초 학 제 연 구 프 로 그 램

<국제학술대회>

# 휴머니즘을 넘어서
# : 인공지능, 정보, 포스트휴머니즘

## Beyond Humanism
## : Artificial Intelligence, Information and Posthumanism

- 일시: 2019년 12월 17일(화) 10:00 – 18:00
- 장소: 고등과학원 1호관 5층 세미나실(1503호)
- 언어: 영어 (한영 동시통역 제공)
- 주관: 인공지능과 포스트휴머니즘 연구단,
  인포스피어 휴머니티를 위한 정보철학 연구단

KIAS KOREA INSTITUTE FOR ADVANCED STUDY NRF

Transdisciplinary Research Team

# International Conference on
# Beyond Humanism : Artificial Intelligence, Information and Posthumanism

- Date: Tuesday 17 December 2019 10:00 – 18:00
- Location: KIAS Bldg#1 Room #1503, Seoul, Korea
- Language: English (Korean interpretation provided)
- Organizer: ‘AI an Post-humanism’ research group, ‘Philosophy of Information for Infosphere Humanity’ research group

KIAS NRF

# 초대의 글

인류 역사에서 17~18세기 근대혁명은 근대적 개인, 근대적 사회, 인본주의를 탄생시켰습니다. 그리고 근대철학은 개인을 주체적 존재로 보고, 그러한 존재에게 인간 본성의 핵심요소라 할 수 있는 이성, 감성, 도덕성, 가치, 자의식, 자유의지 등이 내재돼 있음을 강조하였습니다. 이를 토대로 인간과 인간이 아닌 다른 것들 사이의 경계를 명확히 하고, 다른 모든 것들은 주체인 인간을 중심으로 그 주위에 객체로서 마주하는 존재로 보는 인간중심주의를 탄생시켰습니다.

하지만 21세기에 들어와서 이러한 인간중심주의에 변화가 일어나고 있습니다. 그 동안 인간에게만 고유한 것으로 인식됐던 능력들(감성, 이성, 도덕성 등)이 기계에서도 (제한적이지만) 구현가능하게 되었습니다. 또한 21세기의 인공지능은 인간의 삶에 관한 무한한 정보의 보고인 빅데이터와 자기-주도적인 심화학습 알고리즘을 통해 그 능력이 점차 인간을 능가해 가고 있습니다. 이러한 기술의 발전은 인간의 정신적 활동까지 대신하면서 사회관계 및 생활세계에서의 변화뿐 아니라, 그에 수반하는 윤리적인 법적인 쟁점들의 부상과 더불어, 휴머니즘 및 인간 정체성에도 많은 변화를 예고하고 있습니다. 이는 그 동안 인류사회를 지탱해 온 인간중심적인 휴머니즘을 다시 생각하게끔 만듭니다. 그런 의미에서 21세기는 포스트휴먼의 시대 또는 포스트휴머니즘의 시대라 할 수 있습니다.

이러한 시대적 조망 하에 본 국제학술대회에서는 인공지능을 중심으로 그것이 인류 사회에 던지는 사회적·철학적 함의를 함께 탐색하고자 합니다. 특히 인간-인공지능 간 새로운 방식의 상호작용과 인류 사회의 변화, 이에 수반하는 윤리적 · 법적 · 사회적 쟁점들, 그리고 인공지능의 존재적 본질과 인간 정체성에 대한 이해의 변화에 대해 함께 논의하고자 합니다. 궁극적으로 인간과 인공지능이 함께 공존하면서 발전할 수 있는 적합한 담론으로서 포스트-휴머니즘을 함께 모색하고자 합니다.

본 국제학술대회는 고등과학원의 <초학제 연구프로그램>의 일환으로 올해의 주제인 <인공지능과 포스트휴머니즘>을 연구하고 있는 연구팀, 한국연구재단의 후원을 받고 있는 <인포스피어 휴머니티를 위한 정보철학> 연구팀이 공동으로 개최합니다. 휴머니즘을 넘어서서 미래사회의 변화에 선제적으로 대응할 수 있는 방안을 모색하기 위해 유럽학자 및 일본, 대만, 중국 그리고 국내 학자들을 초청하오니, 많은 관심과 참여 바랍니다.

# INVITATION

The modern revolution of human history in the 17th and 18th centuries gave birth to modern individuals, modern societies, and humanism. In addition, modern philosophy regarded individuals as subjects and emphasized that such beings have reason, emotion, morality, value, self-consciousness and free will which are the core elements of human nature. Based on this, it clarifies the boundary between humans and non-humans and creates anthropocentricism, which sees non-humans as objects around the subject human being.

But in the twenty-first century, this kind of humanism is being challenged. So far, abilities (sensitivity, reason, morality, etc.) that have been perceived as unique to humans in the past have become feasible (albeit limited) in machines. In addition, AI in the 21st century is gradually surpassing human capabilities through big data, an infinite repository of information about human life and self-directed deep learning algorithms. The development of these technologies, replacing human mental activities, signals many changes in humanism and human identity, as well as changes of social relations and the world of life, and then the rise of ethical and legal issues that accompany them. This makes us rethink the anthropocentric humanism that has supported human society. In that sense, the 21st century can be called the post-human age or the age of post-humanism.

Under this perspective, this international conference seeks to explore the social and philosophical implications artificial intelligence puts into the human society. In particular, we will discuss new ways of interaction between human-artificial intelligence, changes in human society, accompanying ethical, legal and social issues, and changes in the understanding of human identity and the existential nature of artificial intelligence. Ultimately, I hope, we would like to explore post-humanism as a suitable discourse where humans and artificial intelligence can coexist and develop together.

This international conference, as a part of <Interdisciplinary Research Program> in the Korea Institute of Advanced Science(KIAS), is jointly held by the research team studying this year's theme “Artificial Intelligence and Post-Humanism” in KIAS together with the another research team on “Philosophy of Information for Infosphere Humanity” sponsored by the Korea Research Foundation. We invite European, Japanese, Taiwanese, Chinese, and Korean scholars to seek ways to proactively respond to changes in future society beyond humanism, so please take a lot of interest and participation

# 프로그램

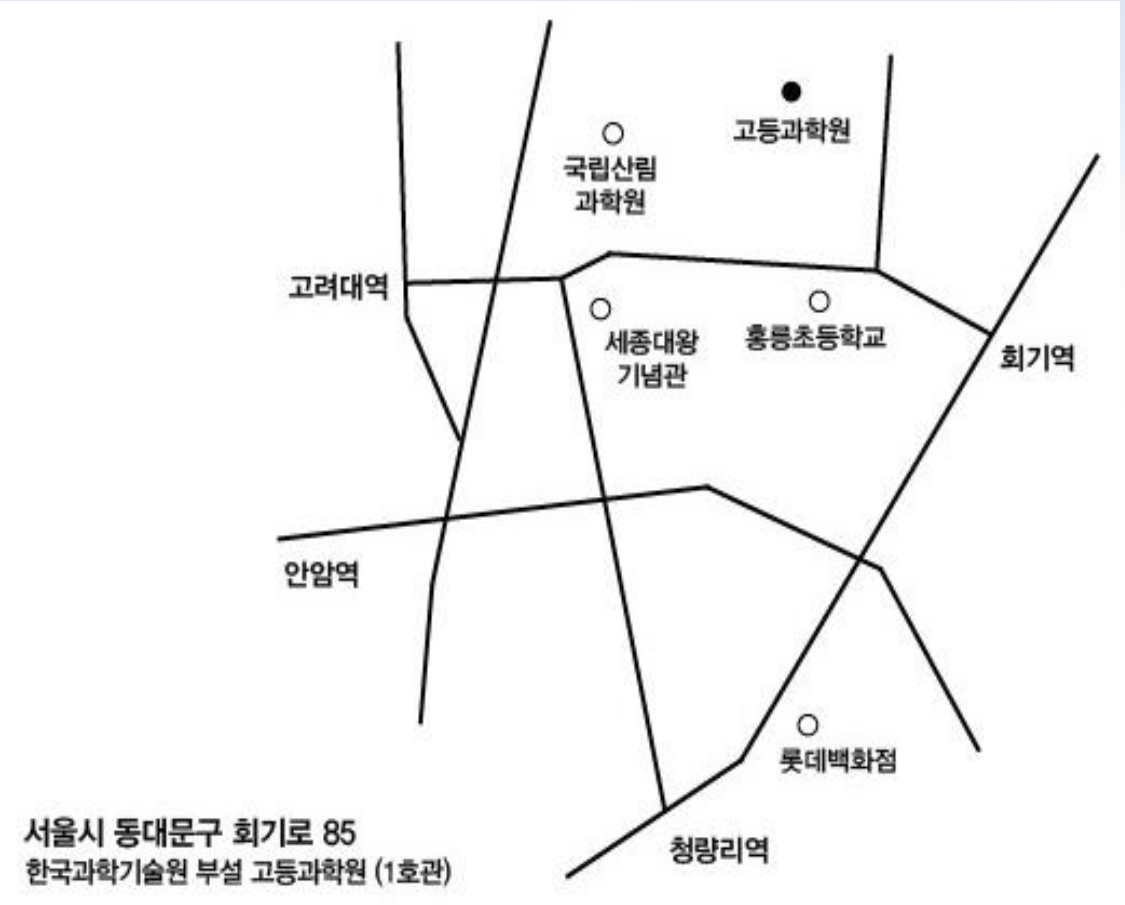
| | |
|---|---|
| **Registration & Reception** | moderator : In-Ryeong Choi |
| 10:00 – 10:20 | Registration |
| 10:20 – 10:30 | Reception |
| **Session 1** | moderator : Young E Rhee |
| 10:30 – 11:00 | **Materiality of intelligence and decentering of anthropocentricism**<br>Kyoung-Min Lee(Seoul National University College of Medicine, Korea) |
| 11:00 – 11:30 | **Body-Conservatism**<br>Kojiro Honda(Kanazawa Medical University, Japan) |
| 11:30 – 11:50 | Discussion |
| 11:50 – 13:00 | Lunch |
| **Session 2** | moderator : Insok Ko |
| 13:00 – 13:30 | **What AI can learn from Husserl's notion of "intentionality"?**<br>Yingjin Xu (Fudan university, China) |
| 13:30 – 14:00 | **Enhancement, Uploading, and Personal Identity**<br>Sangkyu Shin(Ewha Womans University, Korea) |
| 14:00 – 14:30 | **Artificial Moral Agent: its Moral Status and Authority**<br>Tsung-Hsing Ho(National Chung Cheng University, Taiwan) |
| 14:30 – 15:00 | Discussion |
| 15:00 – 15:30 | Coffee Break |
| **Session 3** | moderator : Hyundeuk Cheon |
| 15:30 – 16:00 | **Posthumanism in the Age of AI, Expanding Humanistic Attitude**<br>Sang-Wook Yi<br>(Hanyang University, Korea) |
| 16:00 – 16:30 | **Ethics of AI: Responsibility and Policy**<br>Mark Coeckelbergh<br>(University of Vienna, Austria) |
| **Round Table** | moderator : Sangkyu Shin |
| 16:40 – 18:00 | **Comprehensive Discussion**<br>all speakers |
| 18:30 – 20:30 | Dinner |

# 발표자

- **이경민(Kyoung-Min Lee)**
  서울대학교 의학과 신경과학교실 교수
- **Kojiro Honda**
  Associate Professor, General Education Department (Humanities),
  Kanazawa Medical University (일본)
- **Yingjin Xu**
  Professor, Fudan Philosophy of Science, Fudan university (중국)
- **신상규(Sangkyu Shin)**
  이화여자대학교 이화인문과학원 교수
- **Tsung-Hsing Ho**
  Associate Professor, Department of Philosophy,
  National Chung Cheng University (타이완)
- **이상욱(Sang-Wook Yi)**
  한양대학교 철학과 교수
- **Mark Coeckelbergh**,
  Professor, Department of Philosophy, University of Vienna (오스트리아)

# 고등과학원 오시는 길

- **고려대역에서**
  3번출구 방면,
  도보 약 20분 소요
- **안암역에서**
  안암전철역 정류장
  273번 버스
  (한국과학기술원 하차)
- **청량리역에서**
  현대코아 정류장
  201번 버스
  (한국과학기술원 하차)
- **회기역에서**
  시조사삼거리 정류장
  201번 버스
  (한국과학기술원 하차)




서울시 동대문구 회기로 85
한국과학기술원 부설 고등과학원 (1호관)

# Materiality of intelligence and decentering of anthropocentricism

Kyoung-Min Lee, MD, PhD
Neurology and Cognitive Science
Seoul National University

# Outline of the arguments

- Anthropocentricism is based on the assumption that human beings are special in their unique mental abilities (= intelligence and rationality)
- Human intelligence is produced and constrained by materiality of the world.
- Consciousness, unconscious cognition, and life-organizing processes are in a physical / evolutionary continuum.
- These taken together reject anthropocentricism as a rational thesis, from the viewpoint of epistemic and theoretical rationality.
- Anthropocentric arguments based on the practical rationality perspective also fall short, since the human condition as is and will become by default is not sustainable (sub-optimal), inhumane (contrary to the humanistic claim by anthropocentricists), and just 'bad' (deontologically or sociologically).

# Materiality of human intelligence

- Duality of the brain physics
  - Information-processing system
  - Electrochemical circuits consuming energy
- Exchange between information (I) and matter/energy (ME)
  - At the individual level, accumulation of experience (I) by neuroplasticity (ME)
  - At the species level, accumulation of adaptation (I) by selected individuals (ME)

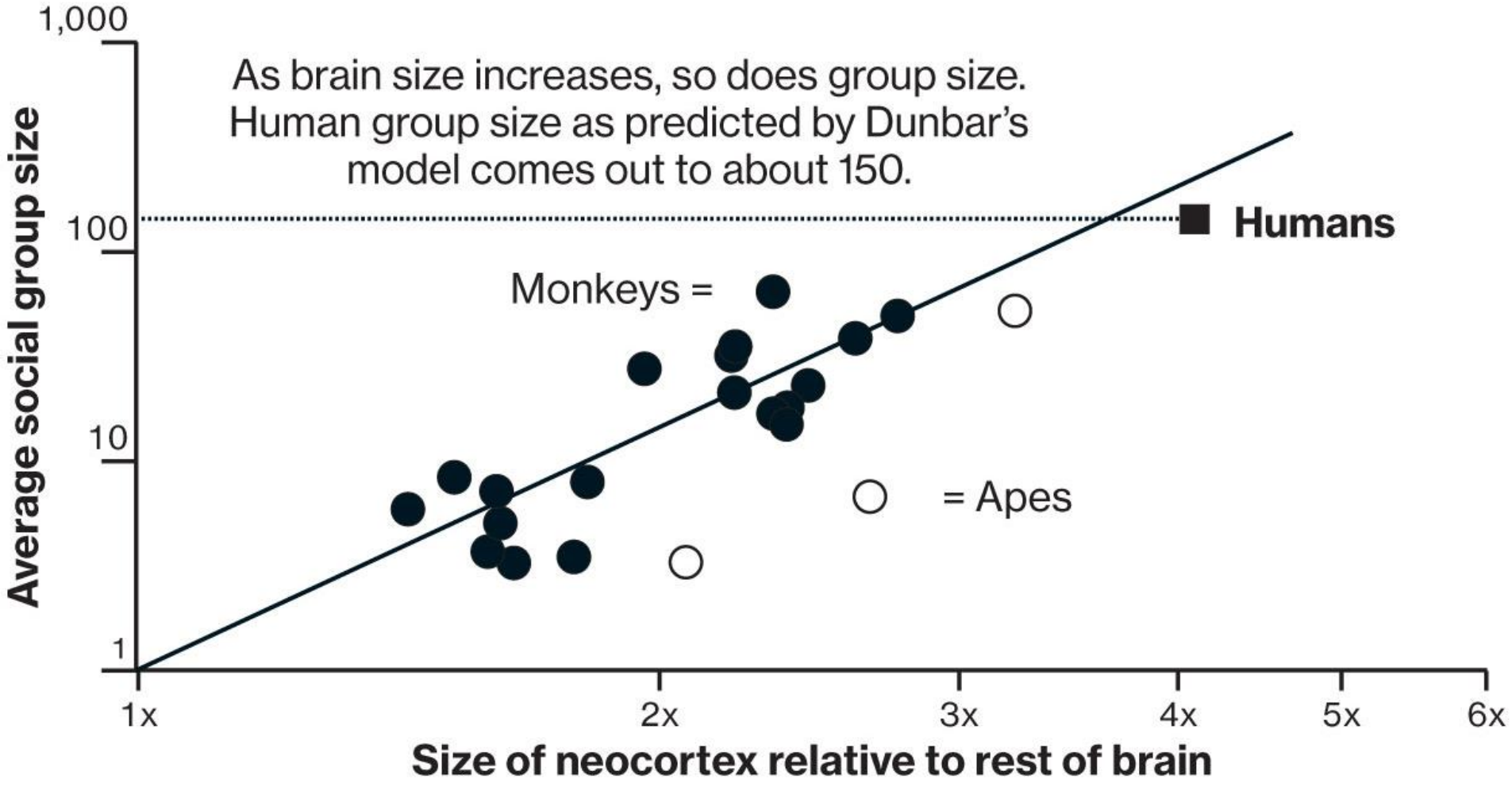
# Some numbers on brain circulation and energy metabolism

- 2% of the body weight
- 15% of the cardiac output
- 20% of total body oxygen consumption
- 25% of total body glucose utilization

- Global cerebral blood flow of 57 ml/100g/min
  → Extract 50% of oxygen and 10% of glucose from the arterial blood
  - Glucose utilization rate: 31 µM / 100 g / min
  - Oxygen utilization rate: 160 µM / 100 g / min

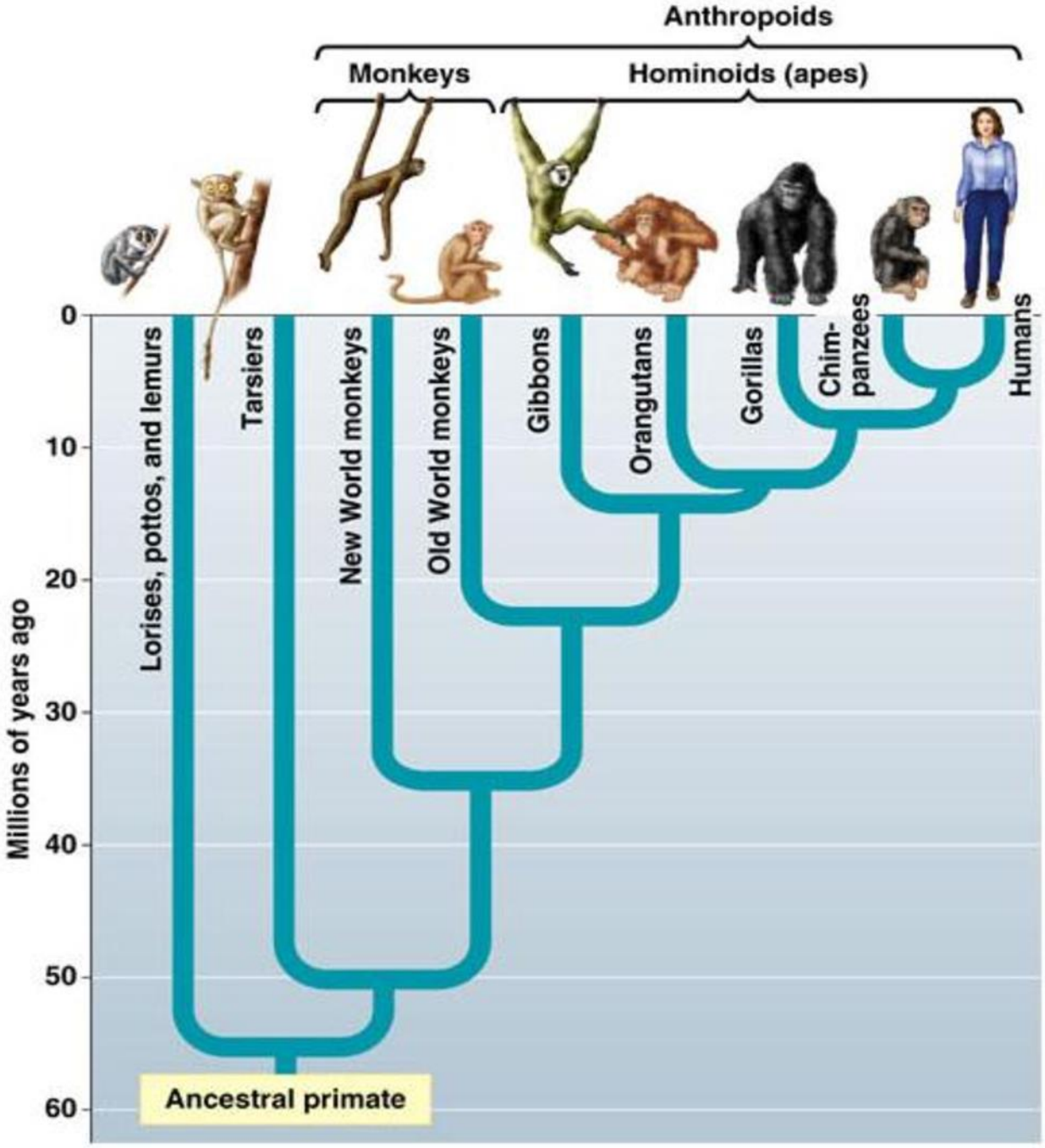
- Complete oxidation of glucose to $CO_2$ and $H_2O$
  - respiratory quotient = 1

