Data Driven & Goal Driven XAI

HaoCheng Ho

ID: 0160617622

24/11/2020

Outline

- Motivation
- Background
- Review Methodologies
- Systematic Literature Review
- Result
- Conclusion

Motivation

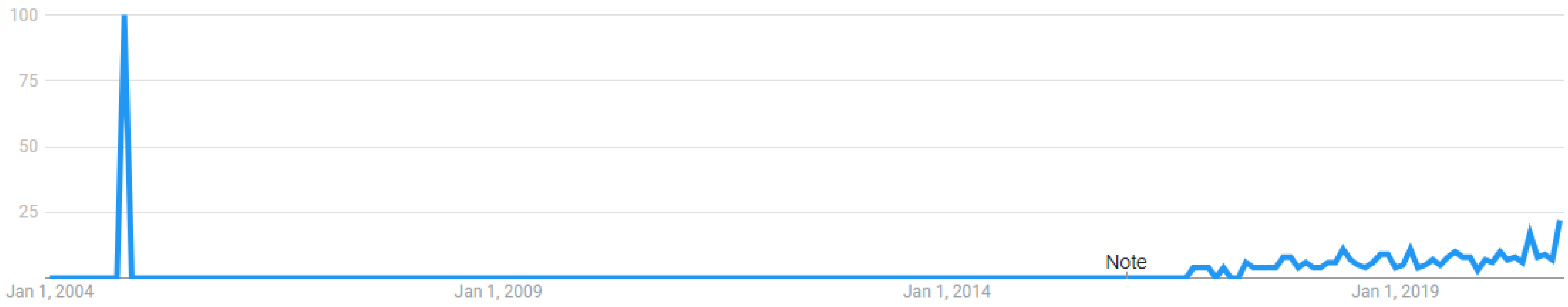
- Global investment on AI:
  12 billion USD (2017) to 52.2 billion USD (2021)
- Revenues from the AI market worldwide:
  480 billion USD (2017) to 2.59 trillion USD (2021)
- AI is an inescapable technology among the Gartner:
  "Top 10 Strategic Technology Trends for 2018"

Motivation (contd.)

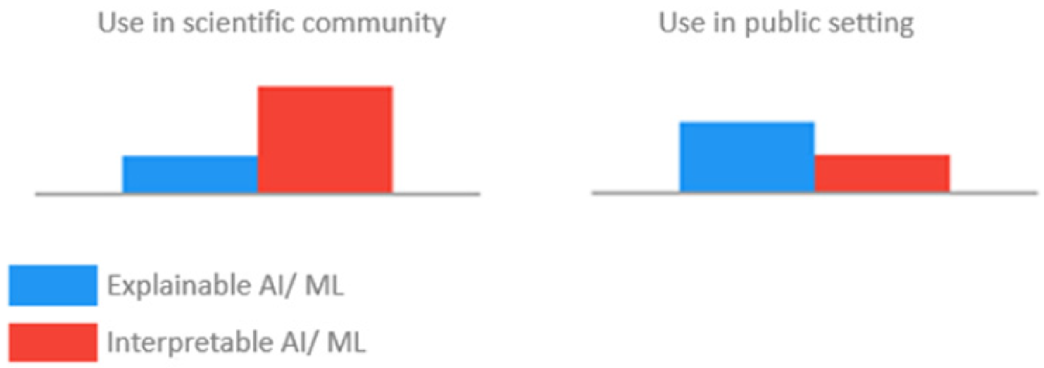
- AI is already present in our daily life
  (Netflix, Amazon, Facebook, Google)
- Important to know the reasoning behind decisions
  (ex: disease diagnosis by AI)
- “Right to explanation” in the <u>**G**</u>eneral <u>**D**</u>ata <u>**P**</u>rotection <u>**R**</u>egulation (**GDPR**), which comes into effect on May 25, 2018 across the EU
- AI algorithms lack transparency (especially ML algorithms)
- E<u>**x**</u>plainable <u>**A**</u>rtificial <u>**I**</u>ntelligence (**XAI**) makes AI more “transparent”

Background

- Google Trends result for the term “Explainable Artificial Intelligence”



- Google Trends results, comparing “Explainable” and “Interpretable” according to the context.



Background (contd.)

