2021/6/19
Japan Association for Philosophy of Science 2021 Annual Meeting

Symposium:
Fairness, Integrity, and Transparency in Formal Systems:
Challenges for a Society Increasingly Dominated by Technology

Transparency in AI: Identifying the Real Issue

Takayuki SUZUKI
(The University of Tokyo)

This talk is based on the research project "Constructing Philosophy of Artificial Intelligence 2.0" funded by JST/RISTEX HITE program.

Recent Progress in AI Research

- Games
- Image recognition
- Machine translation
- Autonomous driving

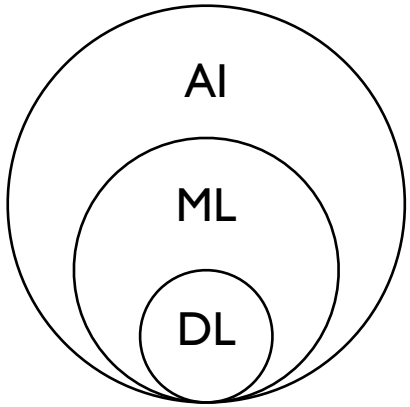
Disputes

- Transparency of algorithms
- Safety of self-driving cars
- Military use of autonomous agents
- Unemployment
- Conflict between human beings and AI

Important Distinctions

- Kinds of AI:
  - Artificial super intelligence
  - Artificial general intelligence
  - Specialized artificial intelligence

Important Distinctions (contd.)



- Kinds of contemporary AI:
  - AI in general
  - Machine learning AI
  - Deep learning AI

Limitations of Contemporary AI

- Needs large data sets for training (in machine learning).
- Needs labeled data (in supervised learning).
- Has adversarial examples (in deep learning).

Limitations of Contemporary AI (contd.)

- Lacks genuine understanding (of words etc.)
- Lacks common sense
- Lacks the sense of relevance

Implications

- The areas where AI, especially deep neural network, performs well are limited.
- Even in the areas where it performs well, it is not clear whether AI and human mind work in the same way.
  - Example: Computer vision and human vision

Implications (contd.)

- Having body and having experience in the real world is important (possibly indispensable) to general intelligence.

Transparency

- Distinctions:
  - Opacity due to social factors (secret) and opacity due to technological factors (inscrutability)
  - Complexity and inscrutability

Transparency (contd.)

- Transparency (scrutability) of algorithms:
  - Some (most) algorithms are transparent.
    - Voting algorithms

Transparency (contd.)

- Transparency (scrutability) of algorithms:
  - Even deep neural networks are not completely opaque.
    - We know general learning algorithm (back propagation etc.).
    - Sometimes we can identify what feature a network responds to.

Transparency (contd.)

- Opacity of algorithms:
  - Deep neural networks are not always transparent.
    - Example: Chess AI.
    - What if AI checks the topological relation of 7 pieces in the last 20 moves?
    - Can understand the mechanism in principle. Cannot predict the response in practice. Cannot give a meaningful interpretation for the feature.

Transparency (contd.)

- Transparency in other areas:
  - Drugs in medicine
    - Transparency to whom?
  - Human decision making
    - Is the truth of explanation important?
    - Should we care more about transparency here?

Transparency (contd.)

- A trade-off between transparency and performance.

Transparency (contd.)

- A trade-off between transparency and performance.
- When, why, and to what extent should AI be transparent?

Biases

- Whose biases?
  - In most cases, biases come from the data.
  - We might be able to use AI to find biases in our society.
    - Example: Word2Vec
- What is the problem of biases that is unique to AI?

Biases (contd.)

- The problem of value-ladenness:
  - 'Fairness' in voting algorithms
  - A more subtle case: Bias in COMPAS (Sumpter 2018, Ch.6)

Biases (contd.)

- The problem of emergent biases:
  - The Google Photo case (Mitchell 2019, Ch.6) :
    - No biases in the data.
    - No biases in the algorithm.
  - Cases like this, as well as adversarial examples, suggest that AI and human mind might not work in the same way.

A Positive Perspective

- Two conceptions of AI:
  - AI as a substitute for human beings
  - AI as a complement for human beings

A Positive Perspective (contd.)

- Overcoming human weaknesses with AI:
  - Human weaknesses: Poor logical reasoning, slow processing speed, poor memory, fatigue, cognitive biases…
  - Where we perform poorly is where AI performs well.
  - We may be able to overcome our weaknesses with the help of AI.

Conclusion

- Morals:
  - We have to distinguish the problems originated in AI itself and the problems originated in human society.
  - We have to look at the cases outside of AI to be consistent about the problem of transparency.
  - We may be able to identify our own biases and other problems through AI.
  - We should use AI so that we can overcome our weaknesses.