

뇌의 4대 고위수준 기능을 모사하는 시스템3 강화학습 기술

(0-1-2)

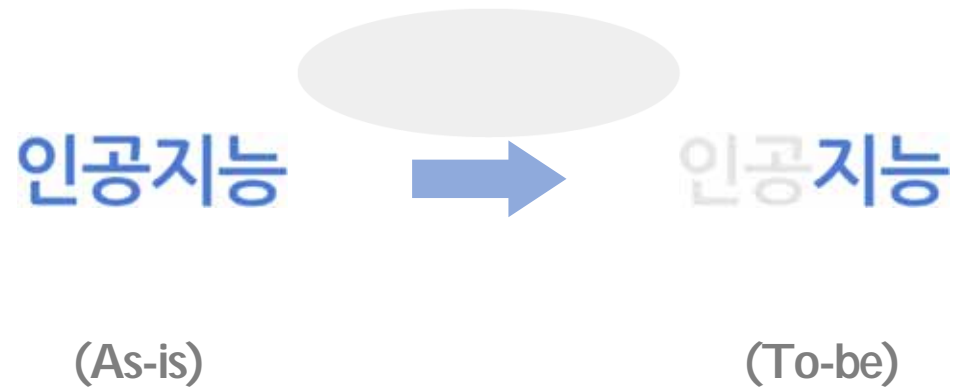
()

KAIST

-

AI /

제안 연구팀의 궁극적인 지향점



인간과 다르게 생각하는 인공지능

(, GPT, GPT-4) 가



GPT-4 surpasses ChatGPT in its advanced reasoning capabilities.

ChatGPT

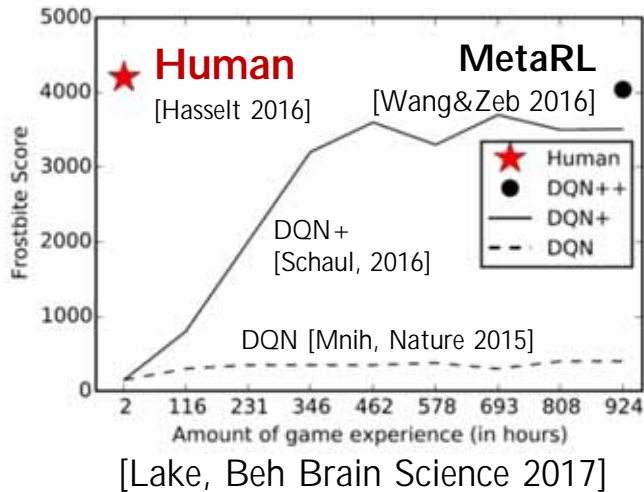
Input
Andrew is free from 11 am to 2 pm, Joanne is free from noon to 2 pm and then 3:30 pm to 8 pm, Hannah is available at noon for just an hour, and then 4 pm to 6 pm. What are some options for start times for a 30-minute meeting for Andrew, Hannah, and Joanne?

Output
Here is a list of their availability:
- Andrew: 11 am to 2 pm
- Joanne: noon to 2 pm, and 3:30 pm to 8 pm
- Hannah: noon to 12:30 pm, and 4 pm to 6 pm
Based on their availability, there is a 30-minute window where all three of them are available, which is from 4 pm to 4:30 pm. So, the meeting can be scheduled at 4 pm.

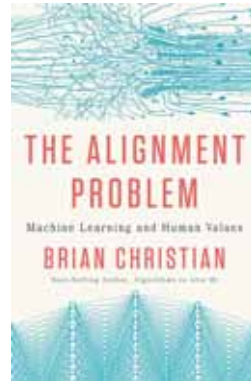
GPT-4

Input
Andrew is free from 11 am to 2 pm, Joanne is free from noon to 2 pm and then 3:30 pm to 8 pm. Hannah is available at noon for just an hour, and then 4 pm to 6 pm. What are some options for start times for a 30-minute meeting for Andrew, Hannah, and Joanne?

Output
Andrew: 11 am - 3 pm
Joanne: 12 pm - 2 pm, 3:30 pm - 5 pm
Hannah: 12 pm - 12:30 pm, 4 pm - 6 pm
Common availability for a 30-minute meeting: 12 pm - 12:30 pm



가 가 가 .



