54, 1996.

Conversational processes and causal explanation. Psychological Bulletin , 107(1):65{81, 1990.

Thomas Eiter and Thomas Lukasiewicz. Complexity results for structure-based causality. Articial Intelligence, 142(1):53{89, 2002.

J Herlocker, J Konstan, and J Riedl. Explaining collaborative filtering recommendations. In Computer Supported Cooperative Work (CSCW) , 2000.

Rashmi Sinha and Kirsten Swearingen. The role of transparency in recommender systems. In CHI EA , 2002.

F. Y. Tzeng and K. L. Ma. Opening the black box - data driven visualization of neural networks. In IEEE Visualization , pages 383–390, 2005.

Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In NAACL-HLT , 2016.

Image Reference

Towards AI Transparency: Four Pillars Required to Build Trust in Artificial Intelligence Systems
https://towardsdatascience.com/towards-ai-transparency-four-pillars-required-to-build-trust-in-artificial-intelligence-systems-d1c45a1bdd59

Mother bird feeding baby bird
https://www.pinterest.ch/pin/491596115550772749/?lp=true

Video credits

- Tobia Albergoni
- Matteo Yann Feo
- Rong Hu