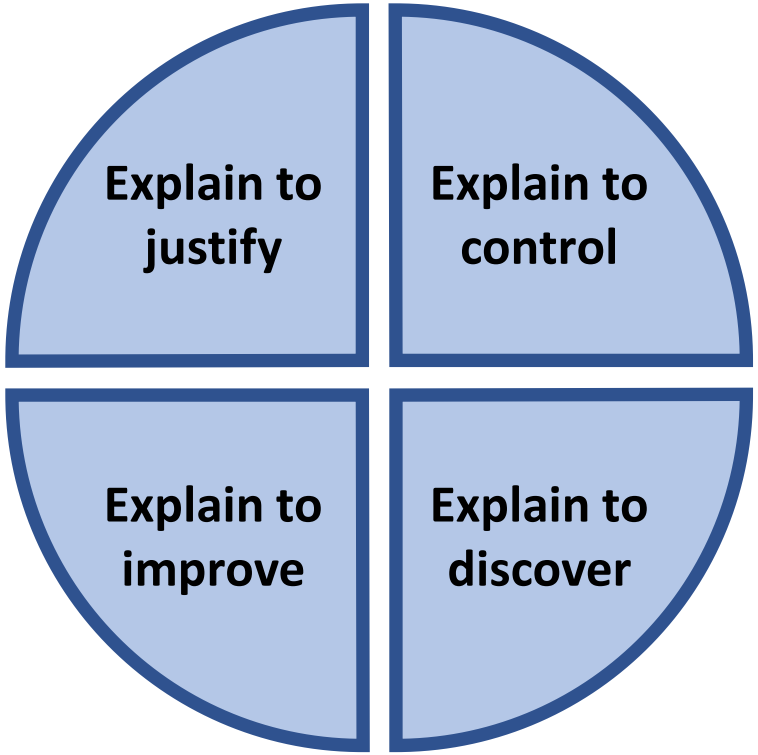
- Recap: What is XAI?
  - Underlying causes to its decisions are understandable by humans.
  - Two types: data-driven & goal-driven

- Data-driven XAI (explaining black-box algorithms)
  - Interpret the decision of ML algorithm given the data used as an input.
- Goal-driven XAI (explainable agency)
  - Explain the actions and reasons leading to their decisions.

Background (contd.)

- Why do we need XAI? (examples)
  - Commercial benefits
  - Ethics concerns
  - Regulatory considerations
  - Essential for users to trust the AI

- 4 categories of reasons:
  - Explain to justify
  - Explain to control
  - Explain to improve
  - Explain to discover

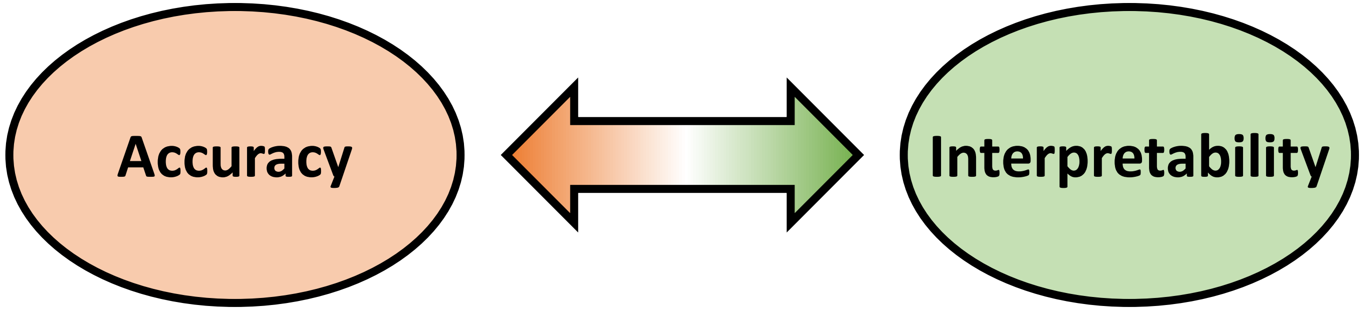


Background (contd.)

- What are the XAI application domains?
  - Transportation
    - Automated/Autonomous Vehicles
  - Healthcare
    - Medical diagnosis
  - Legal
    - Criminal justice
  - Finance
    - Wealth-management
    - Investment advice
  - Military
  - Other domains: Cybersecurity, Education, Entertainment, Government, etc.

Background (contd.)

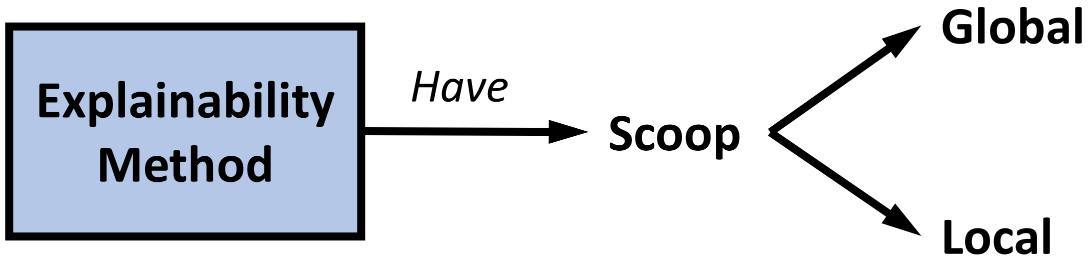
- What are the technical challenges of XAI?
  - We could ask ourself the following:
    - Why the use of XAI is not systematic?
    - Why is not everyone using XAI?
  - Black-box, for example Deep Neural Networks (DNN)
  - “Modern” ML algorithm gets more and more complex
  - For the same set of input, complex ML algorithms can produce different models and the accuracy of the results remains the same.
- There is a trade-off between accuracy and interpretability



Review Methodology

- Complexity related methods
  - More complex -> more difficult to interpret/explain
- “Low” complexity
  - For example: we create a “white-box” AI/ML
  - Simple to explain
    (intrinsic interpretable models)
  - Trade-off between “Accuracy” and “Interpretability”
- “High” complexity
  - This means “black-box” AI/ML
  - Reverse engineering to provide explanations
    (post-hoc explanations by example)
  - Has high accuracy

Review Methodology (contd.)