Consequences of Misaligned AI

Simon Zhuang
Center for Human-Compatible AI
University of California, Berkeley
Berkeley, CA 94709
simonzhuang@berkeley.edu

Dylan Hadfield-Menell
Center for Human-Compatible AI
University of California, Berkeley
Berkeley, CA 94709
dhm@berkeley.edu

[Zhuang, Neurips 2020]

[Christian, Alignment problem 2020]

인간과 다르게 생각하는 인공지능

✓ 고위수준 기능 예시 (인간 ≠ 인공지능)



• 충동

• 직관

-
-
-



• 내재적 보상



• 상황파악-시점변환

인간처럼 생각하는 인공지능의 필요성

Value alignment problem (mismatch between AI and human values)



Consequences of Misaligned AI

Simon Zhang
Center for Human-Compatible AI
University of California, Berkeley
Berkeley, CA 94720
simonzhang@berkeley.edu

Dylan Hadfield-Menell
Center for Human-Compatible AI
University of California, Berkeley
Berkeley, CA 94720
dhadfield@berkeley.edu

Workshop on Human and Machine Decisions
@ NeurIPS 2021

HCAI @ NeurIPS '22
Virtual Workshop on Human-Centered AI Workshop at NeurIPS
Online on 9 December 2022

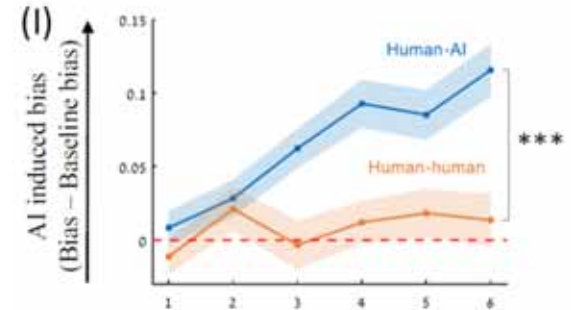
AI-

_____ ('20)

MemARI
NeurIPS 2022 | December 2nd.
New Orleans, USA

AI-

_____ ('21-)



"...AI that aligns with instructions, intentions, revealed preferences, ideal preferences, interests and values." (DeepMind)
"AI should be designed to align with the norms and values of the human." (IBM)

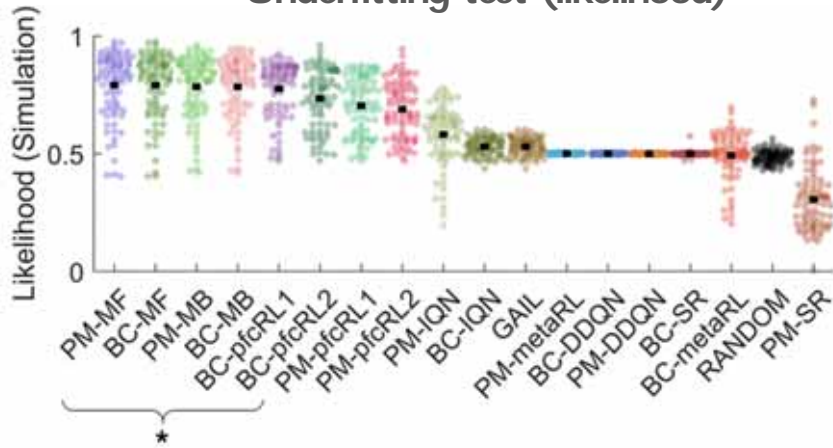
AI-
AI-

_____ ('23)
('21-)
[Cha, AAAI 2021]

기존 기술의 한계

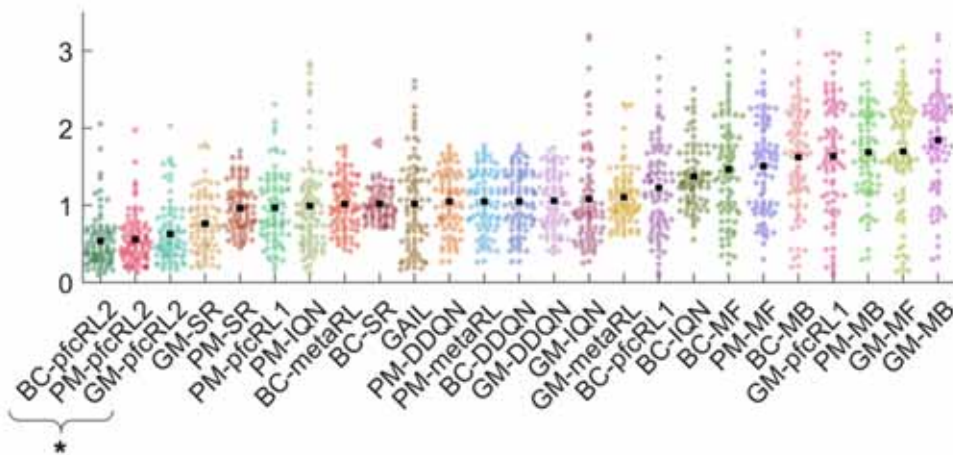
[]