# The Social Cortex



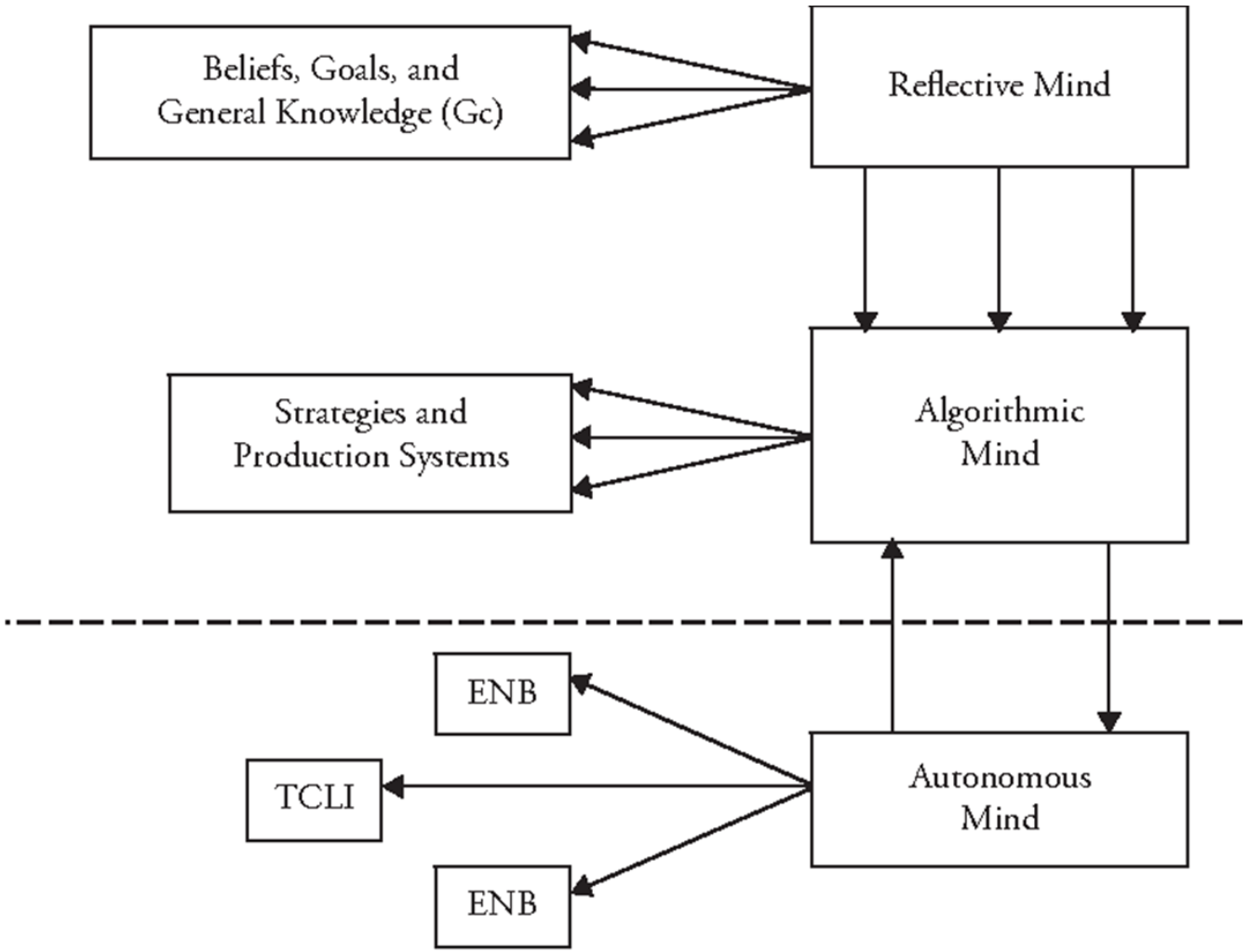
# Continuity of human intelligence

- Phylogenesis: Evolution of the human brain
  - Distinctive features
    - Dual pathways in the cortical processing
    - Lateralization between hemispheres
    - Expansion of the frontal lobe
- Ontogenesis: Development of individuals
  - Genetic programs
  - Epigenetic developments
    - The whole-some of personal experience
    - Through social and cultural interaction

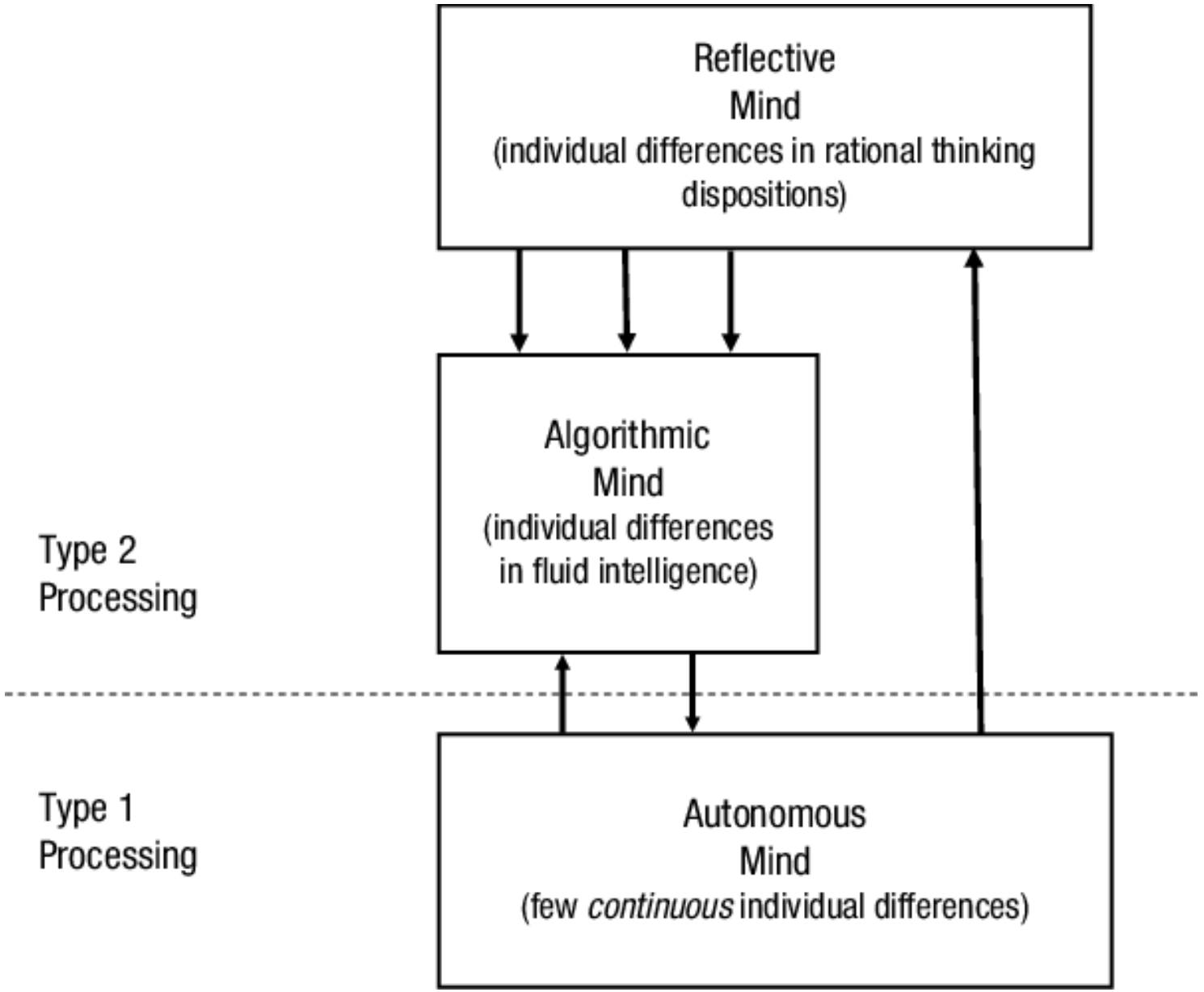

Anthropoids
Monkeys
Hominoids (apes)
Millions of years ago
0
10
20
30
40
50
60
Lorises, pottos, and lemurs
Tarsiers
New World monkeys
Old World monkeys
Gibbons
Orangutans
Gorillas
Chim-
panzees
Humans
Ancestral primate

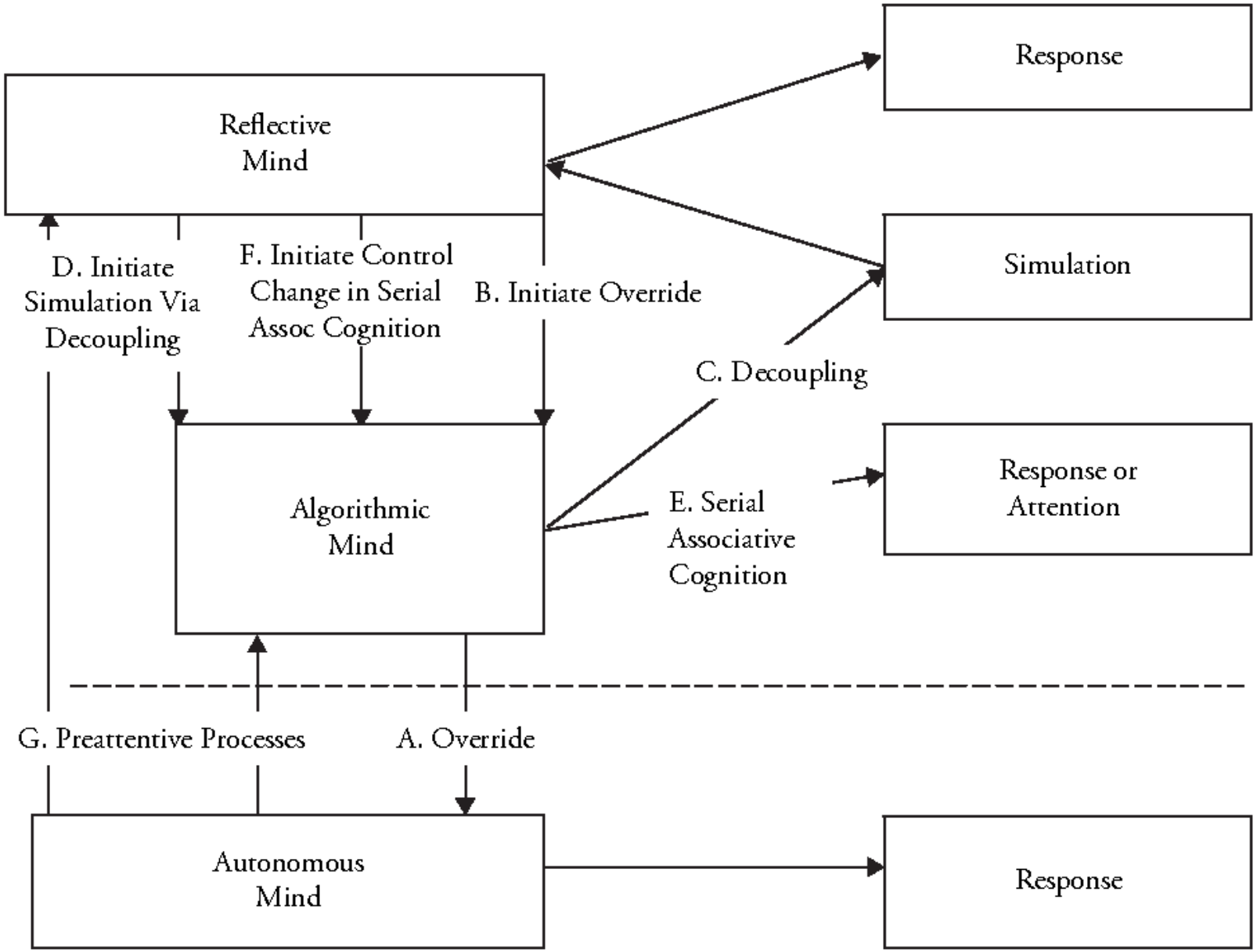
# Continuity of human intelligence

- Tripartite minds (Stanovich)
  - Autonomous mind
  - Algorithmic mind
  - Reflective mind
- The cognitive unconscious (Hayles)
- Origin of life




ENB = Encapsulated Knowledge Base
TCLI = Tightly Compiled Learned Information

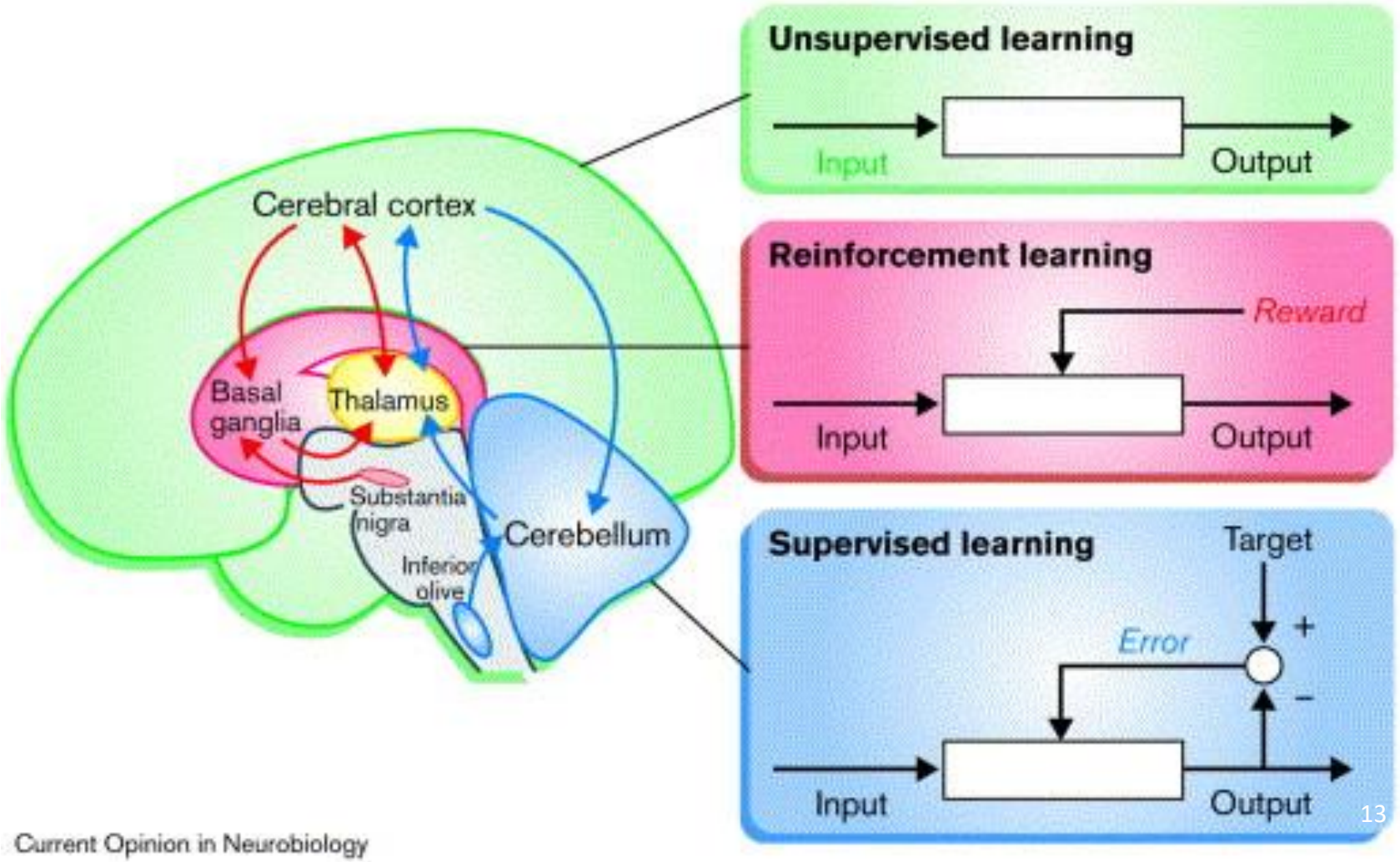

Reflective
Mind
(individual differences in rational thinking
dispositions)
Algorithmic
Mind
(individual differences
in fluid intelligence)
Type 2
Processing
Type 1
Processing
Autonomous
Mind
(few *continuous* individual differences)


Response
Reflective Mind
Simulation
D. Initiate Simulation Via Decoupling
F. Initiate Control Change in Serial Assoc Cognition
B. Initiate Override
C. Decoupling
Response or Attention
Algorithmic Mind
E. Serial Associative Cognition
G. Preattentive Processes
A. Override
Autonomous Mind
Response

# Functional neuroanatomy of human intelligence

- Data processing: Perception and Action
  → spinal cord, brainstem, cerebral cortex
- Modulation of data processing: Learning
  → cerebellum, basal ganglia, limbic system
- Control of perception/action and learning: Intelligence and Rationality
  → prefrontal cortex