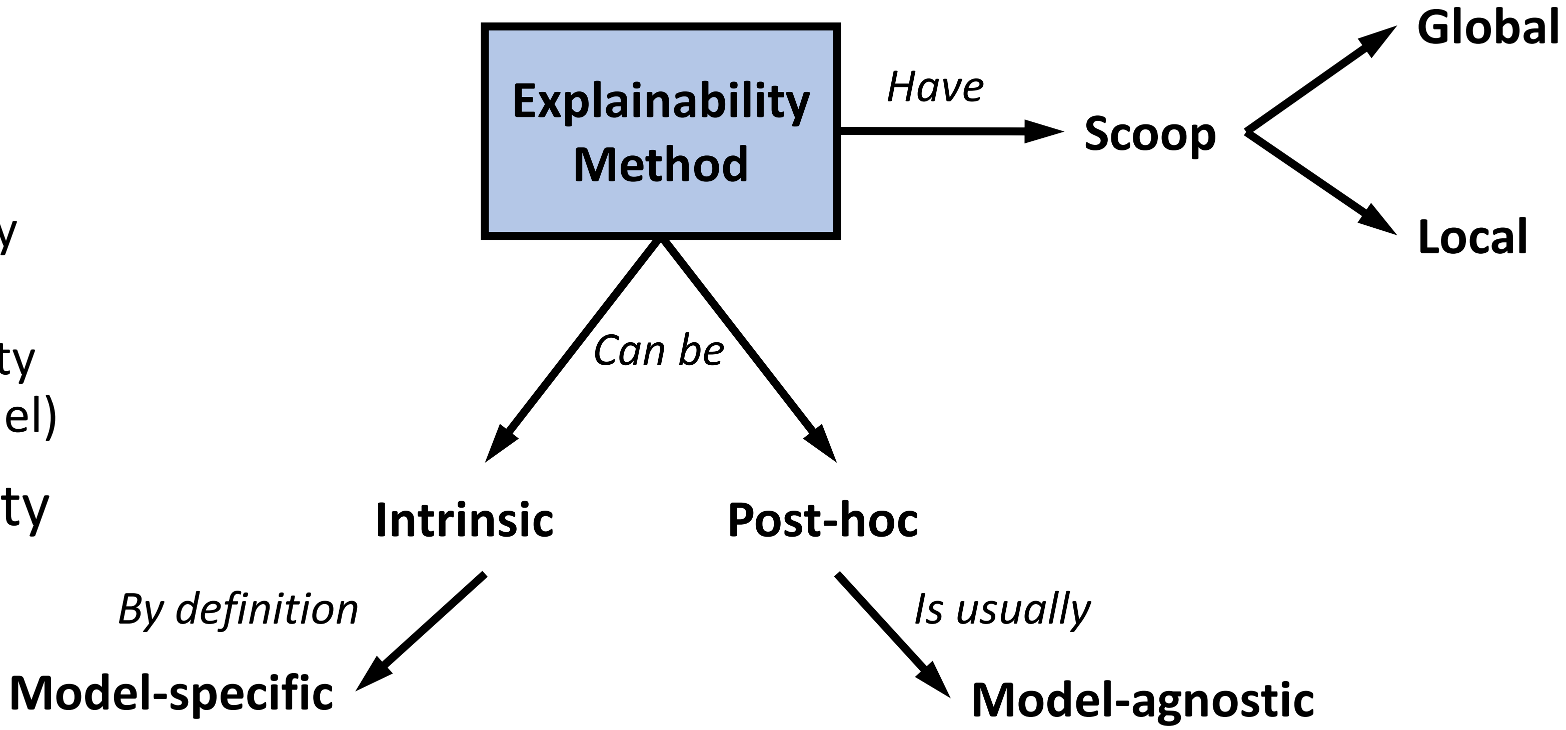


- Scoop related methods
  - Global interpretability (understand the entire model)
  - Local interpretability (understand a single prediction)

- Global interpretability
  - For example: climate change model
  - But limited in predictability if we want interpretability.
- Local interpretability
  - For example: image classification model
  - But limited in interpreting the whole model

Review Methodology (contd.)

- Model related methods
  - Model-specific interpretability (limited to a specific model)
  - Model-agnostic interpretability (not tied to the type of a model)
- Model-specific interpretability
  - Explainable by definition ("low" complexity)
  - More accurate explanation
- Model-agnostic interpretability
  - Explaining using:
    Visualization, knowledge extraction, influence methods and example-based explanation
  - Less accurate explanation



Review Methodology (contd.)