Underfitting test (likelihood)



Overfitting test (contextual behavior recoverability)

Mean L1 norm



(overfitting)

[Broomell, Decision 2014]

[Evans, Beh Res Methods 2019]

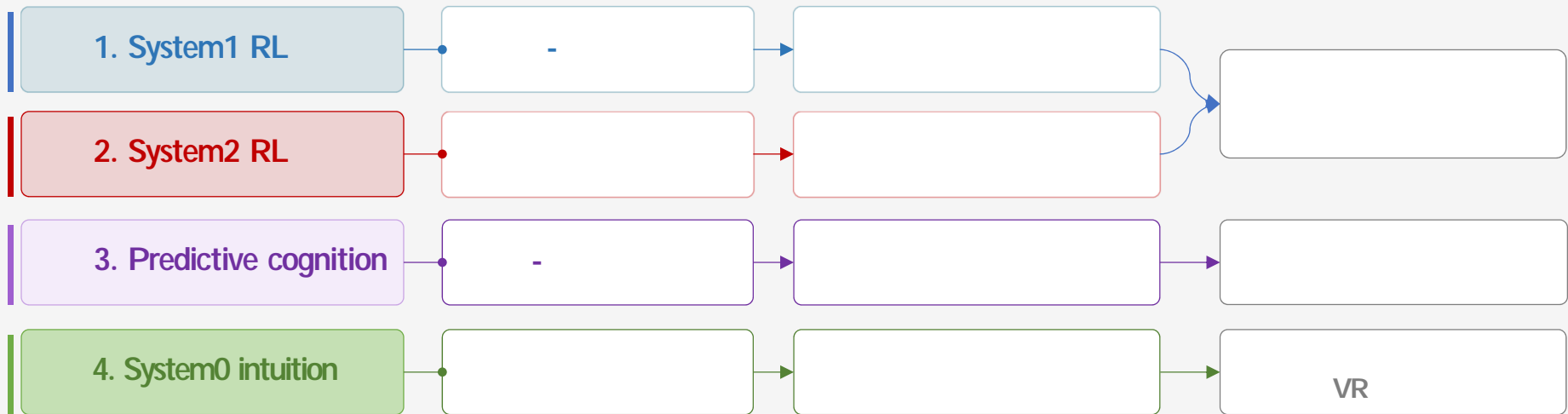
연구목표 (SW 관점)

4 (1-2- -)
3

뇌원형 모델

구현기술 (공개SW)

공개 플랫폼+DB



연구목표 (뇌과학적 관점)

□ 4 (1-2- -)

SW

1. System1 RL

- ()

Distributional RL
Model-free RL

■ Basal ganglia
(movement, reward)

■ Thalamus
(sensory gateway)

■ Hippocampus
(memory)

Successor representation

3. Predictive cognition

- ()

4. System0 intuition

기타 ()