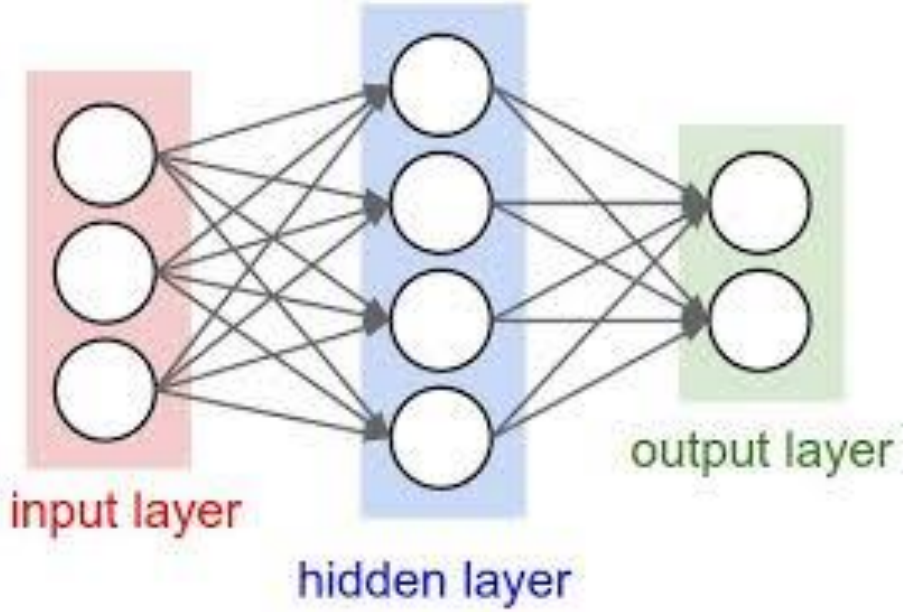
# Neural circuits for learning



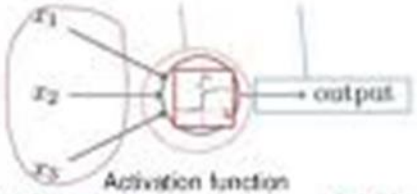

Current Opinion in Neurobiology

# Continuity between human and silicon-based intelligences

- Evolution of silicon-based intelligence
  - Perceptron
  - Connectionist PDP movement
  - Deep learning
    - Rectified linear unit (ReLU)
    - Convolutional neural network
    - Recurrent neural network
    - Big database
    - Computational power
- Technology as expanded human intelligence

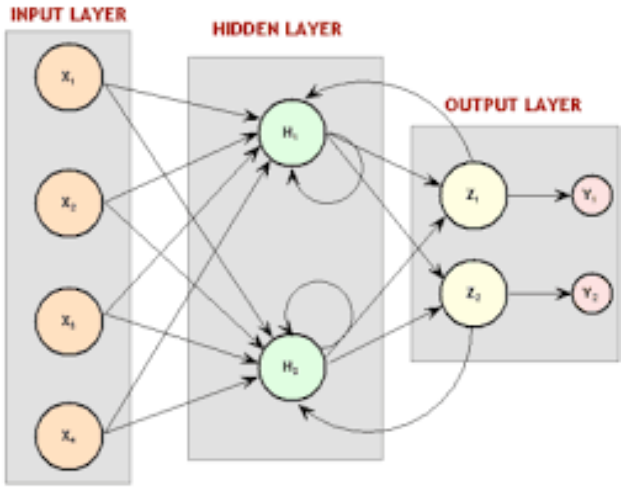


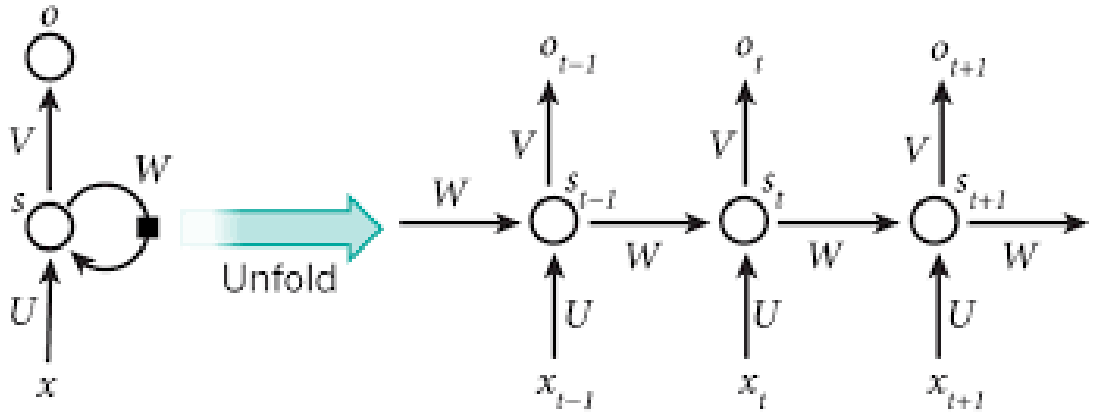

## Rectified Linear Unit (ReLU)



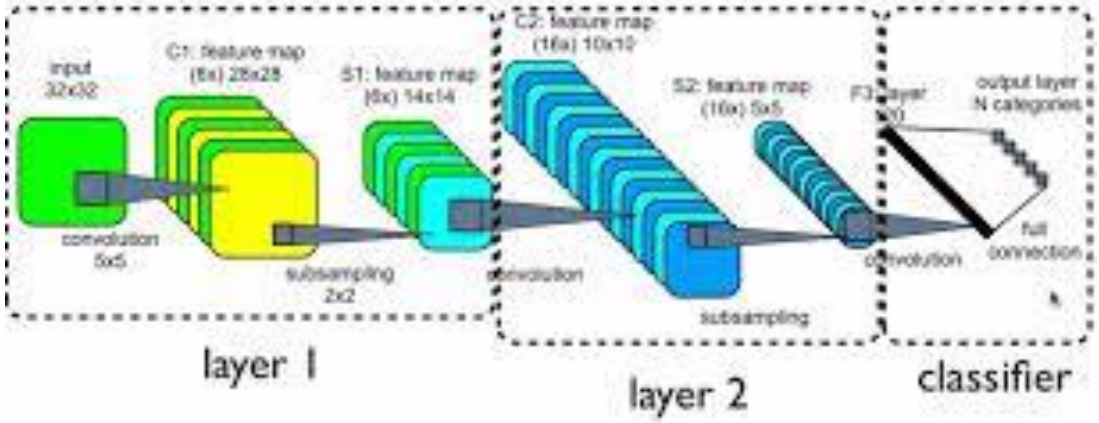

Sigmoid function

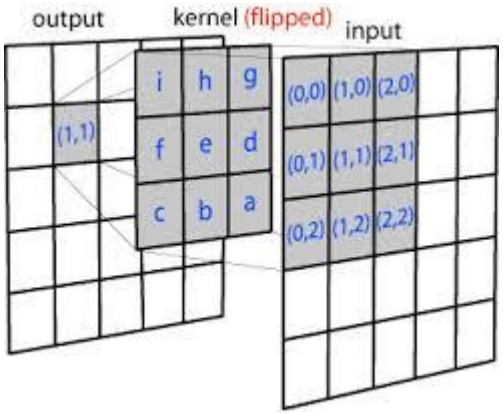
Rectified Linear Unit

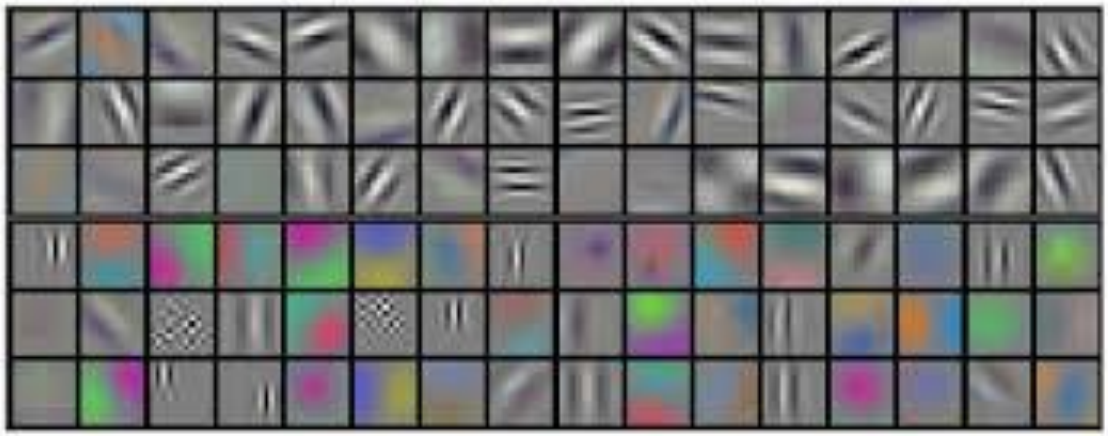

INPUT LAYER
HIDDEN LAYER
OUTPUT LAYER

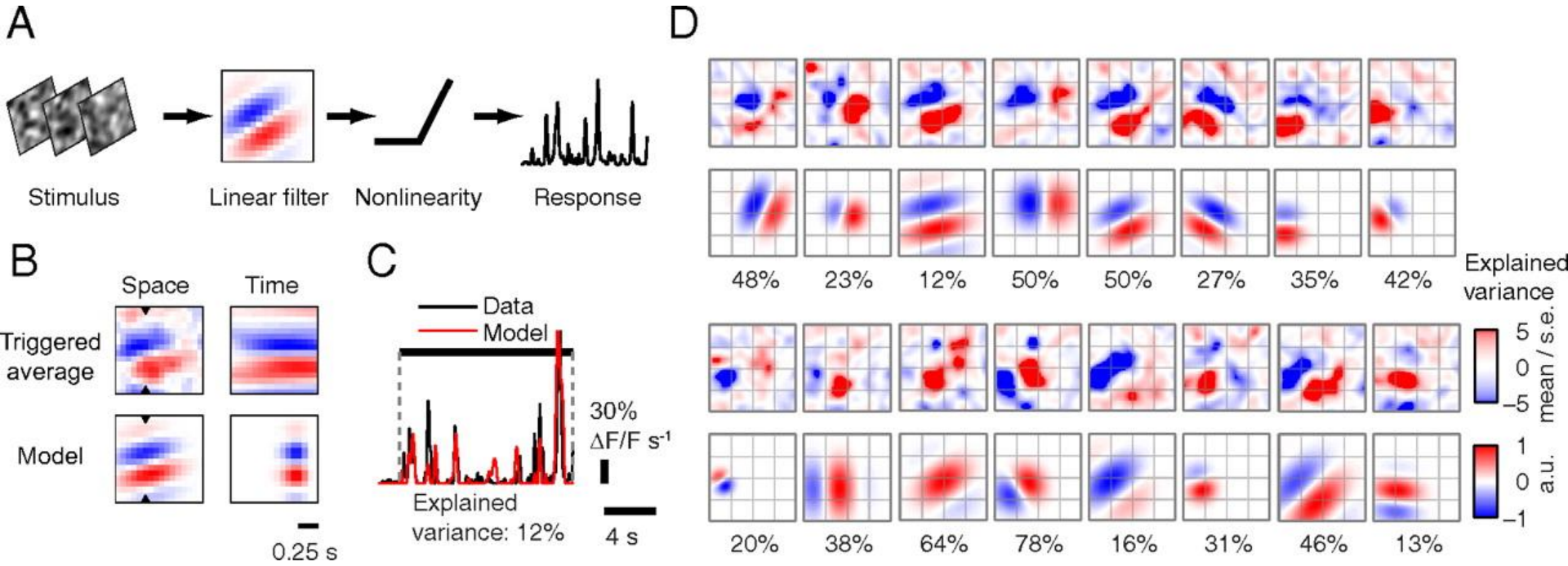

o
V
W
s
U
x
Unfold
W
V
U
W
W
W

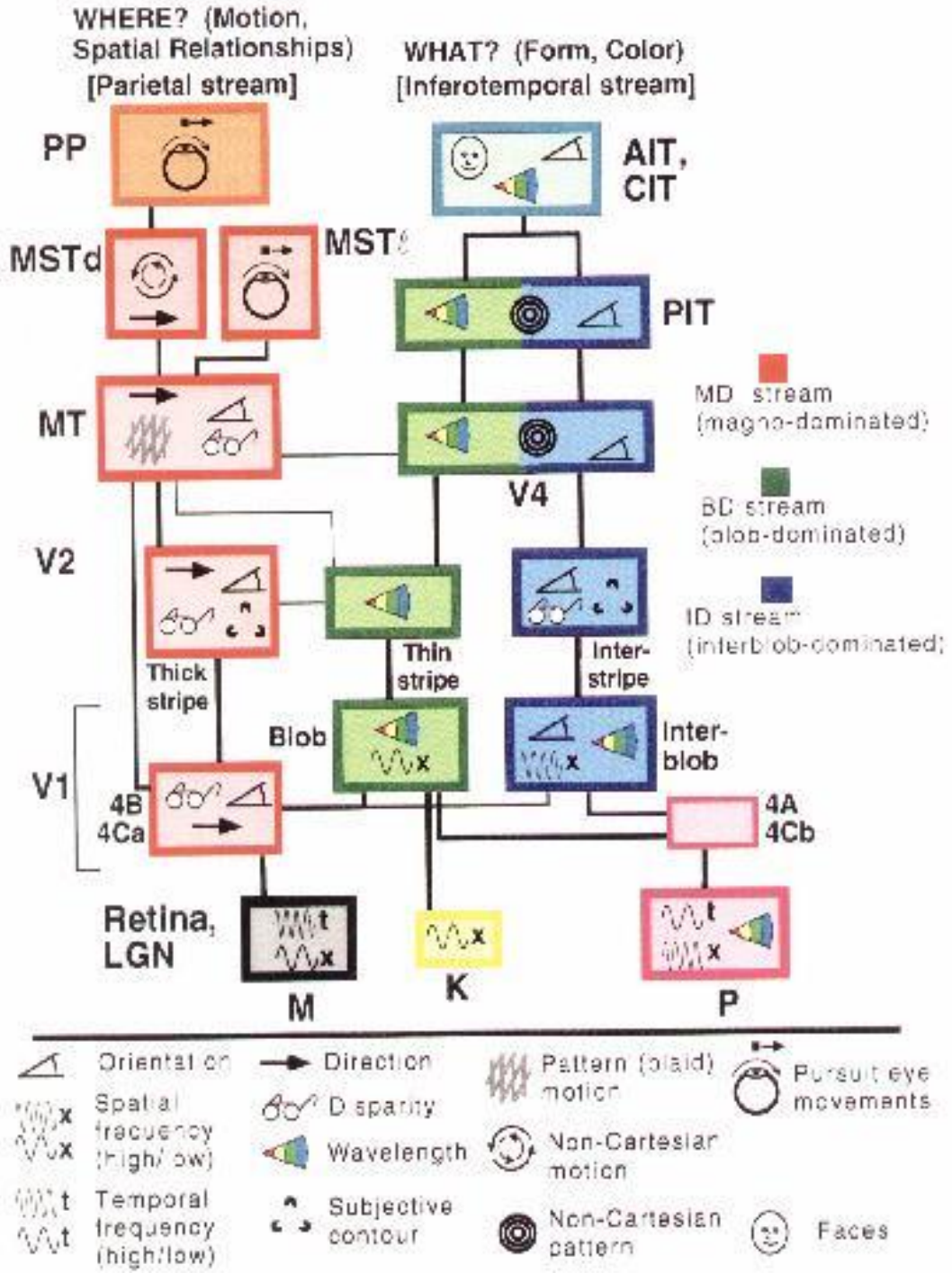
# Convolutional Neural Networks

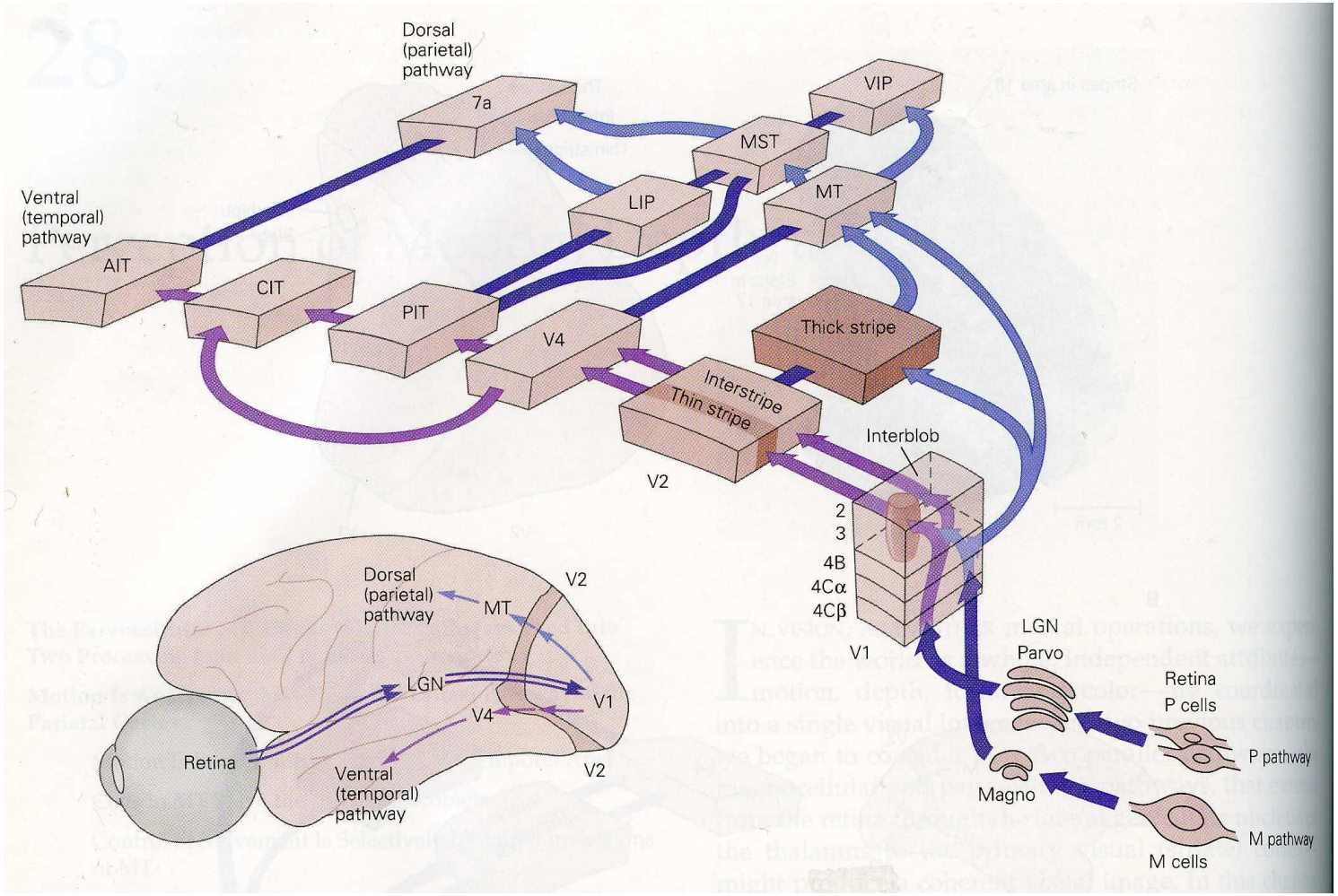



A
Stimulus
Linear filter
Nonlinearity
Response
B
Space
Time
Triggered average
Model
0.25 s
C
Data
Model
30%
ΔF/F s-1
Explained variance: 12%
4 s
D
48%
23%
12%
50%
50%
27%
35%
42%
Explained variance
5
0
–5
mean / s.e.
1
0
–1
a.u.
20%
38%
64%
78%
16%
31%
46%
13%


WHERE? (Motion, Spatial Relationships)
[Parietal stream]
WHAT? (Form, Color)
[Inferotemporal stream]
PP
AIT, CIT
MSTd
MST
PIT
MT
V4
V2
Thick stripe
Thin stripe
Inter-stripe
Blob
Inter-blob
V1
4B 4Ca
4A 4Cb
Retina, LGN
M
K
P
MD stream (magno-dominated)
BD stream (blob-dominated)
ID stream (interblob-dominated)
Orientation
Direction
Pattern (plaid) motion
Pursuit eye movements
Spatial frequency (high/low)
D sparity
Wavelength
Non-Cartesian motion
Temporal frequency (high/low)
Subjective contour
Non-Cartesian pattern
Faces