- How should the AI model be explained to humans?
  - Challenge of designing XAI:
    Communicate a complex computational process to human (with ML expertise?)

- Human-like explanations
  - Three major findings:
    - Why event A happened instead of event B? And not why event A actually happened.
    - Focuses only on 1 or 2 possible causes. (Not all the causes forming the decision.)
    - Explanations are social conversation to transfer knowledge. (Same mental model, explainer & explainee)
- Human-friendly explanations
  - Through simulations, chain of reasoning, multiple examples

Review Methodology (contd.)

- There are three distinct explanation phases:
  - Explanation Generation
  - Explanation Communication
  - Explanation Reception

- “It is not enough to just explain the model, the use has to understand it.”
  - Give the user the possibility to ask questions to the AI model
  - Thus, we need an interaction between human and machine.

Systematic Literature Review

- Selection criteria:
  - Recent Paper
    - (2008 - 2018)
  - Relevance
  - Primary Study
  - Accessibility
    - IEEExplore, Science Direct, ACM, and Google Scholar
  - Explainable Agency
    - Goal-driven XAI
  - Singularity/Originality
  - Explanation as a *Communicative Action*
- 62 papers are selected according to these criteria

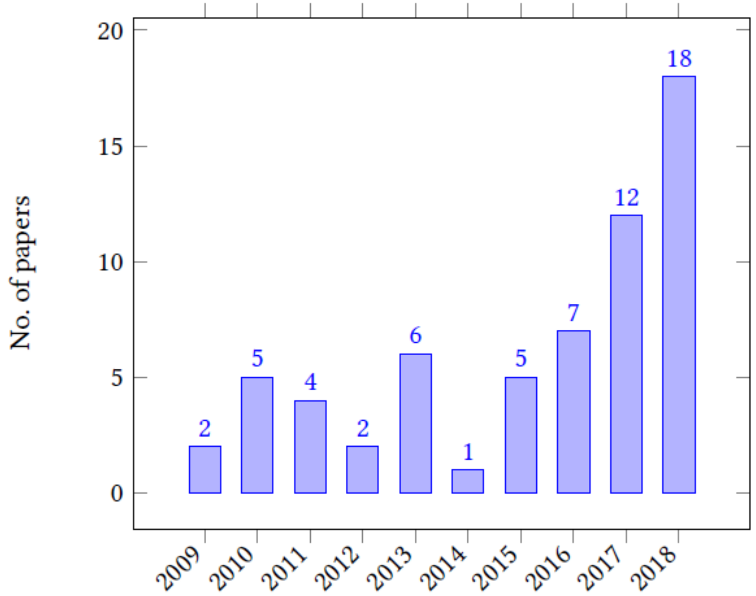
Systematic Literature Review (contd.)

- **S**tructured **R**esearch **Q**uestion**s** (**SRQs**)
  - SRQ1: Demographics
  - SRQ2: Application scenarios
  - SRQ3: Drives (needs)
  - SRQ4: Social science and psychological background
  - SRQ5: Design
  - SRQ6: Dynamics (context-aware, user-aware)
  - SRQ7: Presentation
  - SRQ8: Evaluation/Framework
  - SRQ9: Future challenges