Metamerism
Intuitive physics

2. System2 RL
()

Model-based RL
Metacognitive RL

Cerebral cortex

Forebrain

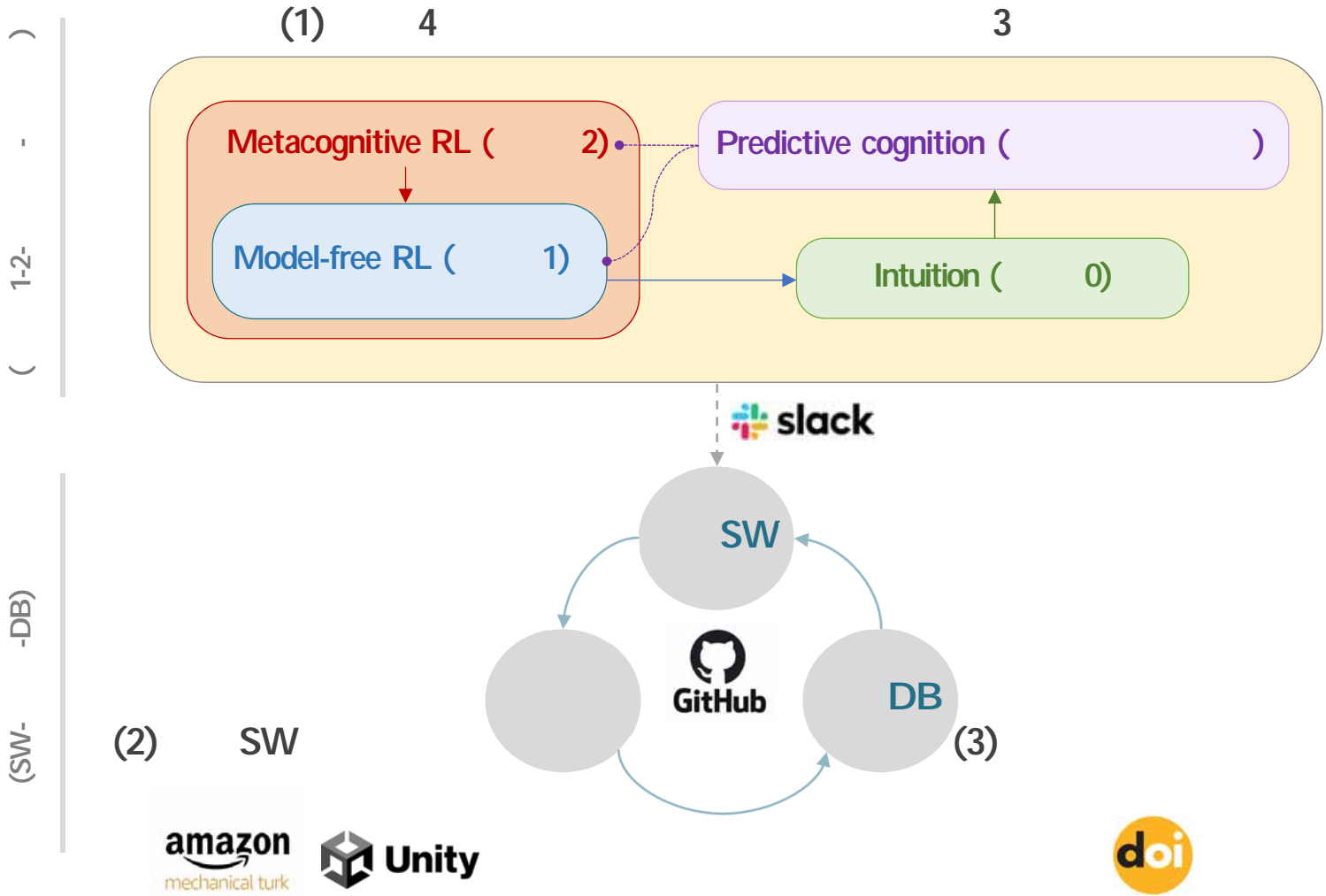
■ Hypothalamus
(regulates body
function)

■ Amygdala
(emotion)

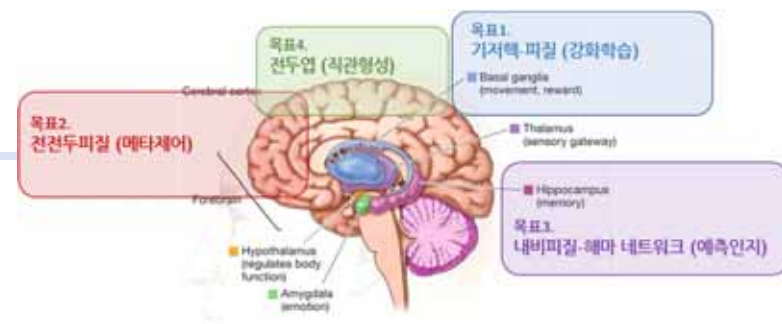
• / / ()
• / / -

연구개발 추진체계

● 4 _____ x 3 _____



제안 연구의 차별성



- [redacted] :
- [redacted] : “ ”(as-is)
 → “ ” / ”(to-be)
- [redacted] : human-likeness +
- [redacted] SW: 4 (1-2- -)
- [redacted] DB: 4
- [redacted] : behavioral stability-flexibility
 ()
 ()
 (: '19 Nature comm)
 가

감사합니다.



뇌모사 인공지능 연구 인력 양성과
세계 최고 수준의 뇌모사 인공지능 SW기술 개발을 위해
끊임없이 노력하겠습니다.