Dorsal
(parietal)
pathway
7a
VIP
MST
LIP
MT
Ventral
(temporal)
pathway
AIT
CIT
PIT
V4
Thick stripe
Interstripe
Thin stripe
V2
Interblob
2
3
4B
4Cα
4Cβ
V1
LGN
Parvo
Retina
P cells
P pathway
Magno
M pathway
M cells
Dorsal
(parietal)
pathway
MT
V2
LGN
V1
V4
Retina
Ventral
(temporal)
pathway
V2

# Materiality and continuity of human intelligence

- Brain is an actualized system
  out of an infinite set of possible intelligent systems
  by physical constraints.
- Unique constraints in the actualization
  - Ex 1: The minimum energy principle
  - Ex 2: The accuracy-speed tradeoff
    Concrete reality vs. abstract representation
  - Ex 3: The exploration-exploitation tradeoff
    Difference across vs. repetition in individuals

# Human intelligence by evolution

- Evolution through material constraints:
  Continuation of an intelligent system
  by reproduction of individuals
- Evolution by production (rather than survival) of the fittest
  = Secure diversity against inevitable unpredictability in the material world

# Evolution and ‘epi-evolution’ of human intelligence

- Trends in the evolution of Homo sapiens
  - Off-line & exteriorized memory (Stiegler)
  - Horizontal & vertical communications, i.e., language and education
  - Hierarchical cognitive system (Stanovich)
- Epi-evolutionary variation of human beings
  - Congenital and acquired variation
  - Socio-cultural: the Flynn effect, Vygotsky & Luria
  - Technology as expanded human intelligence

# Expanded human intelligence

- Expansion in terms of constitution includes brain → body → society → universe
  = Ecological intelligence
- Expansion in terms of contents includes rational thoughts (Stanovich)
  → the cognitive nonconscious (Hayles)
  → intelligent living and non-living processes (Bergson, Whitehead, Maturana, Varela)
  = Posthuman intelligence

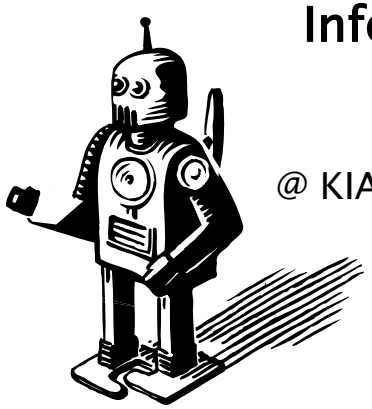
# Summary

- Anthropocentricism is based on the assumption that human beings are special in their unique mental abilities (= intelligence & rationality)
- Human intelligence is produced and constrained by materiality of the world.
- Consciousness, unconscious cognition, life processes, and life-less events are in a physical / evolutionary continuum.
- These taken together reject anthropocentricism as a rational thesis, from the viewpoint of epistemic / theoretical rationality.
- Arguments for anthropocentricism based on the practical rationality perspective also fall short, since the human condition as is and is expected by default is not sustainable (sub-optimal), inhumane (contrary to the humanistic claim of anthropocentricism), and just 'bad' (deontologically and sociologically).

Kanazawa Medical University

International Conference on Beyond Humanism: Artificial Intelligence, Information and Posthumanism

on December 17th, 2019
@ KIAS (Korea Institute for Advanced Study)



# Body-conservatism


Kojiro Honda
Associate Professor
Kanazawa Medical University



Mail to: kh-honda@kanazawa-med.ac.jp



JSPS Grant-in-Aid Fund(B) 16H03343


科研費 KAKENHI

# Index

1. Should we design robots like human body ?
2. The  Co-relation between Shape and Function
3. The Transhumanist Declaration
4. Mistrust in morphological freedom
5. Towards Body-conservatism
6. Concluding Remarks


Kanazawa
Medical
University

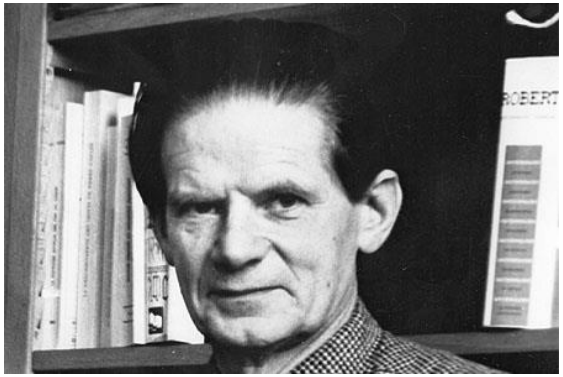
# 1. Should we design robots like human body ?

ロボットを人型にデザインするべきか？

# Technology as externalization of bodily function

- Evolution of tools means process of

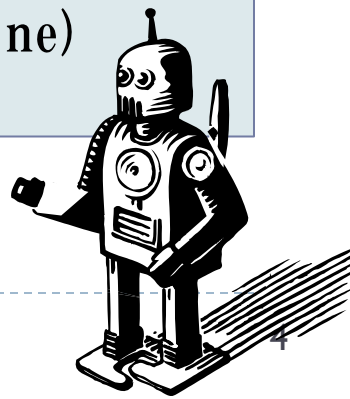
1. Externalization of our organ
2. Emancipation of our memories



André Leroi-Gourhan
(1911-1986)

Stone artifacts ⇒ Animal Machine(horse cart)
⇒ Automatic Machine 1(windmill, waterwheel )
⇒ Automatic Machine 2(steam engine)
……⇒ Information Technology

HISTORY OF TOOLS

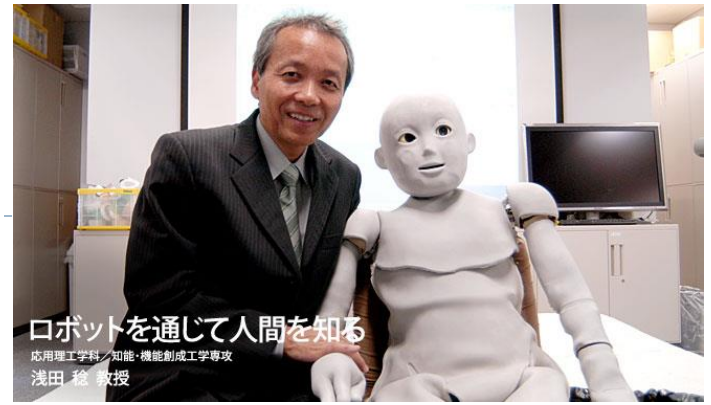


# Why Humanoids?

- Humanoid Boom in Japan
- Why do they try to make humanoid robots?
- This question has not been discussed in a serious manner.


Dr. Ishiguro and his copy geminoid

# Why Humanoids?: **Pro** - ASADA (2010)




- Minoru Asada (Osaka University)
- "Knowing Humanity through Robotics"
- <u>The cognitive developmental robotics approach</u>
- To understand
  - the development of increasingly complex cognitive processes in natural and

    artificial systems
  - how such processes emerge through physical/ social interaction

# Why Humanoids?: **Pro** - ASADA (2010)

- "I have got a inspiration such that, to understand humanity, it is useful to make human being in fact."

- Experiment on a human body is strictly banned

- The cognitive developmental robotics approach
  - making hypothesis about human development of intelligence or body
  - making robot according to the hypothesis
  - observing the robot's development of intelligence
  - validating the hypothesis

# Knowing = Doing

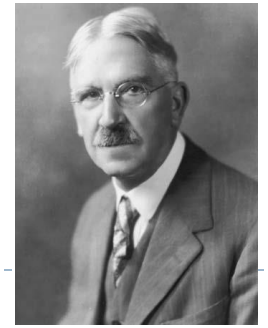
Bachelard(1931), Hacking(1983), Ihde(1991),Cartwright(1999), etc.

- Knowledge and human power are synonymous, since the ignorance of the cause frustrates the effect. For nature is only subdued by submission, and that which in contemplative philosophy corresponds with cause, in practical science becomes the rule.

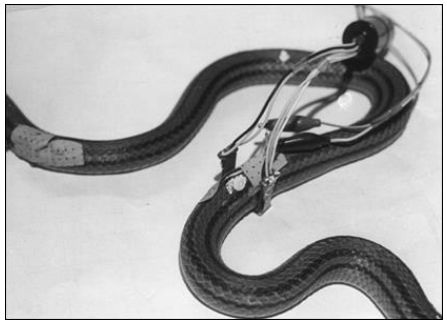
  -Francis Bacon, Aphorism 3, *Novum Organum*, 1620

- Knowing, for the experimental sciences, means a certain kind of intelligently conducted doing; it ceases to be contemplative and becomes in a true sense practical.

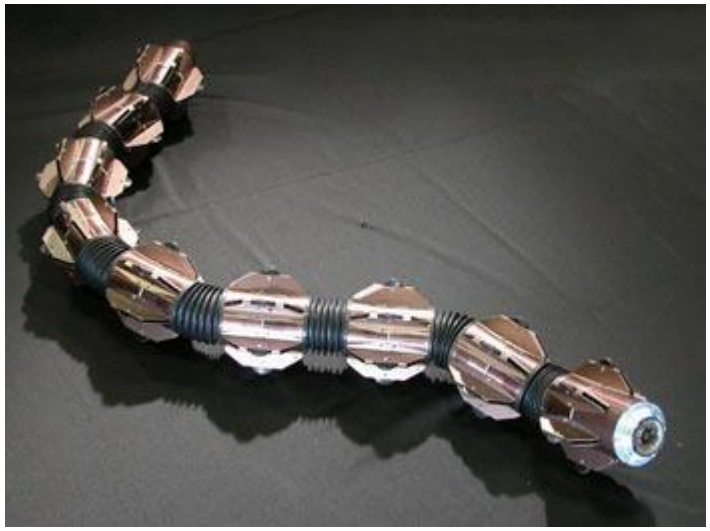
  -John Dewey, *Reconstruction in Philosophy*, p.70, 1920 (Dover 2004)



# Why Humanoids?: **Con** - Hirose (2011)



- Shigeo Hirose (Tokyo Institute of Technology)
- Recipient 2014 IEEE Robotics and Automation Award
- Development of Humanoid is
  - ① Not fitted together with natural evolution of technology
  - ② Not fitted together with the stream of evolution of whole technological system
  - ③ Not fitted together with the future generation's life



# Why Humanoids?: **Con** - Hirose (2011)

Ubiquitous robot

- It is better for us to make every artifact intelligent
- So in the future, robots will be out-of-sight in the nature of things
    - Not be human-like or human-shaped

Design of robots

- Robots should be designed according to their own purpose
- We should throw away **"humanoid fundamentalism"**

# **Why Humanoids?:** Main opinions

1. Because we want to know ourselves (Scientific Interest)

2. Because the human shape is congenial (Aesthetic Interest)

3. Because the human shape is adjustive to domestic environment (Functional Interest)

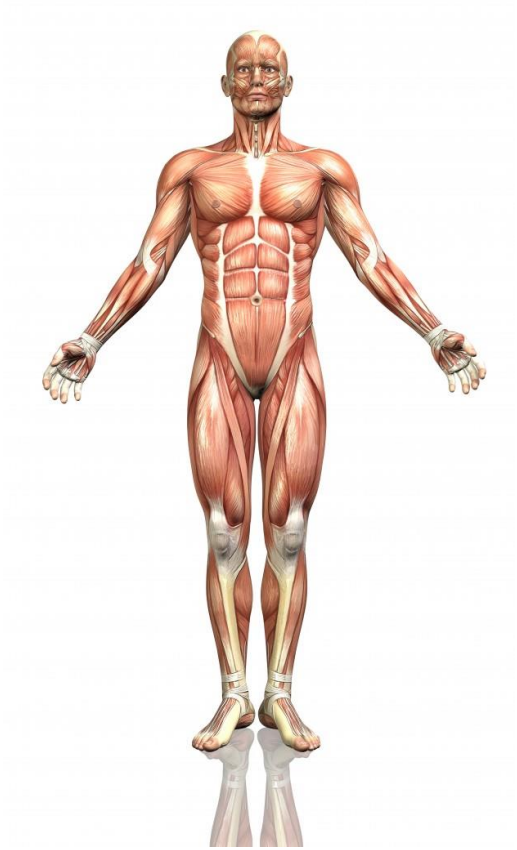
Kanazawa
Medical
University

# 2. The Co-relation between Shape and Function

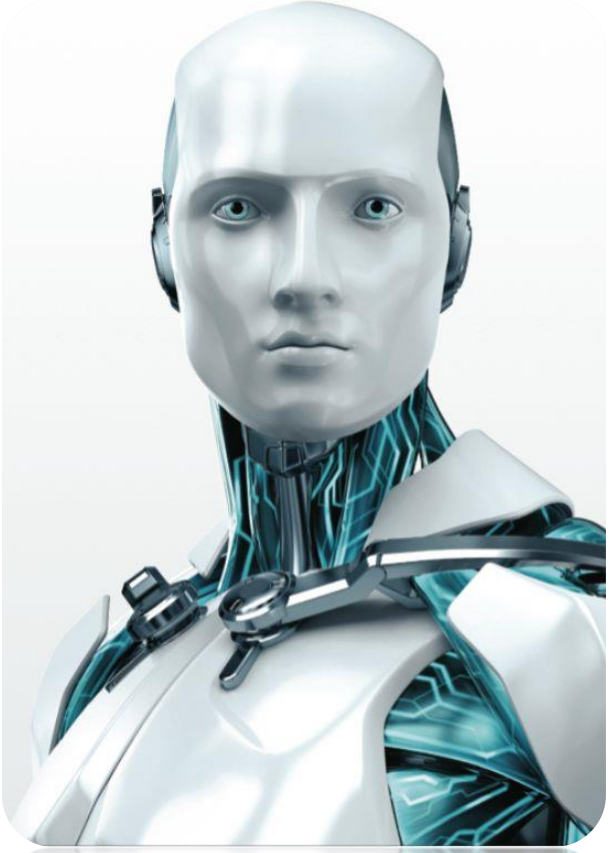
形と機能の相関性

# Humanoid Development

**Knowing-that**



**Knowing-how ?**

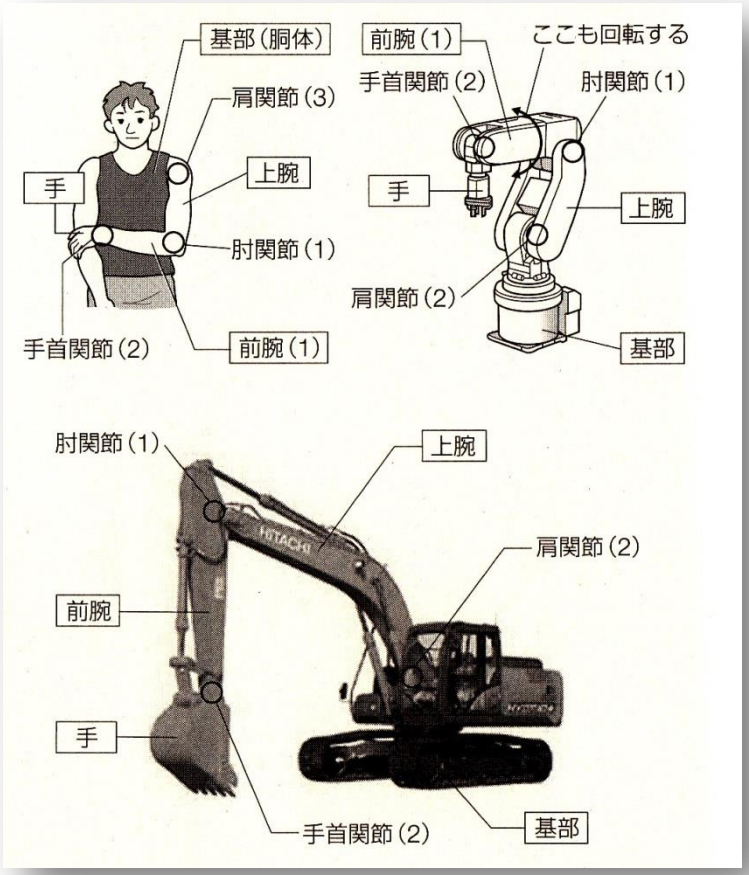


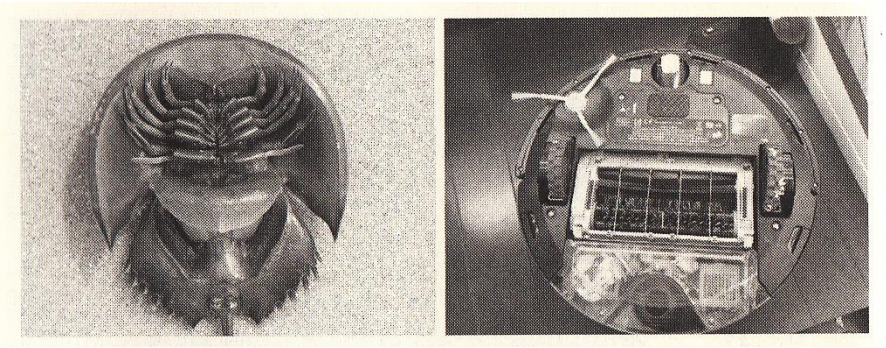
# Bio-mimic robotics

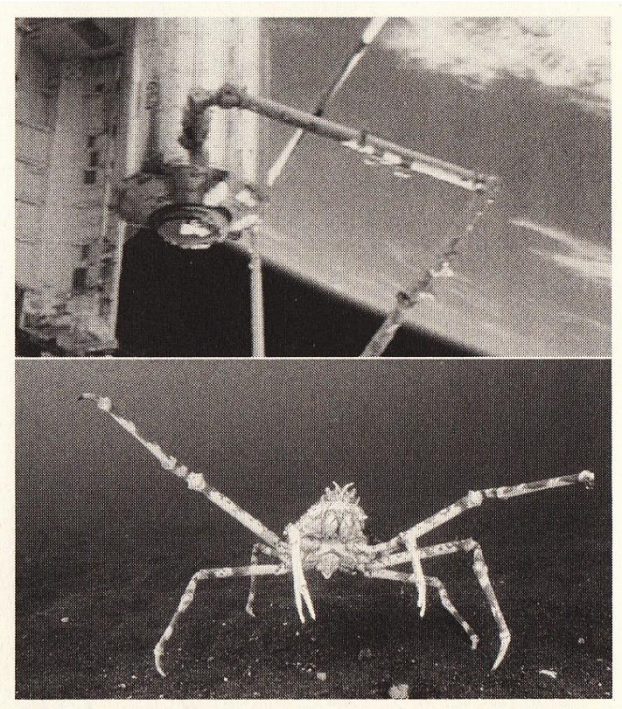
- Koichi Suzuki (Okayama University)
- Robots will mimic bio-structure by necessity

- Because
- Living organism has ultimate mechanical structure
  - Cost
  - Efficacy
  - Compactness