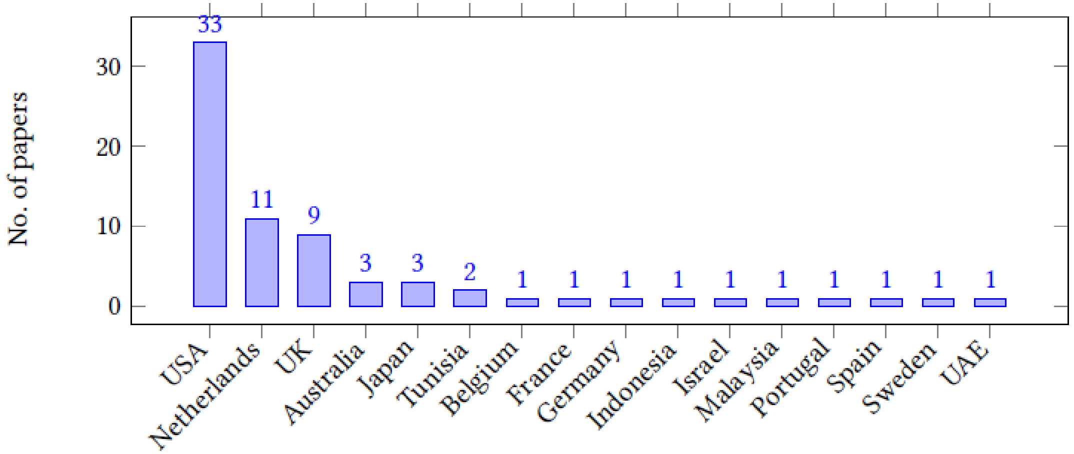
Result

- SRQ1: Demographics

- Increasing growth over the last 5 years

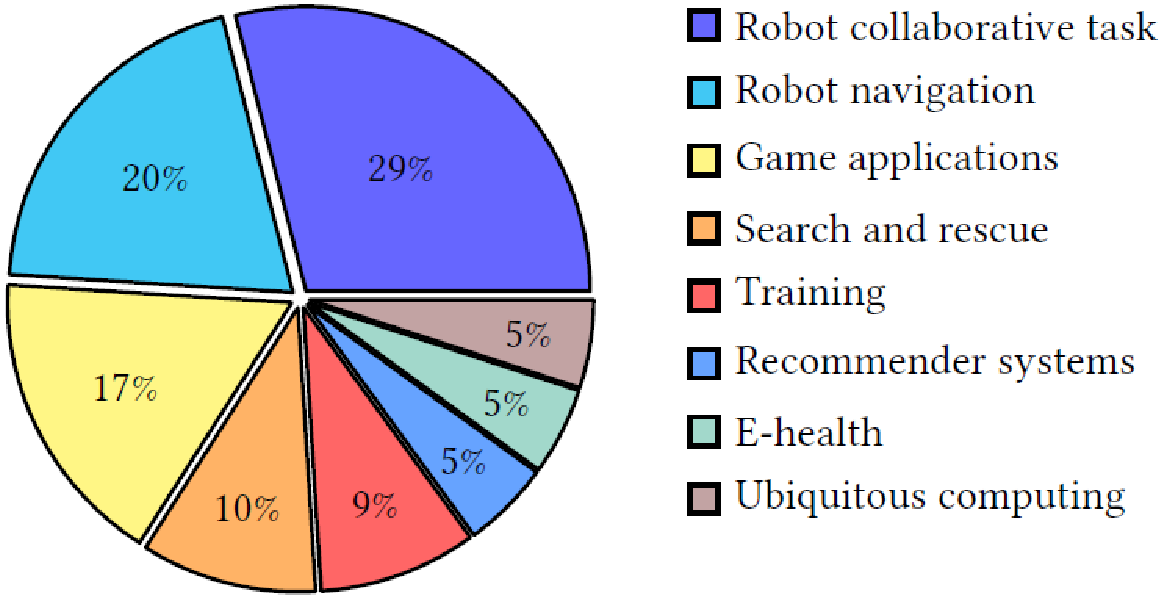


- USA, Netherlands and UK
- European research on this subject might increase (GDPR)



Result (contd.)

- SRQ2: Application scenarios

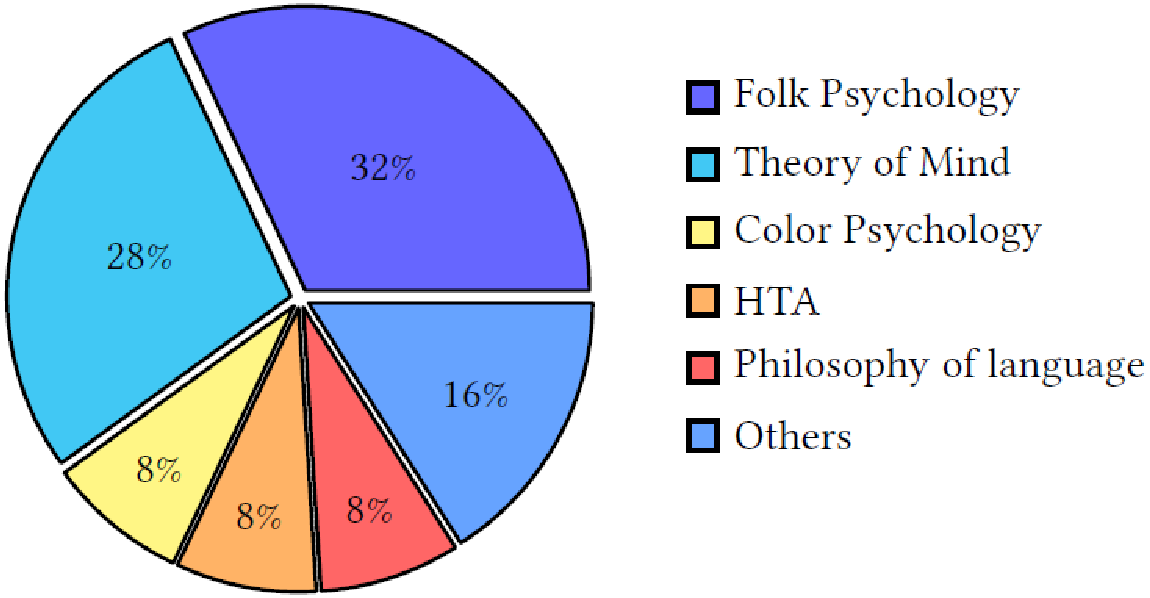


Result (contd.)

