Résumé du Workshop XAI - AAAI21

Séminaire équipe SyCoSMA
Rémy Chaput - 25/02/2021

Contexte

- Explicabilité : quésako?
  - Mot “fourre-tout” : intepretabilité, compréhensibilité, transparence, etc.
  - Plusieurs objectifs
  - intuitivement, “faire en sorte que l’utilisateur comprenne le fonctionnement / les décisions de l’algo” (modèle mental)
    - (quel utilisateur ?)
- Domaine riche parmi les sciences sociales
  - Mais ... lien pas toujours fait.

Quelques techniques d’XAI

- Explaining Agent’s Behavior (Ofra Amir)
  - Identifier les états ‘importants’ (maximisent la différence de récompenses)
  - Expliquer seulement ces états ; pas tous les états
    - Ex: Pacman (tranquille/isolé vs encerclé/coincé)
  - Comment expliquer plusieurs agents ?
    - Comparer les stratégies et se focaliser sur les désaccords

Quelques techniques d’XAI (2)

- Cartes de saillance
  - Afficher visuellement les pixels qui sont les plus importants pour la prise de décision dans une situation donnée
  - Pas toujours compris par l’utilisateur final (non-expert) !
  - Abondamment utilisé, en Computer Vision et en RL (du moins dans les jeux et autres scénarios visuels)

Quelques techniques d’XAI (3)

- Contre-factuels
  - “Pourquoi X et pas Y ?”
  - D’après sciences sociales, probablement une bonne forme d’explication
  - Abondance d’articles à ce sujet

Quelques techniques d’XAI (4)

- Bandits for learning to explain from explanations (Behrens, Teso et Mottin)
  - Apprentissage d’explications (plutôt rare !)
    - Proposition d’une explication
    - Utilisateur dit si c’était utile ou non
    - => Centré utilisateur (mais pas encore tout à fait interaction complète)
  - Kernels spécifiques pour décrire comment généraliser les explications dans l’espace
    - Apparemment coûteux en calcul, et sensible

Quelques réflexions sur XAI

- Explaining AI Well Includes Explaining AI Fairly (Margaret Burnett)
  - Certaines questions réservées aux chercheurs IA => plus de transparence dans les processus par ex.
  - Mais besoin des sciences sociales pour d’autres !
  - Prendre en compte des situations variées
    - ex: utiliser des persona

Quelques réflexions sur XAI (2)

- Explainable, Normative, Justified Agency (Pat Langley)
  - Plusieurs sens à “explication”
    - Une structure (information)
    - Un processus qui construit la structure
    - Interpretive Explanations = construire
    - Communicative Explanations = transférer
  - 3 types de contenu
    - Structural accounts (segments de route jusqu’au but)
    - **Preference accounts** (pourquoi ça et pas autre)
    - Process accounts (déroulement du processus)

Quelques réflexions sur XAI (3)

- Human Centered XAI (Shane Mueller)
  - Explication passe par le “Self-Explaining”
  - Transformation de connaissance ; faire sens
  - Propose également des critères pour noter les systèmes
    - Null ; Features ; Succès ; Mécanismes ; Raisonnement ; Échecs ; Comparaisons ; Diagnostics d’échecs
    - Plus de détails : https://arxiv.org/pdf/2102.04972.pdf

Explanation: what does it mean for humans, for machines, for man-machine interactions?

Rémy Chaput[1], Amélie Cordier[3], Alain Mille[1,2]

February 9, 2021

[1] Université de Lyon, Université Lyon1, LIRIS UMR CNRS 5205
[2] Coexiscience, Lyon, France
[3] Lyon-iS-Ai, Lyon, France

Introduction

Context

There is an important, growing, impact of Artificial Intelligence applications on society.

Even more so with Machine Learning and Deep Learning, but not limited to them.

$\Longrightarrow$ Urgency of research on Explainable Artificial Intelligence

Our claim

Although numerous works begin to take into account end users, they mostly do so by pre-constructing explanations for specific audience profiles.

We posit that explanation is a complex process, and must be co-constructed with the users, within their own context: task, responsibilities, knowledge, mental model of the system, etc.

Explanations: what does it mean for humans?