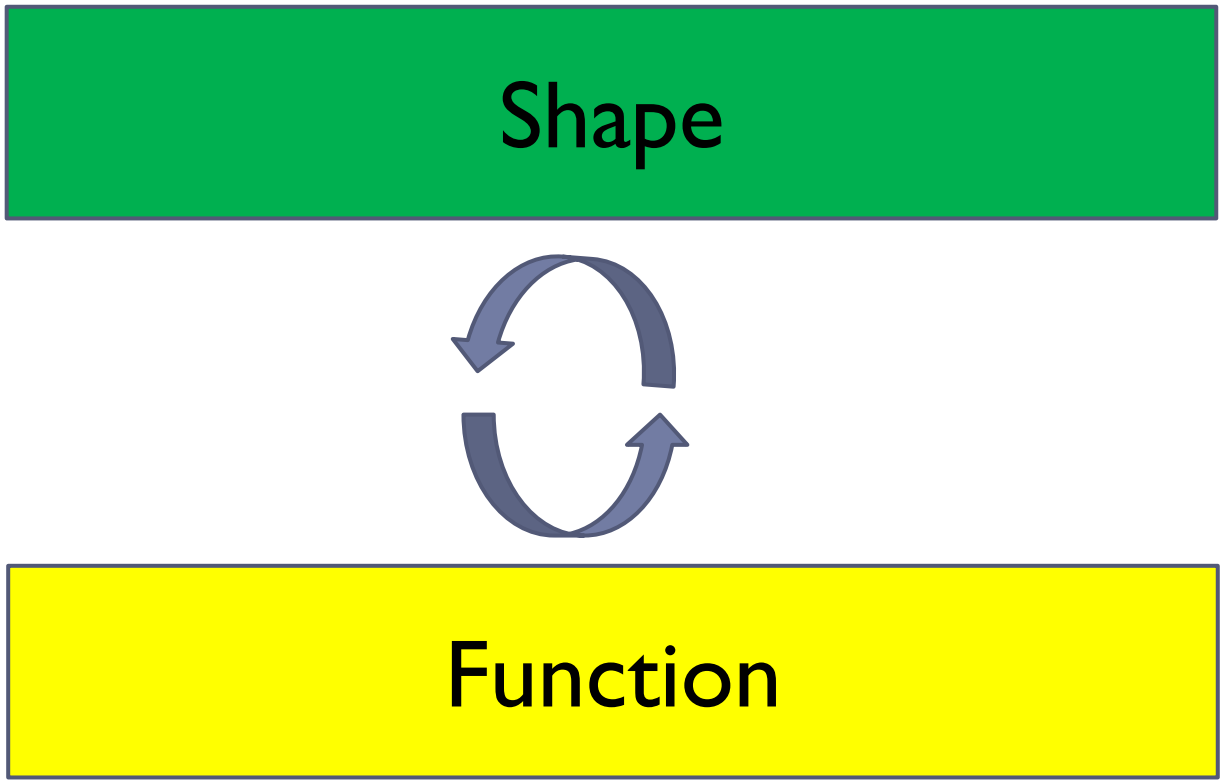

# Bio-mimic robotics　- Suzuki(2012)





# Function and shape are complementary

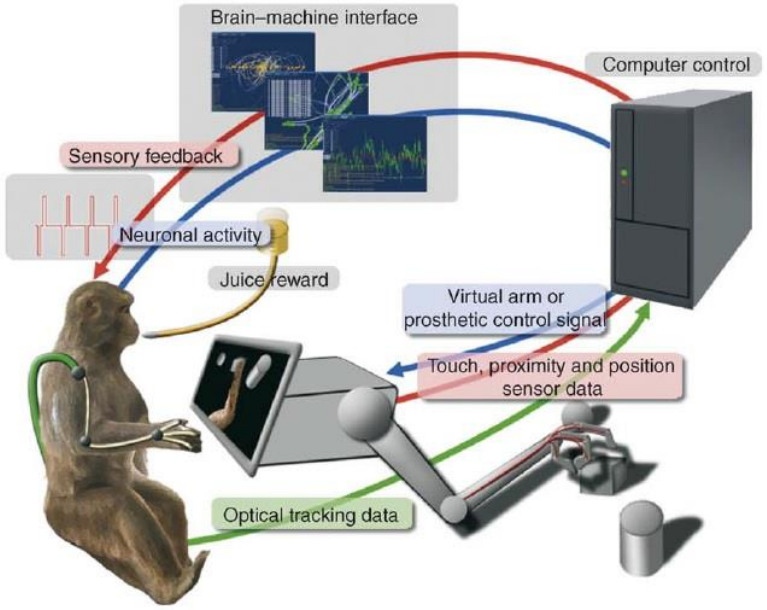


# Function and shape

- If we make human-mimic shape in a robot, then the robot functions like human body

- If we realize human body function in a robot, its shape seems like human being

# BMI - Lebedev & Nicolelis (2006)




# Brain-machine interface (BMI)

- Direct communication pathway between the brain and an external device

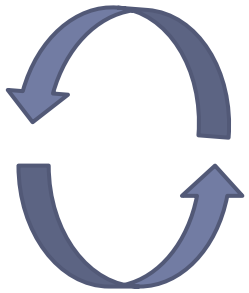
- BMIs have been primarily conceived as a potential new therapy to restore motor control in severely disabled patients

Kim, Park and Srinivasan (2009)

- The problem of BMI involves attempting to produce the intended motion from neurobiological signals.
- In doing so, we try to understand the mechanisms of the underlying neuromuscular system that connect neural signals to motion. Many aspects of the system are modeled in order to recreate the original biological motion. As an interesting byproduct, many of the control strategies proposed for the BMI robot can be applied to bio-mimetic robotics.

# Easy switch in robotics

Shape (Knowing-that)



BMI

Function (Knowing-how)

# Cybathlon 2016


Competition of Athletes with Prosthesis

【item】
ARM: Arm Prosthesis 義手
BCI: Brain Computer Interface
脳コンピューター連絡
FES: Functional Electrical
Stimulation Bike
電気刺激バイク
WHEEL: Wheel Chair 車椅子
EXO: Exoskelton 外骨格
LEG: Leg Prosthesis 義肢

# DARPA opens a new laboratory

## Neural Engineering System Design (NESD)

Dr. Al Emondi

The Neural Engineering System Design (NESD) program seeks to develop high-resolution neurotechnology capable of mitigating the effects of injury and disease on the visual and auditory systems of military personnel. In addition to creating novel hardware and algorithms, the program conducts research to understand how various forms of neural sensing and actuation might improve restorative therapeutic outcomes.

# Corporeal Intentionality and robots

- Bio-mimic structure easily makes artificial intentionality which is comparatively suitable to corporeal intentionality
- If Roboticists realize very human-like shape in a robot ( involuntarily maybe), the robot will be modified easily to artificial limbs, or prostheses
- Development of humanoids is complementary to development of prostheses or artificial body

# **Why Humanoids?:** Main opinions

1. Because we want to know ourselves (Scientific Interest)

2. Because the human shape is congenial (Aesthetic Interest)

3. Because the human shape is adjustive to domestic environment (Functional Interest)

4. Because humanoid technology could be applied to prostheses (Medical or Morphological Interest)


Kanazawa
Medical
University

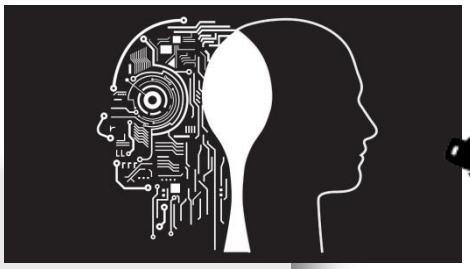
# 3. The Transhumanist Declaration

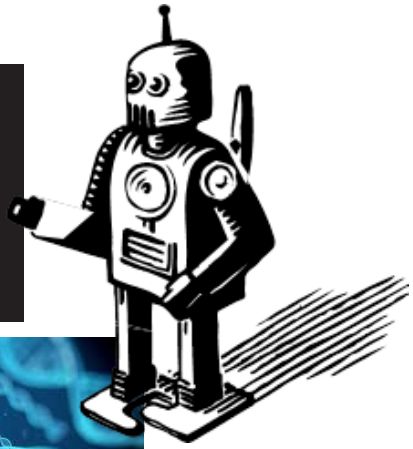
## トランスヒューマニスト宣言

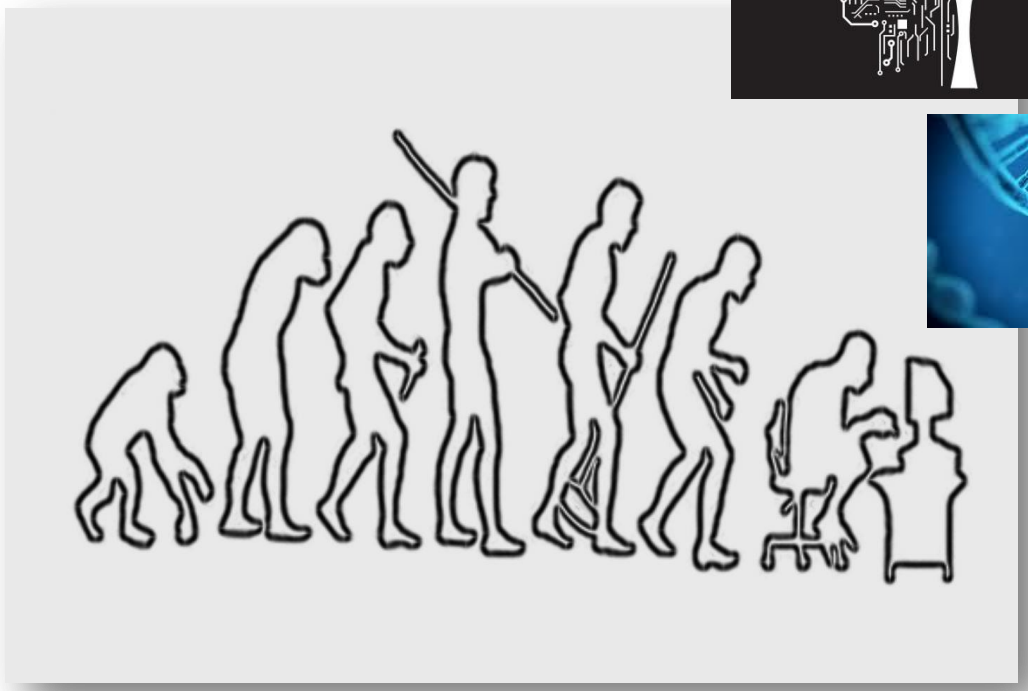
# We are in the era of "Internalization" of Technology

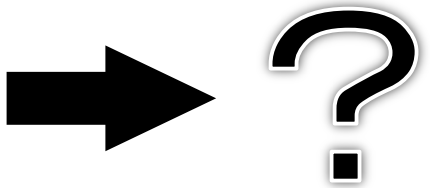
GNR

revolution









# GNR revolution ⇒
# Fusion of Medicine and Technology

- Externalization of bodily function

➔ Internalization of Technology

**Genetics**

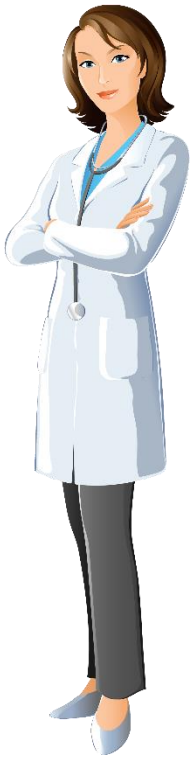
- drug discovery, gene modification, cosmetic

**Nanotechnology**

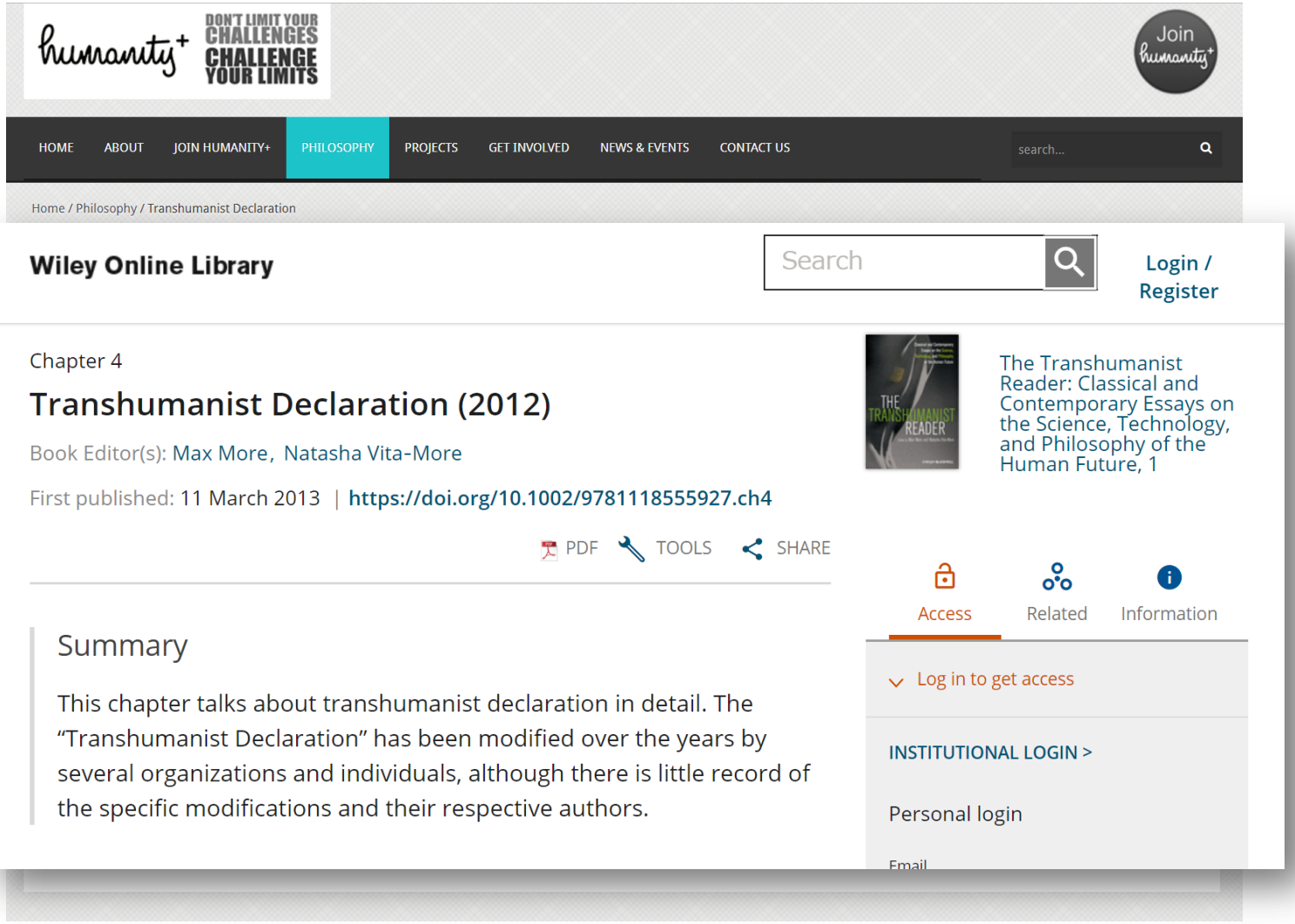
- anti-aging, enhancement of memory

**Robotics**

- prostheses, enhancement



# Transhumanist declaration (humanity+) 2012



# Transhumanist declaration (humanity+) 2012

1. Humanity stands to be profoundly affected by science and technology in the future. We envision the possibility of broadening human potential by overcoming aging, cognitive shortcomings, involuntary suffering, and our confinement to planet Earth.

2. We believe that humanity's potential is still mostly unrealized. There are possible scenarios that lead to wonderful and exceedingly worthwhile enhanced human conditions.

3. We recognize that humanity faces serious risks, especially from the misuse of new technologies. There are possible realistic scenarios that lead to the loss of most, or even all, of what we hold valuable. Some of these scenarios are drastic, others are subtle. Although all progress is change, not all change is progress.

4. Research effort needs to be invested into understanding these prospects. We need to carefully deliberate how best to reduce risks and expedite beneficial applications. We also need forums where people can constructively discuss what should be done, and a social order where responsible decisions can be implemented.

# Transhumanist declaration (humanity+) 2012

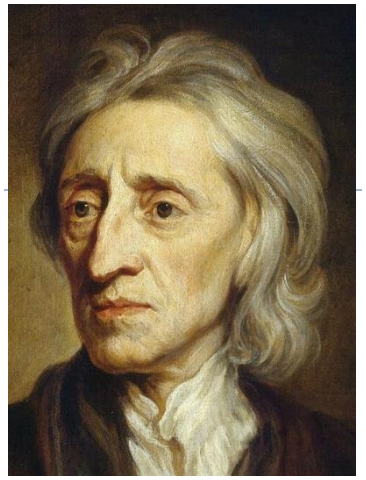
5. Reduction of existential risks, and development of means for the preservation of life and health, the alleviation of grave suffering, and the improvement of human foresight and wisdom should be pursued as urgent priorities, and heavily funded.

6. Policy making ought to be guided by responsible and inclusive moral vision, taking seriously both opportunities and risks, respecting autonomy and individual rights, and showing solidarity with and concern for the interests and dignity of all people around the globe. We must also consider our moral responsibilities towards generations that will exist in the future.

7. We advocate the well-being of all sentience, including humans, non-human animals, and any future artificial intellects, modified life forms, or other intelligences to which technological and scientific advance may give rise.

8. We favor morphological freedom-the right to modify and enhance one's body, cognition, and emotions. This freedom includes the right to use or not to use techniques and technologies to extend life, preserve the self through cryonics, uploading, and other means, and to choose further modifications and enhancement.

# John Locke (1690)
# State of Nature



- State of Nature＝

"That is a state of perfect freedom of acting and disposing of their own possessions and persons as they think fit within the bounds of the law of nature."

(Locke, *Two Treatises of Government*)

Notice: "within the bounds of the law of nature"

# Morphological Freedom:
# **Pro:** -Anders Sandberg (2013)

- If I want to have green skin, it is may own problem – nobody has the moral right to prevent me, but they do not have to support my ambition.

- As a negative right, morphological freedom implies that nobody may force us to change in a way we do not desire ore prevent our change. This maximizes personal autonomy.

# Morphological Freedom: **Pro**

- Ramez Naam

1. No technological boarder between therapy and enhancement
2. If banned, then black market
3. In democratic society, it is individual right to modify our body or mind
4. Impulse to improve ourselves has been natural for human being from the beginning of our races

Claim: We should evolve ourselves to more than human



Rmaez Naam (2010), *More Than Human*


Kanazawa
Medical
University

# 4. Mistrust in Morphological Freedom

## 形態的自由への疑義

# Morphological Freedom: **Con**

- Michael Sandel
  - ✓ Body-enhancing
  - ✓ Memory enhancement
  - ✓ Body height extended
  - ✓ Sexual selection of babies
  - ✓ Anti-aging
  - ✓ Lefe-extension

- Even if these technologies were safe, moral problems remain

Michael Sandel(2007) *The Case against Perfection*

# Morphological Freedom: **Con**

- Doping or Body-enhancement would spoil the virtues of athletes and the value of sports
- Success should not be inherited materially
- We would lower the value of "giftedness of life" by technologies

Michael Sandel(2007) *The Case against Perfection*

# Morphological Freedom: **Con**

- Leon Kass

  - Identity crisis

  - Self-alienation

  - We would lose something precious which deserve human dignity such as natural reproduction, life-cycle, adoration for opposite sex, effort…etc.

Leon R.Kass(2005 ) *Beyond Therapy*

Kanazawa
Medical
University

# 5. Toward Body-conservatism

身体保守主義へ

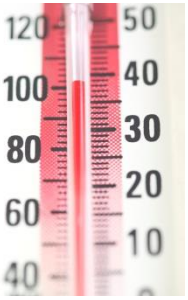
# What Kind of mediation ?

