The Challenges for Interpretable AI for Well-being -Understanding Cognitive Bias and Social Embeddedness-

Takashi Kido
Preferred Networks, Inc.
kido@preffered.jp

Keiki Takadama
The University of Electro-Communications
keiki@inf.uec.ac.jp

Abstract

In this AAAI Spring symposium 2019, we discuss interpretable AI in the context of well-being AI. Interpretable AI is an artificial intelligence methods and systems, of which outputs can be easily understood by humans. Especially in the human health and wellness domains, making wrong predictions may lead to critical judgements in life or death situations. AI based systems must be well-understood. We define “well-being AI” as an AI research paradigm for promoting psychological well-being and maximizing human potential. Interpretable AI is important for well-being AI in senses that (1) to understand how our digital experience affects our health and our quality of life and (2) to design well-being systems that put humans at the center.

One of the important keywords in understanding machine intelligence in human health and wellness is cognitive bias. Advances in big data and machine learning should not overlook some new threats to enlightened thought, such as the recent trend of social media platforms and commercial recommendation systems being used to manipulate people's inherent cognitive bias.

The second important keyword is “social embeddedness”. Cognitive bias will be affected by how the AI is perceived particularly at the community or social level. Social embeddedness is the social science idea that actions of individuals are refracted by the social relations within their community. In our contexts, understanding relationships between AI and society is very important, which includes the issues on AI and future economics (such as basic income, impact of AI on GDP), or “well-being society (such as happiness of citizen life quality).

This paper describes the detailed motivation, important keywords, the scope of interests and research questions in this symposium.

Motivation for Interpretable AI for well-being

Interpretable AI is an artificial intelligence methods and systems, of which outputs can be easily understood by humans. Recently, the European Union’s new General Data Protection Regulation (GDPR) has raised concerns about the emerging tools for automated individual decision-making. These tools use algorithms to make decisions based on user-level profiles, with the potential to significantly affect users. Recent AI technologies (e.g.: Deep Learning and other advanced machine learning methods) will change the world. However, excessive expectations for AI (e.g., the representation of general-purpose AI in science fiction) and threat theory (e.g. AI will lead to unemployment) distort the judgment of many people. Understanding both the potential and the limitations of the current AI technologies is therefore very important.

Especially in the human health and wellness domains, interpretable AI remains a huge challenge. For example, “evidence-based medicine” requires us to show the current best evidence in making decisions about the care of patients. “Why did the system make this prediction?” will be a key question. Even if the system is not accurate, it must be explainable and predictable. Although statistical machine learning predicts the future based on past data, it is difficult to respond to a new event which has never seen in the past. Training data that has outliers or adversarially generated data may lead an AI-based system to make wrong predictions (sometimes with high confidence) in life or death situations in medical diagnoses. For AI to be safely deployed, these systems must be well-understood. One of the important goals in this year's symposium is to discuss the technical and philosophical challenges of interpretability for well-being AI.

Understanding Cognitive Bias and Social Embeddedness

AI also provides the new risk of amplifying our “cognitive bias” through machine learning, as we discussed in our previous AAAI18 Spring symposium on “beyond machine intelligence” (Kido and Takadama, 2018). In the recent trend of big data becoming personalized, corresponding AI technologies for manipulating one’s cognitive bias are starting to evolve; examples of this include social media platforms such as Twitter and Facebook, and commercial recommendation systems. According to the “Echo chamber effect,” people with the same opinion tend to form communities, which makes it felt that everyone else also shares the same opinion. Recently, there has also been a movement to use such cognitive bias in the political world. We welcome discussions on "cognitive bias" in human or personal robot communications.

"Social embeddedness" of AI is also an important keyword in this symposium. We welcome diverse discussions on the relationships between AI and society. The topics on social embeddedness of AI may include such issues as "AI and future economics (such as basic income, impact of AI on GDP)" or "well-being society (such as happiness of citizen, life quality)", etc. Cognitive Bias will be affected by how the AI is perceived particularly at the community (or societal) level. "Social embeddedness of AI" seems likely to become a significant area as AI continues to develop.

Our Scope of Interests and Research Questions.

We expect to discuss important interdisciplinary challenges for guiding future advances in well-being AI. We will have the following scope of interests in this symposium:

(1) "Excessive expectation for AI - understanding possibilities and limitations of the current AI technologies",

(2) "Technical and philosophical challenges on interpretability for well-being AI"

(3) "Cognitive bias" and "social embeddedness of AI" in human/robot communications, from the socio-cultural/political aspects to the technical/practical, accuracy and efficiency issues in health, economics, and other fields.

More technically, we have the following research questions in Interpretable AI for well-being. We need to deepen the understandings of the possibilities and limitations of the Machine Learning and other advanced analyses for Health & Wellness.

1 Interpretable AI/ML

- How can we develop interpretable machine learning methods in well-being AI that provide ways to manage the complexity of a model and/or generate meaningful explanations?
- How can we use the tools of causal inference to reason about fairness in well-being AI? Can causal inference lead to actionable recommendations and interventions? How can we design and evaluate the effect of interventions?
- What are the societal implications of algorithmic exploration? How can we manage the cost that such exploration might pose to individuals?

2 Unintended consequence of algorithms in well-being AI

- Can we use adversarial conditions to learn about the inner workings of algorithms?
- Can we learn from the ways they fail on edge cases?
- Can we achieve accountability in well-being AI?
- How can we conduct reliable empirical black-box testing for ethically salient differential treatment?
- How can we manage the risks that such unintended consequence might pose to users?

Conclusion

In this paper, we described the motivation, technical, and philosophical challenges related to "Interpretable AI for well-being" and explained the two important keywords, "Cognitive Bias" and "Social Embeddedness", as proposers and organizers of this AAAI19 symposium.

This symposium is aimed at sharing the latest progress, current challenges and potential applications related with interpretable AI for well-being. Understanding possibilities and limitations of current AI/ML technologies on interpretability for digital health and wellness will be very important for designing human centric well-being AI.

References

Kido,T., Takadama, K. 2018. WELLBEING AI: FROM MACHINE LEARNING TO SUBJECTIVITY ORIENTED COMPUTING, AAAI *Spring symposium 2018* March, Stanford: https://aaai.org/Library/Symposia/Spring/ss17-08.php

Acknowledgments

We appreciate Amy Ding at Carnegie Mellon University. Research questions on interpretable AI in our scope were coming up with her valuable proposals and discussions. We also thank our program committees of this symposium for their valuable supports.