Key observations

- Explaining is a continuous process
- Explaining is a co-adaptive process
- Explanation must be triggered
- We should facilitate self-explanation
- Explanation is an exploration
- Contrast situations can be explained differently

From Hoffman et al. 2018

Explanations: what does it mean for AI?

Symbolic based expert systems

The symbolic AI theoretically allows to provide the reasoning steps pursued to propose an answer to a request.

The origin and the justification of the knowledge having been used to answer the request are not generally available.

The expert knowledge is not questionable and the explanation is reduced to the trace of the reasoning in a form often only useful for debugging.

See, e.g., Swartout et al. 1985; Chakraborti et al. 2020

Machine Learning

Many methods and techniques to produce explanations; however, the explanation is often considered an artifact and not a process.

In particular, even when toolkits consider several audiences, the ability to construct the explanation by exploring is not focused.

Robotics and IoT

In the field of social robotics and of collaborative robotics, robots are designed to interact with human beings. They have to be designed to ensure that humans trust robots they are working with, to make sure that they can operate safely.

In the field of Internet of Things (IoT), many people express concerns about safety, security, data privacy and resilience. But, it could be much more difficult to explain IoT systems because of the complexity emerging from a vast network of simple devices.

See, e.g., Garcia-Magarino et al. 2019

Bio-inspired systems

Bio-inspired AI methods do not provide symbolic explanations but offer simulators to become familiar with emerging behaviors.

The use of these methods is slowed down by the difficulties of explanation and research communities are active in associating information to key moments of the behavior as a possible support for explanations.

Towards an UXAI model

Rationale

The question of explanation arises in any decision support agent. Artificial Intelligence techniques can help implementing a dynamic explanation process, constructed and conducted by the user of the system.

It is possible to design a dynamic explanation process, based on explanatory agents able to learn in co-construction with users how to explain the behavior of design support agents.

Today

- Researchers build decision models, from their expert knowledge and collected data. These are General Models (GMs).
- Designers use published GMs to build Applied Models (AMs), operationalized in the context of a task.
- Users use the AM in an application as a support for their activity. The interaction traces provide data that can be used to improve the AM. Researchers can also collect data to build the data corpuses that are fed into their own GM.

Tomorrow

- Researchers publish a GM and an associate general model of explanation (UXAI-GM), which contains the necessary knowledge to explain.
- Designers integrate into the explanations applied models (UXAI-AM) the possibility of co-constructing explanations with the users. A first UXAI-AM is deployed along with the AM, and both evolve synchronously.
- Users have not only an application but also an explanation agent with it. They can learn to understand the application's behavior, for example by self-explaining. The explanation agent stores in its memory the various explanations and their contexts, as situated explanations which can be reused.

Conclusion

Main issues for UXAI?

1. ethical problems of misuse;
2. respective responsibilities of AI agent / User;
3. economic consequences of a refusal of use as a precautionary principle.

What does UXAI can provide?

1. the user associates a self-explanation with the explanation provided by the agent;
2. the resulting explanation is contextualized with the help of the user for a similar usage situation;
3. contradictions between design logic and usage logic feed the agent design loop.

Thank You

Any questions?

References

Chakraborti, Tathagata, Sarath Sreedharan, and Subbarao Kambhampati (Feb. 2020). "The Emerging Landscape of Explainable AI Planning and Decision Making". en. In: *arXiv:2002.11697 [cs]*. arXiv: 2002.11697.

Garcia-Magarino, Ivan, Rajarajan Muttukrishnan, and Jaime Lloret (2019). "Human-Centric AI for Trustworthy IoT Systems With Explainable Multilayer Perceptrons". en. In: *IEEE Access* 7, pp. 125562–125574.

Hoffman, Robert R, Gary Klein, and Shane T Mueller (2018). "Explaining explanation for "explainable ai"". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 62. 1. SAGE Publications Sage CA: Los Angeles, CA, pp. 197–201.

References ii

Swartout, William and Johanna D Moore (1985). "Explainable (and Maintainable) Expert Systems". In: *Proceedings of the 9th International Joint Conference on Artificial Intelligence*. Vol. 1. Los Angeles.