Don Ihde

- Don Ihde(1990)

"Technological Mediation"

1. Embodiment relations
   **(I －Technology) → World**
2. Hermeneutic relations
   **I → (Technology －World)**
3. Alterity relations
   **I → Technology(－World)**
4. Background relations
   **I (－Technology ／World)**




# A new Kind of mediation

- Peter Paul Verbeek (2008)

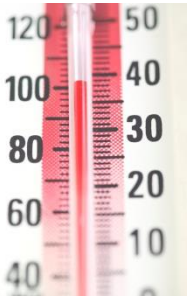
  "Cyborg Relations" & "Cyborg Intentionality"

1. Embodiment relations

   **(Human －Technology) ➔ World**

2. Hermeneutic relations

   **Human ➔ (Technology － World)**




3. Alterity relations

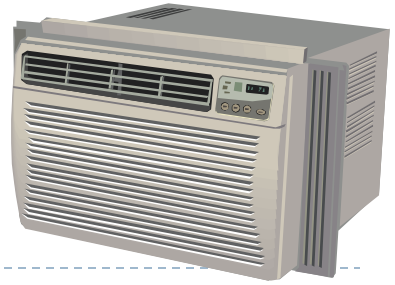
   **Human ➔ Technology(－World)**

4. Background relations

   **Human (－Technology ／World)**

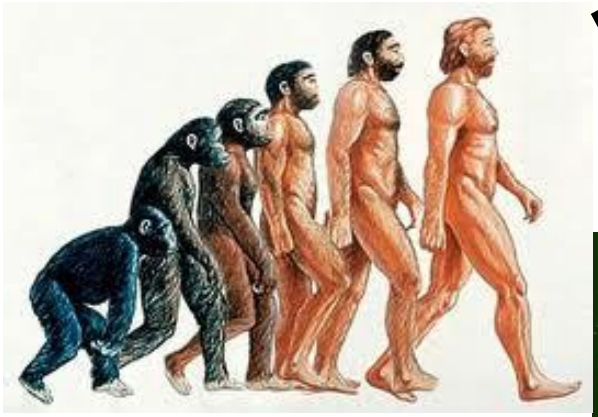
5. Cyborg relations

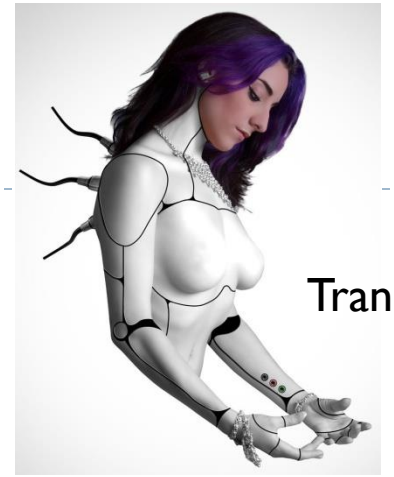
   **(Human／technology) ➔ World**



# Body - Umwelt

Human Body





Trans-human body

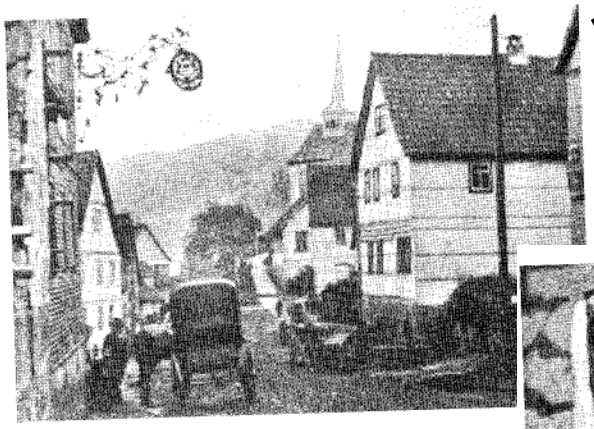
Fly body

Slug body

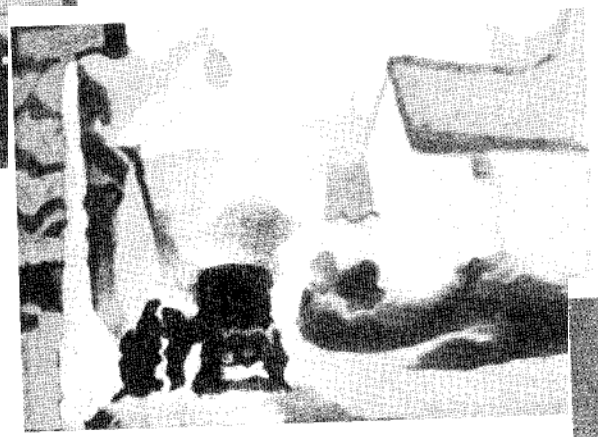
# Body - Umwelt

Trans-human Umwelt

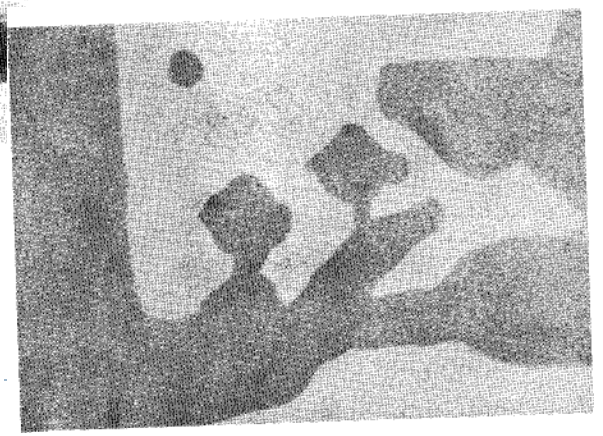
Human Umwelt



Umwelt of Fly



Umwelt of Slug

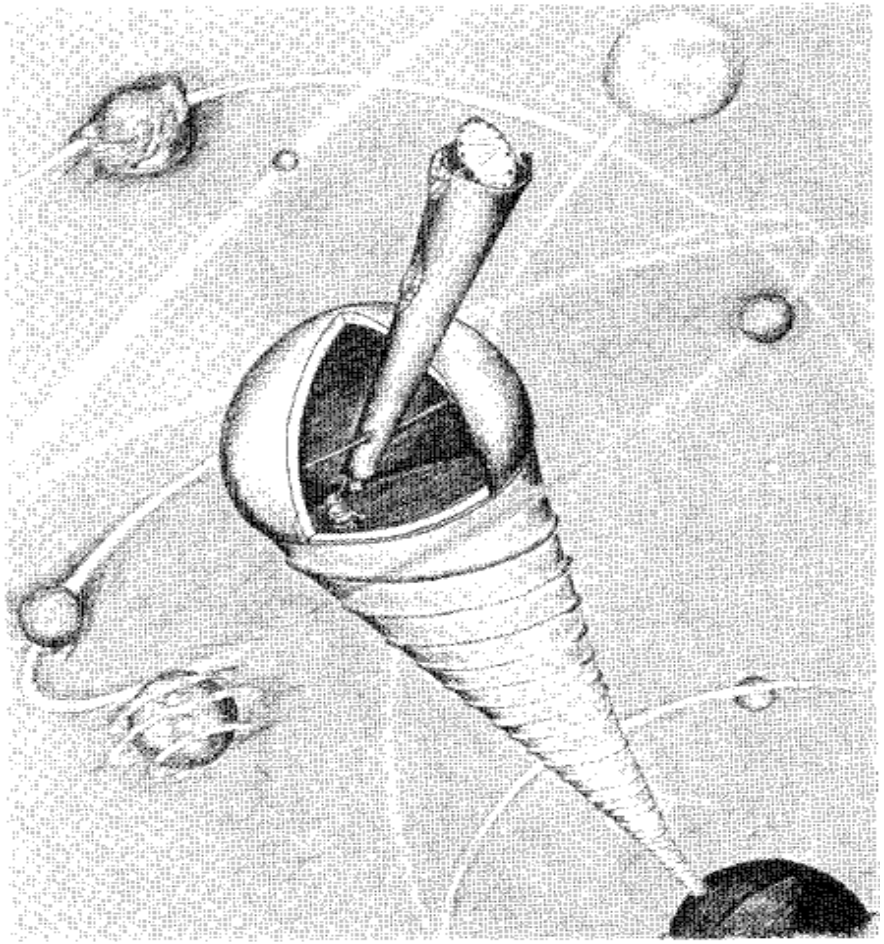


Jakob von Uexküll (1934)

# Umwelt completely mediated by technology

- ex)

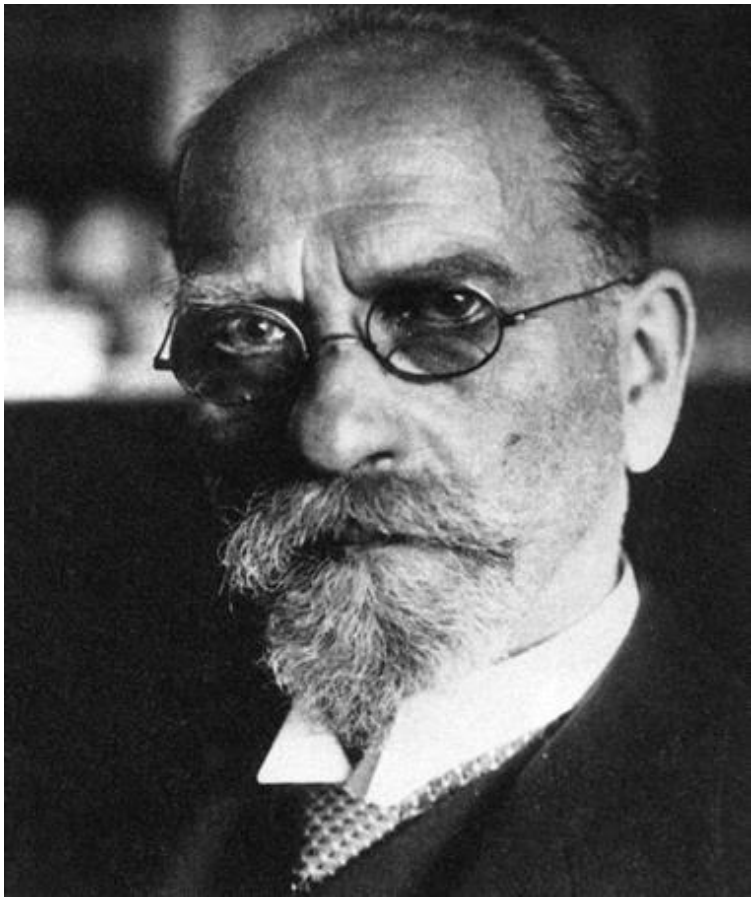
Astronomer's Umwelt



Jakob von Uexküll (1934)

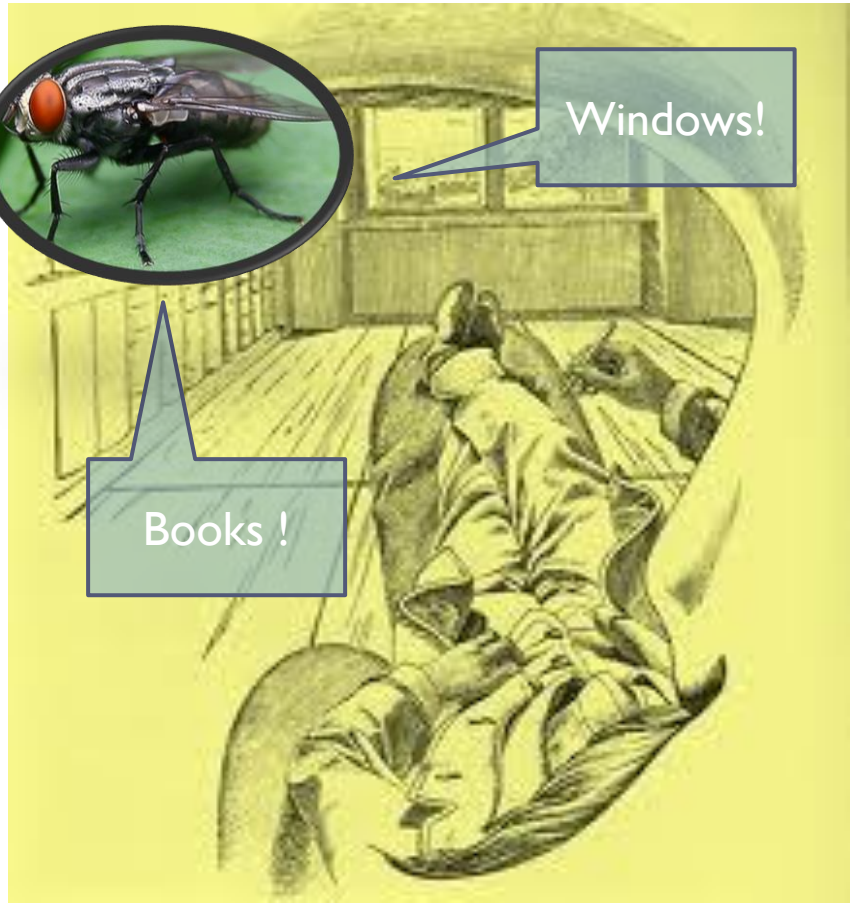
# Human umwelt – Human body

human umwelt

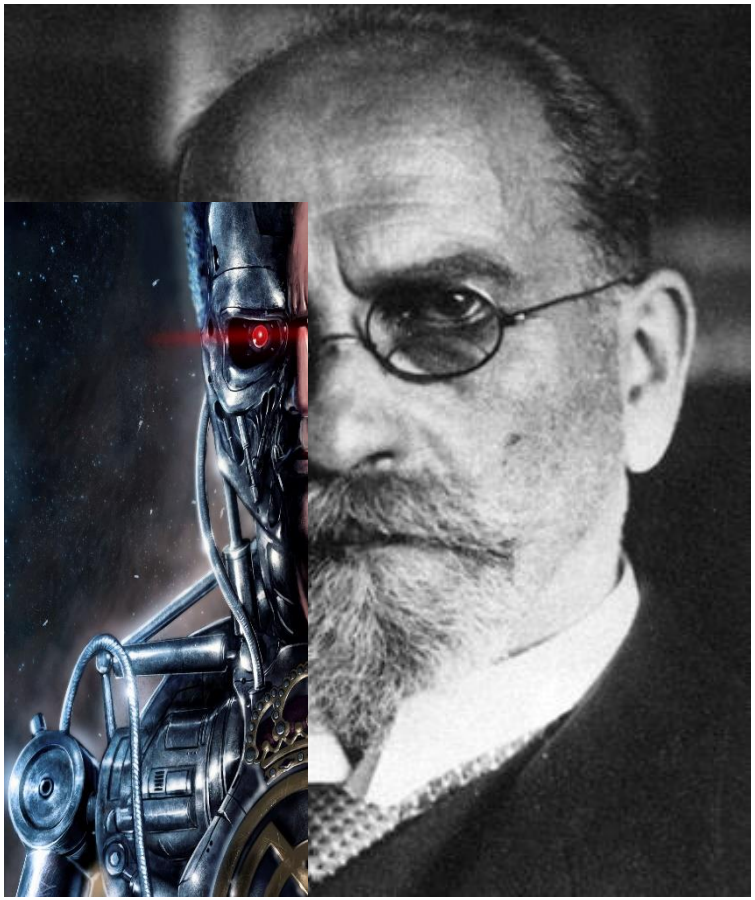


human body

# Transhuman umwelt – Transhuman body




trans-human umwelt



trans-human body

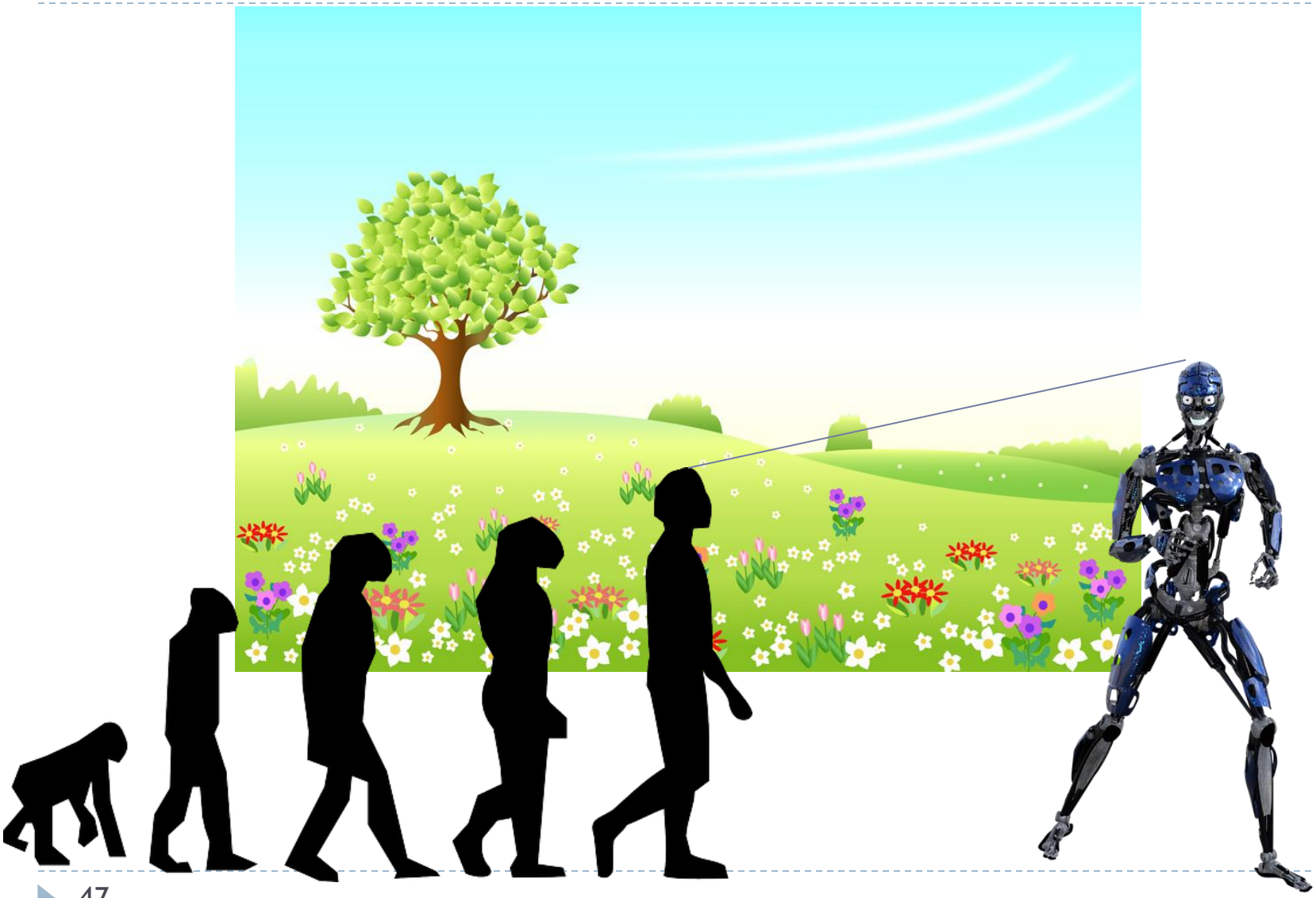
# Cyborg Intentionality (CI)

