AI and Human Nature

Jinho Kang
Seoul National University
jhkang@snu.ac.kr

Global HR Forum 2020
“AI and Human, How Can We Coexist?”
November 11, 2020
Grand Walkerhill, Seoul

The AlphaGo Shock: Machine Beats Human!

Why Were We Shocked?

- There are already machines in many areas that are superior to humans in performing required tasks.

- Question
  - So **why** did we react so differently to AlphaGo, being greatly shocked and even feared by its victory?

Rationality as the Essence of Human Nature

- My Answer
  - We suppose
    (1) that the capacity of **rational thinking** is the essential characteristic of humans, and
    (2) that we are not only different from all other animals but also **superior** to them precisely **because** we have the capacity of rational thinking.

- Playing the game of Go well requires an exceptionally high level of rational thinking, as it is the most complex board game in the world.

- Playing such an intelligent game, AlphaGo won a sweeping victory against Lee Sedol, one of the world's top Go players.

- Natural Reactions
  * "AlphaGo is **better** than us in **rational power**, the distinctive feature of humans!"
  * "We now made a machine that is **genuinely superior** to human beings!"
  * "AI machines might become our **masters** in the future, just as we have been masters of all other animals precisely because we have ability to think rationally!"

Two Views on Human Nature: The Classical View

- The Classical View
  - The essence of human nature lies in its rational capacity.

- **Plato** (BC 428/7-348/7)
  - The soul of human beings is divided into three parts: appetite, spirit, and reason
  - Reason plays the central role, guiding and controlling appetite and spirit.

- **Aristotle** (BC 385-322)
  - Plants, animals, and humans all have souls.
  - The soul of plants has capacities for nutrition and growth.
  - The soul of animals has further capacities for perception, desire, and movement.
  - The soul of humans has still further capacity for rational thinking, which enables them to pursue knowledge and perform good actions.

- A Neoplatonic philosopher **Porphyry**(234-305) defines human beings as **"(mortal) rational animals**", which was widely accepted for a long time.

Two Views on Human Nature: The Humean View

- **David Hume** (1711-1776)

- Human beings are fundamentally controlled by **“passions”** and **“sentiments”**, not by reasons.

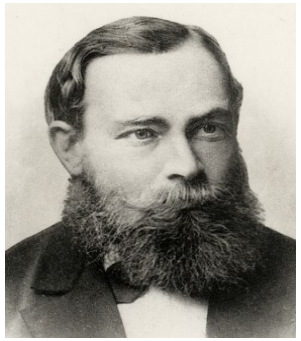
> “Reason is, and ought to be, the slave of the passions, and can never pretend to any other office than to serve and obey them.”
>
> (Hume (1739), *A Treatise on Human Nature*, p.415)

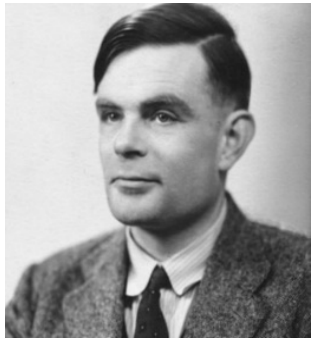
- What kind of “passions” and “sentiments” (= desires and emotions) we have is a purely **subjective** matter, and therefore cannot be the object of rational evaluation.

> “’Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. ’Tis not contrary to reason to choose my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me”
>
> (Hume (1739), *A Treatise on Human Nature*, p.416)

The Birth of “Rational” Machines

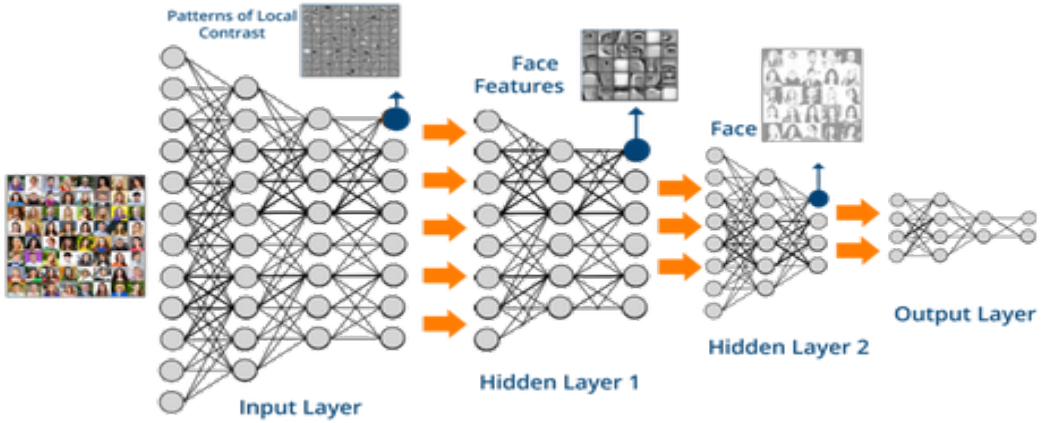
- **Gottlob Frege** (1884-1925) revolutionizes **logic**, whose achievement enables us to **formalize** logical and mathematical reasoning.
- **Alan Turing** (1912-1954) proposes the idea of “Turing machine”, which enables us to devise a machine that can implement any **computational** processes including those of formalized deductive and mathematical reasoning.