- SRQ3: Drives (needs)
  - Transparency
  - Trust
  - Collaboration
  - Intent communication
  - Control
  - Education
  - Debugging

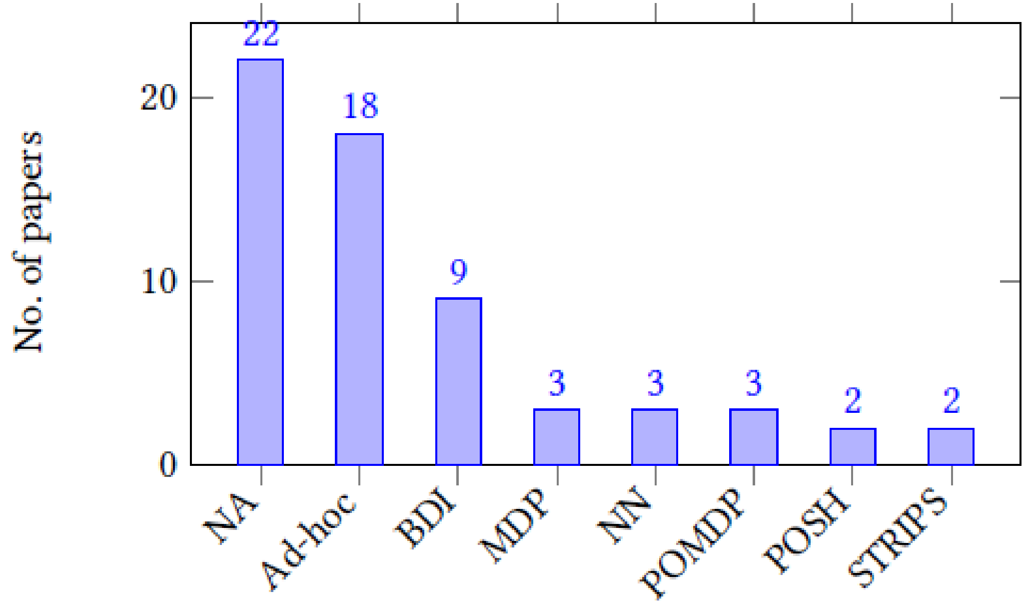
Result (contd.)

- SRQ4: Social science and psychological background



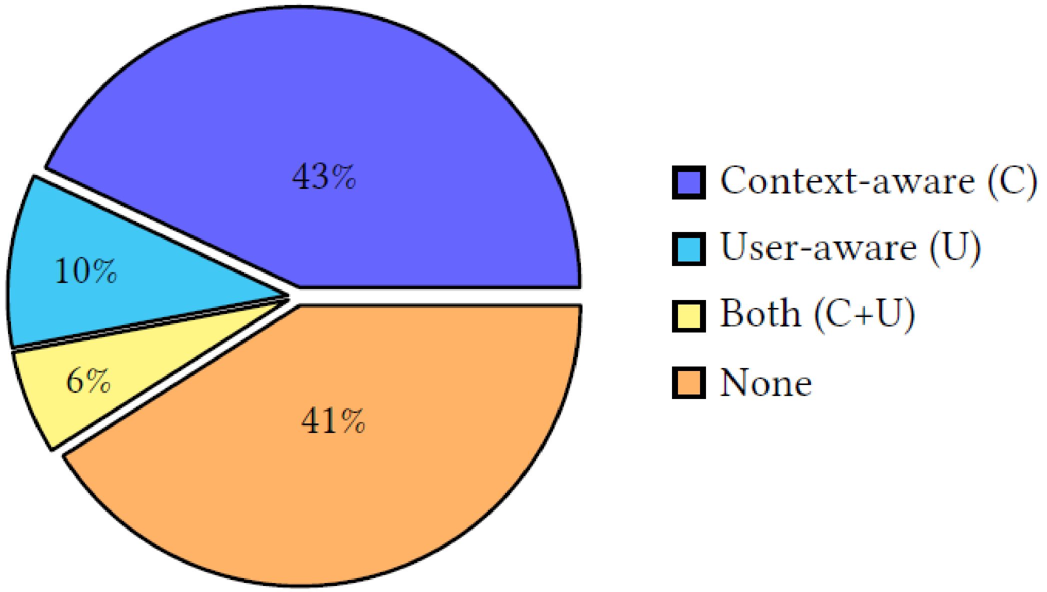
Result (contd.)