- = Body-intentionality +
  Technological intentionality

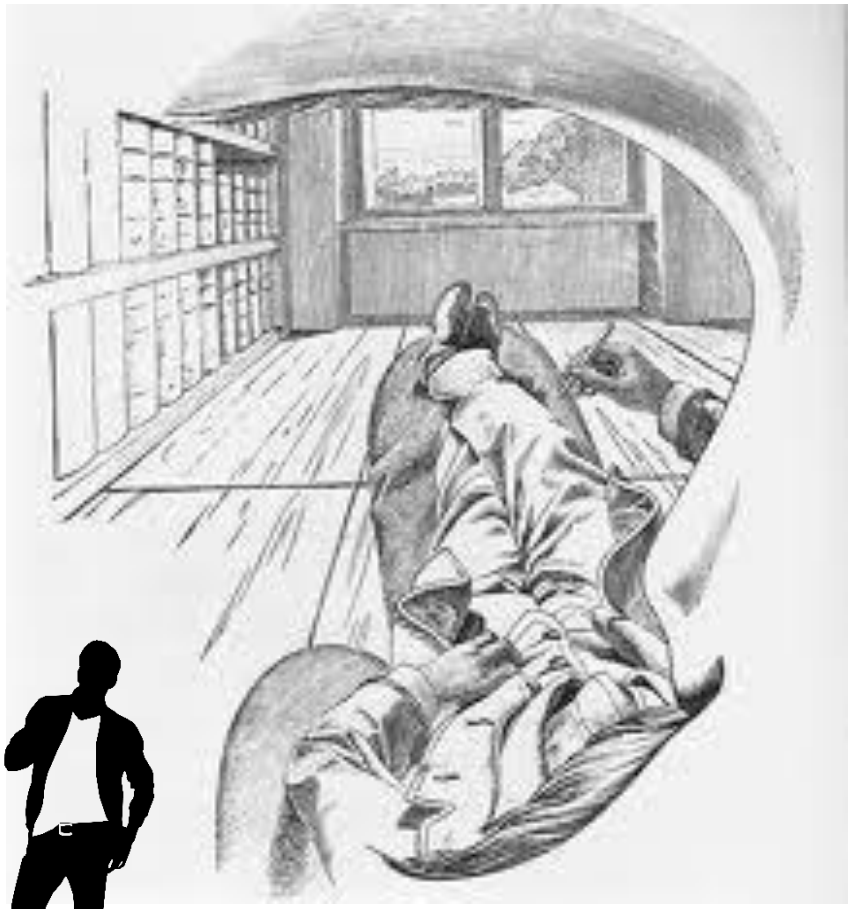

- CI will transform our own existence and appearance of the world

- CI will let us disable to go back to the common platform of our experience

# World:
# Exactly the same as before?

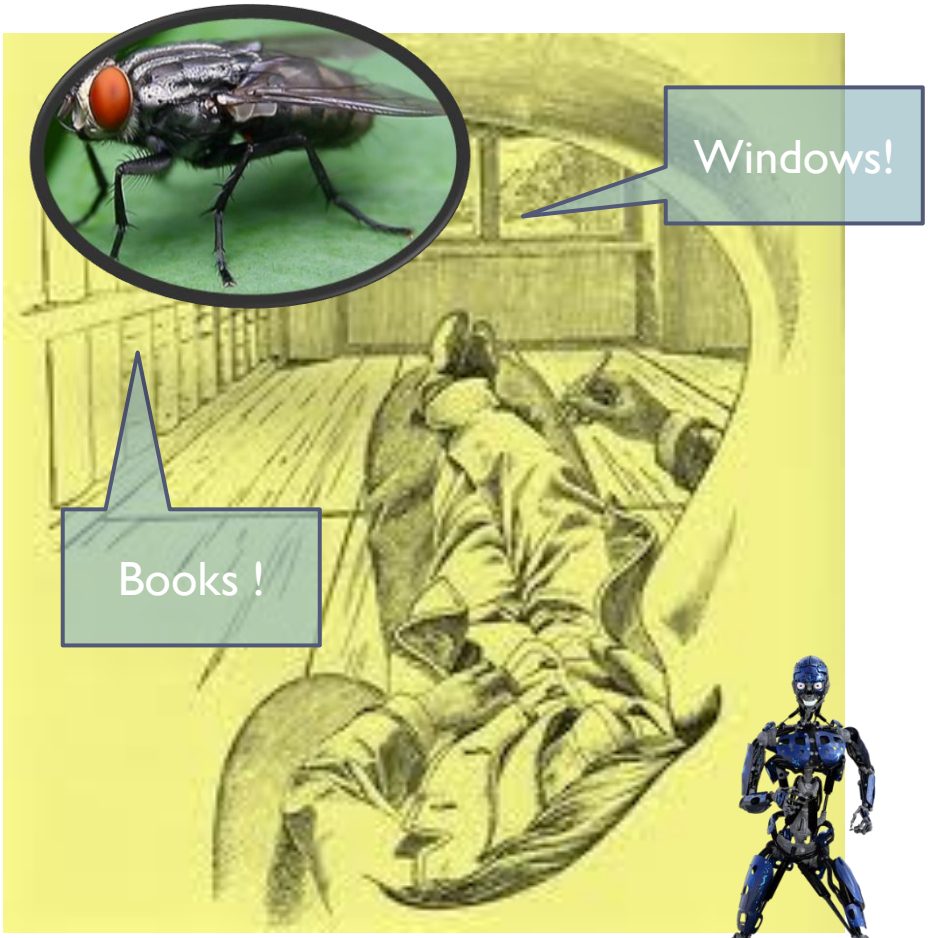


# Enhancement technology could result in: **World divide**





human umwelt

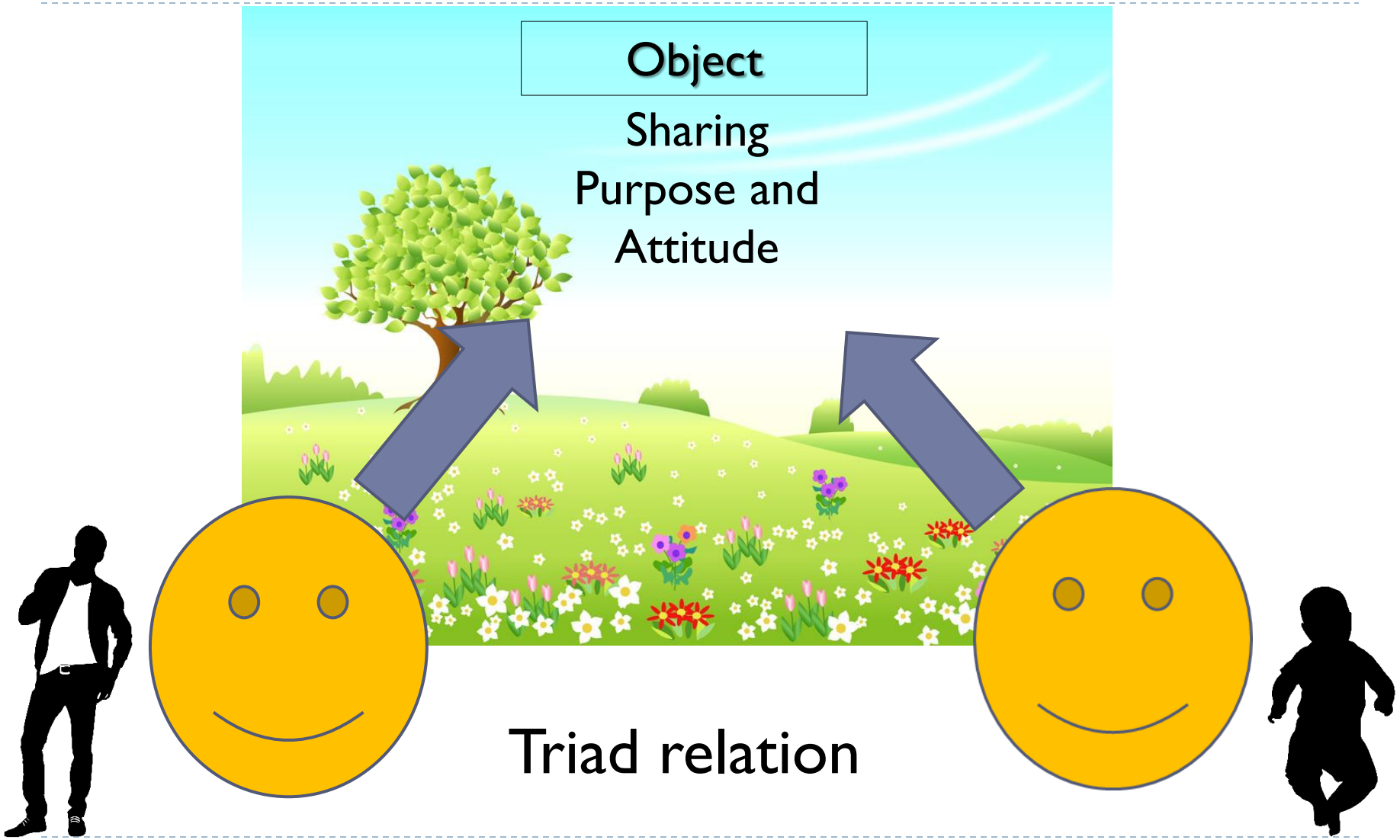





trans-human umwelt

# Cultural Inheritance based on bodies

## **We need the same world**



# Gesture and "Imitative Community"

The sense of the gestures is not given, but understood, that is, recaptured by an act on the spectator's part. The whole difficulty is to conceive this act clearly without confusing it with a cognitive operation.

The communication or comprehension of gestures comes about through the reciprocity of my intentions discernible in the conduct of other people. It is as if the other person's intention inhabited my body and mine his.

Maurice Merleau-Ponty, 1945


Kanazawa
Medical
University

# 6. Concluding Remarks

# Result of Transhumanism

- Modifying our body could result in "World Divide"

- When we can not share the  same world, we will also lose our platform of compassion.  At that time we are not residents in "imitative community" any more

# Body-conservatism　身体保守主義

- Body-conservatism is necessary to prevent **world-divide**

- Limits of body-modification　人体改造に一定の制限を設ける。
- Limits of social implement of enhancement technology

人体改造技術の急激な社会実装を制限する。

- Enhancement technology as therapy

人体改造技術の使用を「治療」の範疇におさめる。

- Discretion for irreversible modification

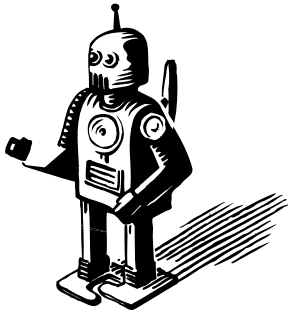
治療するにしても、元の自分が何であったか、分からなくなるような治療はしない。

- Big problem : What is the standard of human body ?

標準的人間身体の基準

  - We don't have standard ➡ So we can modify （Transhumanism）
  - 〃 ➡ So we should consign our body to nature (Body-conservatism)

# Thank you for your attention!



**kh-honda@kanazawa-med.ac.jp**

**AMED**
希少難治性脳・脊髄疾患の歩行障害に対する生体電位駆動型肢装着型補助ロボット(HAL-HN01)を用いた新たな治療実用化のための多施設共同医師主導治験
【研究代表:中島孝(国立新潟病院)】

JSPS Grant-in-Aid
Fund(B)JP16H03343
日本型「ロボット共生社会の倫理」のトランスディシプリナリーな探求と国際発信(研究代表:神崎宣次)

# Why is the Husserlian Notion of "Intentionality" Needed by Artificial General Intelligence?

Yingjin Xu

**E-mail**: yjxu@fudan.edu.cn
**Affiliation**: School of Philosophy, Fudan University
**Address**:
220 Handan Road,
School of Philosophy, Fudan University, Shanghai, China, 200433

# Why is the Husserlian Notion of "Intentionality" Needed by Artificial General Intelligence?

**Abstract**

Intentionality is required by any intelligent system, given that intelligence requires intentionality-presupposing capacities of revising beliefs in accordance with environmental changes. However, mainstream Anglophone philosophical theories of intentionality is not illuminating for Artificial General Intelligence (AGI) because they either appeal to external environmental factors which cannot be internally modeled or they cannot handle gradual transitions among different cognitive states. Hence, the needed theory of intentionality has to view mental contents as something which could be detached from external reality on the one hand, and view psychological modes as something permitting gradual mutual transformations among them on the another. The two requirements will naturally lead us to Husserl's notion of "phenomenological *epoché*" and an inferentialist interpretation of his notion of "noema", both of which could be algorithmically reconstructed via Non-Axiomatic Reasoning System (NARS).

**Key Words and Phrases**

General Artificial Intelligence (AGI); the box-approach; externalism; internalism; inferentialism; psychological mode; Non-Axiomatic Reasoning System (NARS); phenomenological *epoché*; noema

## 1. Introduction

Although there is a sizable body of literature at the intersection of phenomenology and cognitive science, [①] there are not so many studies intended to clarify the relationship between Edmund Husserl, the founding figure of the entire phenomenological movement, and Artificial General Intelligence (AGI) ,[②] which bears affinities with cognitive science in many aspects. [③] The main motivation for marginalizing Husserl in the circle of the so-called "naturalized phenomenologists" seems to be based on the following syllogism:

1. The most promising way to build the alliance between phenomenology and cognitive science or AGI is to appeal to notions like "embodiment", "embeddedness", "extendedness" and "enactedness", as summarized as "4E-ism" by Mark Rowlands[④], and all of these notions *cannot* be well treated in the framework of symbolic AI or "good-old-fashioned AI" (abbreviated as

---

[①] Representative literature include: Francisco J. Varela. *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: The MIT Press, 1992; Evan Thompson. *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*, Cambridge, MA: Belknap Press, 2010; Jean Petitot et al. eds. *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*, Stanford, CA: Stanford University Press, 2000; Shaun Gallagher. & Daniel Schmicking. eds. *Handbook of Phenomenology and Cognitive Science*, Dordrecht: Springer, 2010, etc..

[②] AGI literally means an attempt to build a machine that can perform any intellectual task that human beings could do. Although it is the goal of first generation of AI researchers and synonymous to "AI" in both future studies and science fictions, it is not the primary concern of most AI researchers nowadays, who are only interested in building machines which can perform a specific sort of task. Obviously AGI is more philosophically interesting than AI, given that the goal of AGI will inevitably involve high-level speculations about the nature of intelligence.

[③] Surely there are some exceptions in the literature as the follows. According to Anthony Beavors (cf. Anthony Beavors.'Phenomenology and artificial intelligence'. *Metaphilosophy* (2002) 33: 70-82), the Husserlian phenomenological reduction may lead to a re-description of cognitive processes, which is potentially valuable to AI/AGI. But he still owes readers a more detailed account of how to make phenomenological reduction relevant to *any given technical approach* in AI/AGI. The same criticism could be also applied to James Mensch (cf. James Mensch. 'Phenomenology and Artificial Intelligence: Husserl learns Chinese'. *Husserl Studies* (1991)8: 107-127), whose purpose of citing Husserl is mainly to refute John Searle's Chinese Room Argument, rather than to make positive proposals for practices in AI. That aside, there is still some research which does attempt to reconstruct Husserl's ideas with more technical details. For instance, Manuel Gustavo Issac provides a Husserlian phenomenological foundation of mathematical logic by carefully reconstructing Husserl's notion of "pure logic" and "semiotic intentionality" in his *Logic Investigations* (cf. Manuel Gustavo Issac. 'Towards a Phenomenological Epistemology of Mathematical Logic'. *Synthese* (2018) 195: 863-874.). Nonetheless, his study is not informative enough to build the desired bridge from Husserl to AI/AGI, given that (1) philological speaking, his characterization keeps the core notions of the Husserlian theory of intentionality, e.g., "noema", untouched, and (2) more generally speaking, his formalization of Husserl's ideas via mathematical logic *in an axiomatic manner* is still far remote from the AGI-oriented goal of building an artificial agent capable of flexibly responding to an open-ended environment.

[④] Mark Rowlands: *The New Science of the Mind,* Cambridge (MA): MIT Press, 2010, p. 3.

"GOFAI", as coined by John Haugeland [①]).

2. Husserl's notion of "noema", according to Hubert Dreyfus, [②] is a philosophical equivalent of AI scientist Marvin Minsky's notion of "frame" [③](since both include a pre-fixed data-structure for symbolically representing a stereotyped situation) and hence belonging to the tradition of GOFAI.
3. Therefore, Husserl's legacy concerning the nature of intentionality is not illuminating enough for a naturalized phenomenologist today.

However, besides the controversy involved in the first premise that we will address in section 3, at least the second premise of this argument is doubtable, since there is a relatively new tendency of interpreting the Husserlian notion of "Noema" not in terms of Minskian frames or Fregean "senses" but by virtue of Robert Brandom's inferentialism, and it is this reading that attributes more dynamic features to Husserl's theory of intentionality (more on this in section 5). Therefore, mainstream naturalized phenomenologists' marginalization of Husserl (which is in sharp contrast with their preference of Heidegger and Merleau-Ponty) is not warranted.

But the preceding claim itself does not imply that the relevance of Husserl to AGI is self-evident. The revelation of this relevance requires some further arguments, which are supposed to be provided in this article. To be more specific, these arguments are supposed to be supporting the following sub-claims, which constitute the route-map of this research:

1. Intentionality is required by any intelligent system, no matter whether it is artificial or natural, given that intelligence requires intentionality-presupposing capacities of revising beliefs in accordance with environmental changes.
2. The mainstream externalist treatment of *mental contents* (which is one component of intentionality) is to appeal to the correlation between them and external factors, but this approach is not beneficial to the modelling of intentionality in the sense that to directly model external factors is not feasible for any AI/AGI system.
3. The mainstream externalist treatment of *psychological modes* (which is another component of intentionality) is to metaphorically view them as "boxes" which apply different algorithms on contents emplaced in themselves, but this treatment is not beneficial to the modelling of intentionality either in the sense that it has assumed the discreteness among different modes and hence goes against the intuition that there are gradual transitions from this mode to another.
4. Hence, the needed theory of intentionality has to view mental contents as something which could be technically detached from external reality on the one hand, and view psychological modes as something permitting gradual mutual transformations among them on the another. The two requirements will naturally lead us to the Husserlian notions of "phenomenological *epoché*" and

---

[①] John Haugeland. *Artificial Intelligence: the Very Idea*, Montgomery: Bradford Books, 1985, pp. 112.

[②] Hubert Dreyfus: *What computers still can't do*. Cambridge, MA: The MIT Press, pp. 34-35.

[③] Marvin Minsky: A Framework for Representing Knowledge, in J. Haugeland, Ed., *Mind Design*, Cambridge (MA):MIT Press, 1981, pp. 95-128.

"noema", both of which are expected to be algorithmically reconstructed.

The main purpose of doing this research is not only to persuade naturalism-oriented AI/AGI researchers to acknowledge the values of Husserl's phenomenology, but also to reconstruct Husserl's phenomenology from a new perspective, namely, a perspective different from mainstream naturalized phenomenology by keeping distance from 4E-ism. And explorations in this direction will hopefully save Husserl's reputation out of the shadows of Heidegger and Merleau-Ponty, who have long been favored by mainstream naturalized phenomenologists.

## 2. **Intentionality is required by intelligence**

Here we will elude the complicated problem on how to strictly define the term "intelligence" [①] and begin with a simpler question: given that no reasoning system can get its conclusion which are practically useful without premises encoding empirical contents, and that *prejudices* are usually (albeit perhaps not inevitably) involved in these premises, what kind of reasoning machine we need to build if it is supposed to be bearing the mark of "intelligence"? Prima facie we have four options on the table:

**Option 1**: To build a system which reasons with premises which are all *true* and is capable of revising its beliefs in accordance with new environmental changes.

**Option 2**: To build a system which reasons with premises involving *prejudices* and is capable of revising its beliefs in accordance with new environmental changes.

**Option 3**: To build a system which reasons with premises involving *prejudices* and is *not* capable of revising its beliefs in accordance with new environmental changes.

**Option 4**: To build a system which reasons with premises which are all *true* and is *not* capable of revising its beliefs in accordance with new environmental changes.