- Frege’s and Turing’s works provide theoretical foundations for modern **computers** (= machines that compute).
- Today’s computers have much better capacities of **formal logical reasoning** and **mathematical calculation** than humans, the two capacities that had been thought to be paradigms of human rational powers for a long time in the Western tradition since Plato and Aristotle.

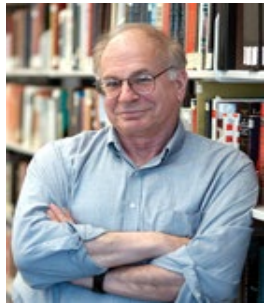
- The **AI research**, which began in the 1950s, has been undertaking the ambitious project of implementing the **whole human intelligence** by machines.
- The rapid progress of research on **machine learning**, in particular on **deep learning** since the late 2000s, is fundamentally changing the field of AI research.

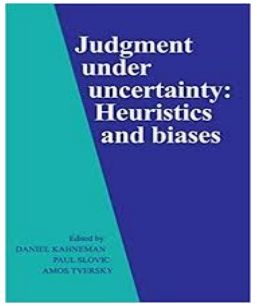


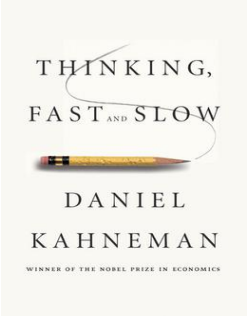
The Victory of the Humean View?

- The classical view of human begins as rational animals has been recently challenged by various researches in various fields, including biology, neuroscience, and psychology.

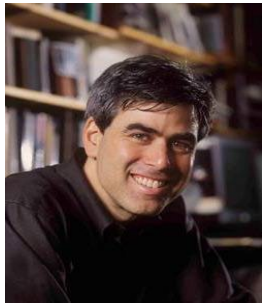
- Daniel Kahneman (1934-)
  - Human judgments have systematic **biases** created by **heuristic thinking**.
  - In our judgments, “System 1” (fast, automatic, intuitive) plays a larger role than “System 2” (slow, conscious, logical).

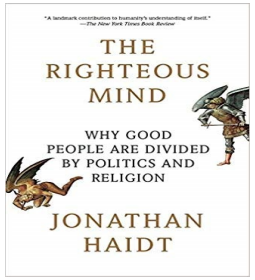






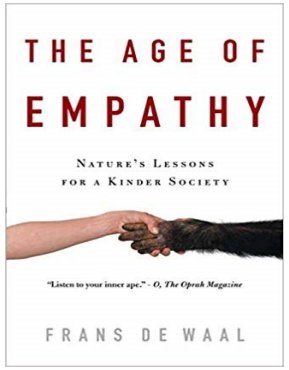
- Jonathan Haidt (1963-)
  - Our moral judgments primarily come from **intuitions** or **emotions**, not from reason.

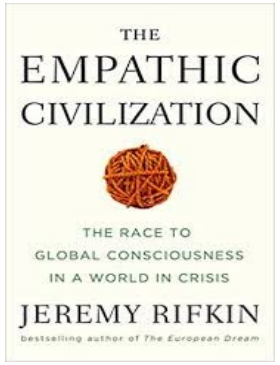


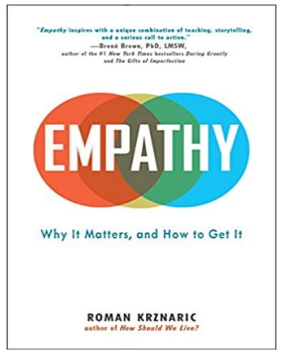


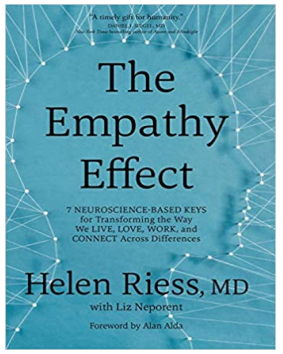
The Age of Empathy?