- SRQ5: Design
  - 22 NA (not available)
  - 18 Ad-hoc (customized methods)
  - 9 BDI (Belief, Desires, Intentions)
  - 3 MDP (Markov Decision Process)
  - 3 NN (Neural Networks)
  - 3 POMDP (Partially Observable Markov Decision Process)
  - 2 POSH (Parallel-rooted-ordered Slip-stack Hierarchical Action Selection)
  - 2 STRIPS (Stanford Research Institute Problem Solver)



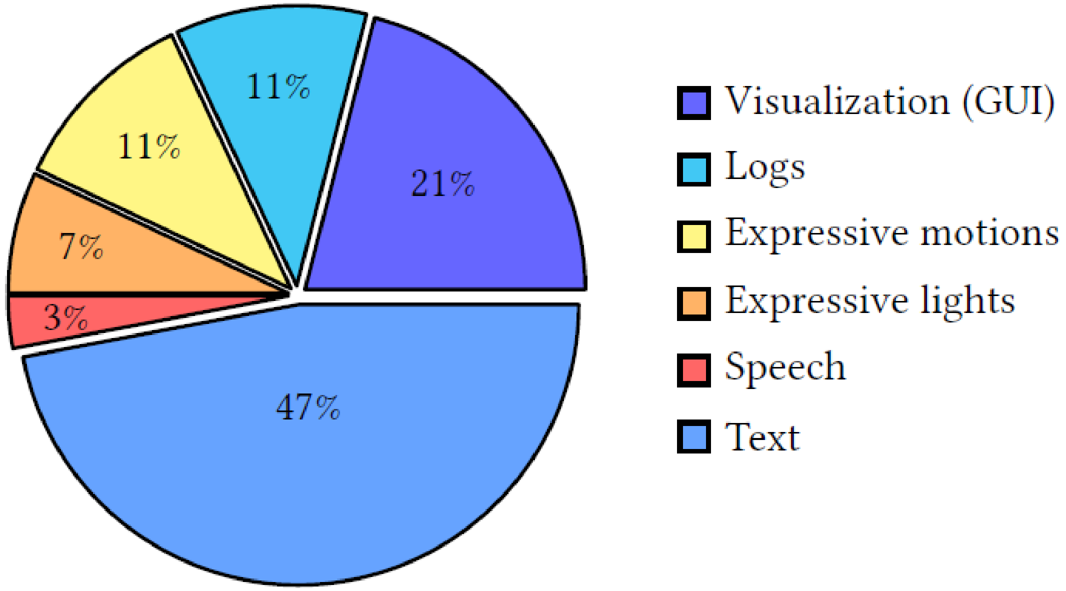
Result (contd.)

- SRQ6: Dynamics (context-aware, user-aware)



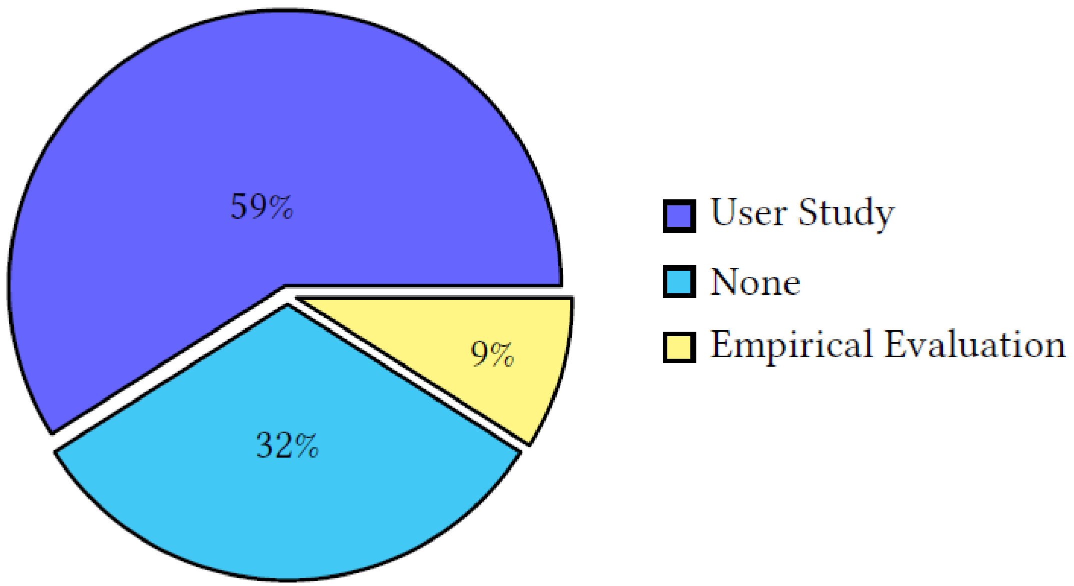
Result (contd.)