- Many people, including researchers in biology, psychology, and philosophy, now claim that **empathy** is a key to what makes us truly human.









- Frans de Waal (1948-) on the "Age of Empathy"
  - Our empathy has **deep evolutionary roots**, having originated before the order Primates came into existence.
  - It is likely that our distinctive capacity for **altruism** and **fairness** has also come from our power of empathy.
  - It is only after we understand the evolutionary value of empathy that we can have a proper understanding of human nature and progress from the "Age of Greed" to the "Age of Empathy".

Critical Comments

- In spite of recent critiques, I believe that the classical view of human beings as rational animals is still **fundamentally right**.

- It is our rational capacity that makes us truly human and distinguishes us from other beings.

- The Humean view of human nature is on the wrong track with its emphasis on the **emotional** or more broadly **non-rational** capacities of human beings ("empathy", "imagination", "creativity", etc.).

- The Humean view of human nature is not only on the wrong track but also potentially **dangerous**, for our emotional capacity, including the capacity of empathy, is **parochial**, **biased**, **short-sighted**, and **easily manipulatable**.

    (cf.) Comparison of **empathy** and **anger**

    (cf.) Distinction between **empathizing with others** and **understanding others**

Four Main Theses

- Question

  "If you are right that what makes us truly human is our rational capacity, then AIs such as AlphaGo are **already human-like beings** given their superior power of rational thinking!"

- Answer: No!

- Four Main Theses

(1) The core of human rationality lies in our **reflective capacity**, or the capacity of rational deliberation.

(2) All mental states of human beings, including perceptions, beliefs, desires, and emotions, are transformed into a **higher** level of mental states thanks to our reflective capacity.

(3) AIs in its current form are **unable** to implement our reflective capacity. And it is **far from certain** whether AIs in the future can do so.

(4) The classical view of human beings as "rational animals" will become **even more important** in the age of AI, as it is crucial to cultivate and improve our reflective capacity in order to develop a **proper symbiotic relationship** with AIs.

The Origin and Meaning of the Word ‘Reason’

- Origin of the Word ‘Reason’

  ‘logos (λόγοζ)’(Greek)

  ➔ ‘ratio’ (Latin)

  ➔ ‘reason’(English), ‘raison’(French), ‘Vernunft’(German)

  ➔ ‘理性’ (Japanese translation of 'reason'), ‘이성’(Korean)

- Meanings of the Greek Word ‘Logos’

  - Original meanings: ‘speech’, ‘calculation’

  - Broadened meanings: ‘ratio’, ‘judgment’, ‘conceptual thinking’, ‘inference’, ‘reasoning’, ‘explanation’, ‘ground’, ‘principle’, ‘thesis’, ‘proof’, ‘theory’, etc.

Reflective Capacity as the Core of Human Rationality

- In the Western tradition, human being's 'rational capacity' includes the following:
  (1) capacity of logical reasoning
  (2) capacity of mathematical calculation
  (3) capacity of speaking and understanding language
  (4) capacity of expecting the best outcome and achieving it [= utility maximizing]
  (5) capacity of explaining why something is a case
  (6) capacity of introspecting and examining one's own mind [=reflective capacity]

- Among them, the **core** of human being's rational capacity, which makes us truly human, is (6),
  (cf.) "The unexamined life is not worth living" (Plato, *Apology,* 38a5-6)

- The reflective capacity is the capacity of asking and finding **reasons** that can **justify** our mental states such as perceptions, beliefs, desires, and emotions.
  ➔ The core of human **reason** is the capacity of asking and finding **reasons** for why we **should** have such and such mental states.

Non-Human Animals Do Not Have Reflective Capacity

- Lion
  - Belief: <There is a dog in front of me>
  - Desire: <I want to eat the dog>

- Human Being
  - Belief: <There is a dog in front of me>
  - Desire: <I want to eat the dog>
  - **Reflection**: <Do I have a **reason** to believe that there is a dog in front of me?>
    <I do have a **reason** to believe so!>
  - **Reflection**: <Do I have a **reason** to desire that I eat the dog?>
    <I don't have a **reason** to have such a desire!>

Two Understandings of Humans as ‘Rational Animals’

- First Understanding
  - Human beings have mental states such as perceptions, memories, beliefs, desires, emotions, feelings, etc. **just like other animals**.
  - Human beings are different from other animals in that they have an **additional** capacity, i.e. the rational capacity, that cannot be found in other animals.