- SRQ7: Presentation



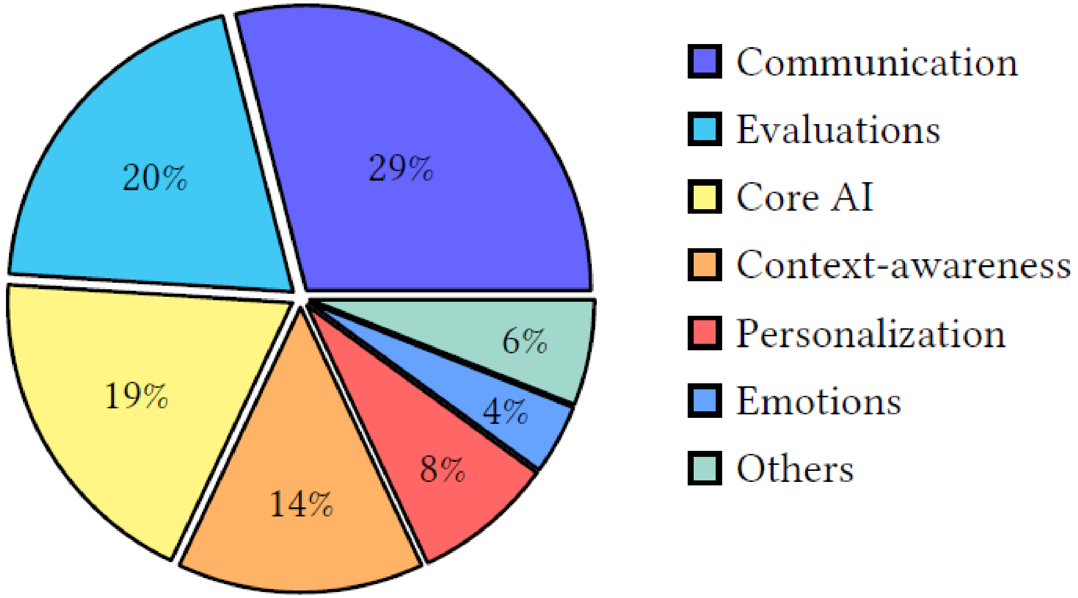
Result (contd.)

- SRQ8: Evaluation/Framework



Result (contd.)

- SRQ9: Future challenges

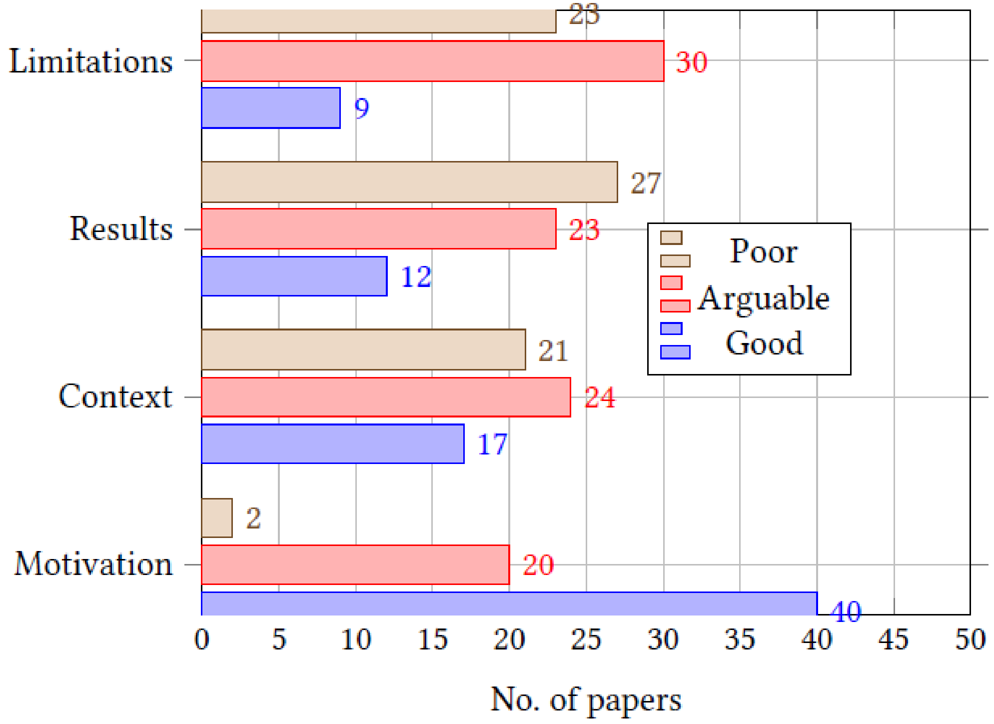


Quality Criteria Assessment

- Quality criteria:
  - Motivation
  - Context
  - Results
  - Limitations

- Graded at 3 levels:
  - “Good”, “Arguable” and “poorly presented”

- Each paper was evaluated by at least 2 reviewers and their results averaged.



Conclusion

- We focused on the 5 W's (What, Who, When, Why, Where) and How to cover all aspects related to XAI
- This survey reviewed a portfolio of explainability approaches and organized them from different perspective.

Future works

- Considerable effort will be required in the future
  to tackle the challenges and open issues with XAI
- Human’s role is not sufficiently studied in existing XAI
- Most of the existing works focus on interpretability in ML
  - But this is just one type of AI
- In the era of Internet of Things (IoT)
  - We also need machine-to-machine XAI
  - Likely that future XAI approaches,
    provide both kinds of explanation

Thank you for listening! ☺

References (Papers)

- Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Framling. 2019. **Explainable agents and robots: Results from a systematic literature review.** In 18th International Conference on Autonomous Agents and Multiagent Systems
- A. Adadi and M. Berrada, **“Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI),”** IEEE Access, vol. 6, pp. 52138–52160, 2018.