- Second Understanding
  - Thanks to the rational capacity, the mental states of human beings such as perceptions, memories, beliefs, desires, emotions, feelings, etc. are **transformed** into a **higher** level of mental states, a level that **cannot** be found in other animals.
  - Therefore, the **“animality”** of our human beings should be understood as a **higher** level of animality.

- I believe that the second understanding is correct.

The Mental States of Humans Are at a Higher Level

- Thanks to our reflective capacity, our mental states are transformed into a “higher” level of mental states in the following three senses:

(1) Because of our reflective capacity, we can ask about all of our mental states whether they are **justifiable** by some **reasons**, and accordingly we can distinguish between **correct** and **incorrect** mental states.

(2) A reason must be something **universal** and applicable to everyone. Therefore, the “reason criterion” by which we distinguish between correct and incorrect mental states is a **universal one** that can be applied to everyone.

(3) We can **improve** our mental states in such a way that they form a more **consistent** and **coherent** system, and this enables us to form a more and more **unified self**.
(➔ the emergence of a **reflective self**)

(cf.) the reflective self vs. the biological self

Can AIs Have Reflective Capacity?

- Varieties of Human Rational Capacities and the Current Status of AI Research

Human Rational Capacities	**Current Status of AI Research**
(1) (deductive) logical reasoning	well done!
(2) mathematical calculation	well done!
(3) speaking and understanding language	making progress
(4) achieving the best expected outcome	making progress
(5) explaining why something is a case	having just started [**"explainable AI"(XAI)**]
(6) introspecting and examining one's own mind	no research currently being done
[= finding reasons that can justify one's mental states]	

- In order to create AIs that can successfully implement (5), we must be able to **formalize** the notion of **explanation**. But we don't yet have a good idea of how to do so.

- In order to create AIs that can successfully implement (6), we must be able to **formalize** the notions of **reasons** and **justification**. But we don't yet have a good idea of how to do so.

The Difficulty of Formalizing Abductive Reasoning

- **Abductive reasoning**, or **abduction**, is a form of reasoning in which one reaches a conclusion that **best explains** the premises.

- Abductive reasoning is also called the "**Inference to the Best Explanation(IBE)"**, which is widely used in both everyday lives and scientific activities.

(e.g.) (P1) Young-hee is not in her room.
(P2) It is Wednesday 9am in the morning.
(P3) Rice is newly cooked in the pressure cooker.
(P4) An emptied rice bowl, a spoon, chopsticks, and some side dishes are on the table.
Therefore, Young-hee cooked her breakfast and left home.

- Unlike deductive reasoning, however, it turns out to be extremely difficult to formalize abductive reasoning due to the elusiveness of the notion of explanation.

"Abduction really is a terrible problem for cognitive science, one that is unlikely to be solved by any kind of theory we have heard of so far."
- Jerry Fodor (2001), *The Mind Doesn't Work That Way,* p.41

What If AIs in the Future Have Reflective Capacity?

- If we manage to create AIs in the future that do have genuine reflective capacity, we will have to treat them as human-like beings, for it is this capacity that makes us truly human.

- Question
  - If we create AIs in the future that are much better than us in their reflective capacity and hence make much better reflective judgments about what we ought to believe, desire, and feel, **should we rely on these AIs in deciding what to believe, desire, and feel?**
    (cf.) The increasingly wide use of “recommendation AIs”

- Answer: We should not!
  - If we let AIs decide what we ought to believe, desire, and feel, then we will **lose** our (reflective) **selves**.

  - If AIs make all decisions about what to believe, desire, and feel for us, then **we will no longer live our lives**.

    (cf.) Thought Experiment: Chulsoo and the reflective AI "DM 21"
    (cf.) Socrates’s emphasis on the Delphic maxim “Know thyself!”

Conclusions

- The core of human rational capacity is the **reflective capacity**, i.e. our ability to ask and find **reasons** that can **justify** our mental states such as beliefs, desires, and feelings.

- It is this capacity that makes us truly human. Hence the **classical view of human beings as "rational animals"** is fundamentally right.

- The classical view of human beings as rational animals will **become even more important in the age of AI**, for we will be able to resist our strong temptation to let AIs make decisions about what we should believe, desire, and feel for us only if we understand that
  (1) it is our rational capacity as reflective capacity that makes us truly human, and that
  (2) it is the exercise of our reflective capacity that makes our lives worth living and enables us to form our true selves.

- It is only when we realize the central importance of our rational capacity as reflective capacity that we will be able to develop a proper symbiotic relationship with AIs.