Option 1 is quite weird in the sense that it looks unnecessary for a system to revise its belief if its starting premises are all true. Surely the set of all true premises of a system could be fairly small so that it is still necessary for such a system to enlarge the scope of its true beliefs in order to be more adaptive to the environment. But to include more new true beliefs does not mean that those older ones have to be *revised*, unless they can be proven to be untrue. Thus, option 1 still remains weird. Option 3 is weird too, since it is not so practically useful to build a machine which can only transfer falsities from premises to conclusions rather a machine which can automatically recognize falsities and separate them from truth. As to option 4, it is theoretically a bit more acceptable than 1&2, since a system with no false starting

[①] A systematic survey of this topic will involve considerations from the perspectives of AI, human psychology and even animal psychology. A recent attempt of doing this research is provided by .José Hernández-Orallo. Cf. .José Hernández-Orallo: *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*, Cambridge: Cambridge University Press, 2017.

premises would theoretically require *no* revisions of its beliefs. But it is still practically too challenging to build such a system, given that no programmer, who can be any one but an omniscient being, can guarantee that all premises that she feeds into the system will not be proven to be untrue in the future, unless the premises in question encode only trivial truth and hence are not potentially relating to any interesting implications. Hence, only one option, namely, option no. 2, is left on the table. That is to say, any intelligent system, whether artificial or natural, has to be able to revise its initial beliefs, some of which will be proven to be untrue.

And it is this option that makes the modelling of intentionality an indispensable part of the modeling of any artificial agent, if it is supposed to be minimally intelligent. Here goes the argument for saying so:

1. The design of an artificial intelligent system has to assume that it is capable of revising its stored beliefs (as what option 2 tells).
2. The revisions of beliefs have to be accompanied with changes of psychological modes, for instance, changes from a mental state of believing *p* to that of suspecting *p* and further to that of disbelievin*g p*, etc.(as what a folk psychologist would predict).
3. Intentionality is usually construed as a mental capacity to make the agent directed at "something", no matter whether this "something" exists in the physical world (a commonsensical view of what intentionality is).
4. Hence, intentionality is composed of both the manners of directing the agent and the "something" to be directed. Or in another way around, it is composed of psychological modes and mental contents.
5. Although it is a bit hard to judge whether the existence of mental contents conceptually assumes the existence of corresponding psychological modes (like "belief" or "desire"), the contrary case should hold, that is to say, the existence of psychological modes have to be based on corresponding mental contents, which do constitute the core part of intentionality (given that higher-order mental properties have to be based on first-order properties, although not necessarily vice versa).
6. Hence, from (2)&(5), it can be inferred that the requirement of the variety of psychological modes will eventually lead to the modelling of full-fledged intentionality in artificial systems.
7. Therefore, from (1)&(6), it can be inferred that the requirement of intelligence will eventually lead to the modelling of full-fledged intentionality in artificial systems.

We believe that this argument, which is sound, can make any reasonable AGI scientist seriously consider the problem of modeling intentionality, no matter whether the term "intentionality" has to be construed in a Husserlian manner. However, some readers may still ask: if the modelling of intentionality is so urgent for the design of any intelligent system, why do most AI scientists seem to be dismissive of this issue?

The answer is fairly simple: they are mostly AI scientists rather than AGI scientists; or in another way around, most AI systems that they built are too specific to certain tasks to satisfy the general requirement of option 2. Actually, these systems are

merely intended to satisfy option 4, according to which premises fed into the system are at least *supposed* to be *all* true. An exemplary case to footnote this point is Edward Feigenbaum's expert system (which is fairly representative of GOFAI), namely, a system usually designed to emulate the decision-making processes of human experts in a certain domain of knowledge.[①] Such a system is routinely composed of a knowledge base, which represents empirical state of affairs which are supposed to be facts, as well as an inference engine, which applies the inference rules to given "facts" to yield new "facts". But such a system can work well only when the "facts" stored in its knowledge base do encode genuine facts and hence immune to further revisions, and this condition itself is hard to satisfy since the progress in any domain of human scientific inquiries will routinely force human experts to update what they did believe, whereas it is technically challenging to make an expert system to automatically update its knowledge base as what a human expert would do with less efforts. Surely an AI scientist may try to design an expert system which literally has the capacity of automatically acquiring genuine knowledge from a large body of information including falsities, but this move is tantamount to the adoption of option 2, which eventually leads such as designer to the modelling of intentionality, as the preceding 7-step argument predicts.

Advocates of connectionism may wonder why option 2 is also compelling for connectionists, given that artificial neural networks that connectionists appeal to are not directly encoding mental contents on the symbolic level and hence seemingly not relevant to any option from 1 to 4. But it is noteworthy that the the mapping relationships between the training data fed into a typical neural network and the ideal outputs of the whole network are still analogical to the "knowledge base" of an expert system in the sense that they still crystallize knowledge of how an human programmer determines what type of inputs have to be mapped onto what type of outputs. Thus, analogical to an expert system, an neural network still needs to revise these mapping relationships when an human programmer finds it practically necessary to do so. However, still analogical to a typical expert system which cannot automatically update its knowledge base, a typical neural network, once having been trained to be adaptive to a certain type of mapping relationship, is also hard to be adaptive to a new relationship as well.[②] Therefore, in order to be more intelligent, even a neural network needs to exhibit intentionality by taking option 2 seriously.

Now philosophers should do their job by providing a plausible theory of intentionality to guide the modelling of intentionality, given that the abstractness of the term "intentionality" itself can be only philosophically construed. However, not all philosophical theories of intentionality are suitable to guide AGI researches. Against many readers' intuition that analytic philosophy bears more affinities with AI/AGI than continental philosophy, we will immediately argue that the notion of

---

① A systematic introduction to expert systems can be found in the follow textbook: Joseph Giarratano and Gary Riley: *Expert Systems: Principles and Programming*, Boston: Thomson Course Technology, 4rd edition, 2004.

② The technical jargon for this problem is "overfitting", which means that the "regularity" the system finds cannot be applied to a wider scope of tasks, since it is already trained to be too adaptive to a specific sort of training data.

intentionality provided by mainstream analytic philosophy is less preferable to its counterpart in Husserl's phenomenology.

## 3. Mental contents cannot be treated externalistically in AGI/AI

As aforementioned, besides psychological modes, the core part of intentionality is mental content, and in this sense, the problem of intentionality-with-a-t is also correlated with that of intensionality-with-an-s and hence relevant to semantic considerations. For any reader sympathizing with the tradition from Brentano to Husserl, it looks natural to view the existence of mental content as "inexistence", namely, a mode of existence which has to be confined within one's *internal* mental life and hence not directly relevant to external reality. By contrast, the mainstream Anglophone treatment of mental content is of an externalistic flavor, especially after Hillary Putman's twin earth case[①] became the standard thought experiment for participating the debate between semantic internalism and externalism. However, to be formally involved in this four-decade-long debate is not on the agenda of this research; rather, what is more relevant to our basic concern is the discussion of which side of the debate looks more acceptable *from the perspective of AI/AGI*. And our conclusion is that internalism has to preferred since externalism cannot be compatible with any conceivable from of practice in AI/AGI. Here goes the argument:

1. The formal framework of semantic externalism is two-dimensional semantics, by which the *external* dimension of meaning has to be detached from its *internal* dimension. To be more specific, such semantics allows one to distinguish the the primary intension from the secondary one: The primary intension is the method by which the agent attempts to pick up her desired object in a cross-worldly manner and to which she has the epistemic access, whereas the secondary intension is the information imbedded in the external object which she actually picks out in a certain possible world by using certain linguistic tools but to which she may have no epistemic access.
2. Hence, two-dimensionalism has assumed that there is an omniscient being's perspective from which the secondary/external intension could be presented, e.g., a perspective allows one to refer to the chemical composition of water even when modern chemistry is entirely out of the mind of the agent in question.
3. From (1) & (2), it can be inferred that any attempt to model intentionality in accordance with externalism has to encode the secondary intension from an *omniscient* being's perspective.
4. However, it is not feasible for any AI system to present an omniscient being's perspective, given that the knowledge of AI is ultimately derived from human-beings, who are not omniscient beings.
5. Hence, from (4) & (3), it can be deduced that semantic externalism cannot provide a feasible framework for AI.

① Hillary Putnam. "The meaning of meaning", in his *Mind, Language and Reality. Philosophical Papers*(Vol. 2). Cambridge: Cambridge University Press, 1975, pp. 215–271.

6. Therefore, semantic *internalism* is more appealing than externalism for AI, given that internalism and externalism have exhausted the logical space for semantic constructions.

Some readers may doubt the acceptability of step 2 by denying the necessity of introducing an omniscient being's perspective for fixing the secondary intension. They may contend that a high-level ascriber who knows more than the agent in question may suffice for ascribing the secondary intension to the target representation. But the question is: to know how much more is more enough for such an ascriber? Advocates of two-dimensionalism simply *cannot* say that "the ascriber only needs to know that the chemical composition of water is $H_2O$" in the twin earth case, since it would be too *ad hoc* to explain why this ascriber is so lucky to acquire the right piece of knowledge, among others, for picking up the right sort of secondary intension just in this case. Given that luck will routinely undermine the reliability of ascribing the secondary intension, luck has to be precluded in such processes, and the best way to preclude it is to appeal to an idealized ascriber who delivers semantic knowledge steadily and reliably. Obviously only an omniscient being can perfectly satisfy this condition, whereas no artificial system can stimulate such a being.

Some readers may also doubt the acceptability of step 4. Although for GOFAI, as they may contend, it looks necessary to deliberately avoid introducing an omniscient being's perspective by constructing "micro-worlds"(namely, partial representations of worlds which could be processed by a certain configuration of computing machinery[①]), GOFAI is not the only game in the town. It seems that both connectionist and enactivist systems are irrelevant to the problem posed by step 4 by avoiding building such micro-worlds.

But we don't think so. Actually, even in a connectionist system, it still makes sense to view "neuronal activation space" as another form of micro-worlds, although elements of these worlds are points, regions, or trajectories rather than symbols in their GOFAI-counterparts. [②] Moreover, according to AI scientist Ian Goodfellow *et al*., in a deep learning system (which is an updated form of connectionism), increasing amounts of raw data equivalent to fragments of certain mirco-worlds do go hand in hand with the increasing complexity of the micro-world-building mechanisms.[③] Hence, just like GOFAI, even in connectionism, there is no place for an omniscient being who is not constrained by any micro-world-building mechanism either.

There is no such a being in any enactivist system as well. Enactivism is a trend of thought in both philosophy of cognitive science and AGI/AI which claims that cognition arises as the result of interplays between an acting organism and environmental factors. One of the philosophical doctrines of enactivism is formulated in terms of AI scientist Rodney Brooks' "physical grounding hypothesis", according

---

① This term itself is coined by Hubert Dreyfus. Cf. H. Dreyfus. "From Micro-Worlds to Knowledge Representation: AI at an Impasse". In J. Haugeland, Ed., *Mind Design*, Cambridge (MA):MIT Press, 1981: pp. 161-204.

② Cf. Paul Churchland. *Neurophilosophy at Work*. Cambridge: Cambridge University Press, 2007, pp. 43; 128.

③ Goodfellow, I. *et al*.. *Deep learning*. Cambridge, MA: The MIT Press, 2016, pp. 18-23.

to which to build a system that is intelligent is necessary to have its representations grounded in the physical world.[①] But our philosophical worry of this remark is: Is it really possible for any cognitive system to be connected to the "physical world" without the mediating role of a certain micro-world which is epistemologically assessable to the system in question? We don't think so, and we even believe that Brooks' own following comment cannot be conceptually precluding such an mediating micro-world:

> The key observation is that the world is its own best model. It is always exactly up to date. It always contains every detail there is to be known. The trick is to sense it appropriately and often enough…. To build a system based on the physical grounding hypothesis it is necessary to connect it to the world via a set of sensors and actuators. Typed input and output are no longer of interest. They are not physically grounded.[②]

This observation is self-defeating because the term "physical grounding" seems to indicate the identity between the external world which "contains every detail to be known" and the world perceived by "a set of sensors and actuators", whereas actually they cannot be the same. The pictures taken by a sensor, say, a digital camera simulating the operation of the compound eyes of a dragonfly, should be different from another sensor, say, another camera simulating the operation of the eyes of an owl, and different visional inputs of two sensors are themselves subject to different constructing rules of different micro-worlds, none of which is unbiased towards the physical reality. Hence, Brooks' "physical grounding hypothesis" at most implies that language-like representations of the external world are unnecessary, rather than that *any* somehow *biased* presentation of the external world is unnecessary. And this implication is definitely not powerful enough to introduce an omniscient (and hence entirely unbiased) point of view of the physical world.

Another representative enactivism-inspired AI research deserving mentioning is provide by Randall D. Beer, who attempts to build a framework in which an agent and its environment are modeled as two coupled dynamical systems whose mutual interaction is in general jointly responsible for the agent's behavior. [③] But the epistemological problem involved here is still salient: how could a programmer model the external environment of the agent in a perspective-free manner? Actually there is no way to do so, and Beer's own design of an artificial agent simulating insect-like walking is also based on the construction of a continuous-time recurrent neural network, which can perceive the external environment only in accordance with what its internal structure allows it to perceive. Hence, there is no omniscient view of reality involved even in Beer's enactivist model.

Here we simply have no space to make comments on all enactivism-inspired AI researches. But the philosophical problem that they face are basically the same. They

---

[①] Rodney Brooks: "Elephants don't Play Chess". *Robotics and Autonomous Systems* (1990). 6: 139–159.

[②] Ibid., p. 141.

[③] Randall D Beer. "A Dynamical Systems Perspective on Agent-Environment Interaction". *Artificial Intelligence* (1995). 72: 173–215.

all assume that there are "information" stored in the external environment and that either mental representations or perceptions of agents can be modeled as teleologically oriented to this "information". The more abstract form of this assumption is a teleological account of information processing of agents, which is proposed by mainstream Anglophone philosophers like Fred Dretske,[①] Ruth Millikan[②] and developed by Karen Neander. To be more specific, according to Neander, a representation *R* has the content *C* if the subject has the function of producing *R*-type representations to respond to *C*-type environmental factors. This definition is patently attempting to introduce an omniscient being's view in the sense that it allows to formulate "*C*-type environmental factors" not from the subject's perspective.[③] However, even though this teleosemantic account were philosophically plausibly, when algorithmically realized, it would still have to appeal to internalism, because a subject-independent encoding of "*C*-type environmental factors" will presuppose a further encoding perspective in which these factors have to be emplaced in accordance with a certain format, and thereby "internalized" on a deeper level. To be a bit more formally, although the computing language $L_e$ for representing environmental factors outside the agent may be *superficially* different from the language $L_r$ for presenting representations of the agent, $L_e$ has to be *substantially* expressive enough to make all $L_r$-expressions translatable into their $L_e$-equivalents in order to maintain the uniformity of the entire computing platform. The resulting matryoshka-like structure of this world will still assume an underlying internalizing perspective.

Therefore, semantic internalism has to be assumed for modeling intentionality.

## 4. Psychological modes, directions of fit, and the box-approach

Another critical component of typical intentionality is psychological modes. If intentionality can be construed as any mental state which is essentially or at least potentially directed at anything that can be mentally presented, then different psychological modes can be accordingly viewed as different pathways through which the agent in question can direct herself at her mental target. John Searle has a long but still incomplete list of these modes in his widely-cited research of Intentionality, including belief, fear, hope, desire, love, hate, aversion, liking, disliking, doubting, wondering, whether, joy, elation, depression, anxiety, pride, remorse, etc..[④] However, as the analysis in this section will immediately show, the treatments of psychological modes proposed by mainstream Anglophone philosophers, like John Searle and Jerry Fodor, are not satisfactory enough even due to pure philosophical reasons, needless to say AGI-based considerations.

We will start with Searle. His characterization of psychological modes is based on the notion of "direction of fit". To be more specific, according to Searle, modes

---

① F. Dretske . *Knowledge and the Flow of Information*. Cambridge, MA: The MIT Press, 1981.

② R.G. Millikan. *Language, Thought and Other Biological Objects*, Cambridge, Mass.: MIT Press, 1984.

③ Karen Neander. *A Mark of the Mental: in Defense of Informational Teleosemantics*, Cambridge, MA: The MIT Press, 2017, p. 151.

④ John Searle: *Intentionality: An Essay in the Philosophy of Min*d, Cambridge: Cambridge University Press, 1983, p. 4.

like belief has a "mind-to-world" direction of fit in the sense that "it is the responsibility of the belief, so to speak, to match the world, and where the match fails I repair the situation by changing the belief". By contrast, modes like desire has a "world-to-mind" direction of fit in the sense that it is the responsibility of the world to match the desire, and when the world fails to do so, "I cannot fix things up by saying it was a mistaken intention…Desires and intentions…cannot be true or false, but can be complied with, fulfilled, or carried out…"[①]

But we don't think that a theory of psychological modes based on "directions of fit" is untenable. Firstly, Searle's description of directions of fit cannot be always fitting our linguistic intuitions in ordinary discourses. For instance, it looks intuitively unacceptable to say that "the world has to take its responsibility" if one's desire cannot be fulfilled, *when the content of such desire is utterly ridiculous*, e.g., a desire that "I want to be landing on the sun." (Hereafter I will simply call this desire as the "sun-desire"). Obviously there is nothing wrong for the sun, which has no free choice, to be a huge sphere of hot plasma which makes any attempt to land on it unrealizable, and in this case, against Searle's suggestion, the speaker in question has to take the responsibility of having such a ridiculous sun-desire.

Secondly, it may be implausible to attribute responsibility to the world even when intentions involving non-ridiculous contents cannot be fulfilled. For instance, if Tom fails to fulfil his desire of having a cup of Japanese tea by doing X, the whole situation can be more naturally interpreted as a failure based on his *wrong* belief, say, that "I can have a cup of Japanese tea *by doing X*.", and this interpretation quickly transfers the target of responsibility-attribution back to the agent again. This pattern of analysis can be even applied to those evaluating attributes intended to replace truth-values in Searle's list, such as "fulfilment" or "being carried out", etc.

More generally, the failure of fulfilling a desire of content *p* can be analyzed as a compound state of three *internal* components: (1) the agent *recalls* that she did believe that *p* would happen if she could do *X*; (2) She *recalls* that she did *X*; (3) She *observes* that *p* does not happen. Surely responsibilities of not being able to make *p* to happen have to be attributed to the agent again if either of the two cases occurs: (1) the *belief* that doing X would cause *p* to happen is false; (2) The agent did not successfully complete the task of X whereas she still *believes* that she did. In both cases the world itself is still innocent.

The moral of our analysis of Searle's treatment of psychological modes is that the desire/belief distinction cannot be treated in terms of directions of fit, which assume that these modes are based on relationships between mental entities and external entities (otherwise it would make no sense for him to talk about the direction either of "mind-to-*world*" or "*world*-to-mind"). Moreover, even seemingly world-oriented actions like "carrying out X" can be also viewed as something based on (although perhaps not reducible to) internal states and hence still more relevant to agent's internal mental life. This perspective-based analysis of psychological modes is perfectly compatible with the internalist treatment of mental contents, which was proposed by the last section, whereas Searle's perspective-free view is conflicting

① Ibid. p. 8.

with it. Hence, if the conclusion of last section is sound, Searle's treatment of direction of fit cannot be acceptable.

Compared with Searle, Jerry Fodor's treatment of psychological modes, which is part of his Language of Thought Hypothesis (LOTH), is more internalism-oriented. According to LOTH, thinking is a processes in which mental representations are "tokened" by some lexicon-like mental entities with the aid of a combinatorial syntax which gives these items an appropriate structure. Since the rules guiding the operations of this syntax are determined by the internal features of the cognitive architecture rather than external environmental factors, on the LOT-level, Fodor is not so interested in "whether what the oracles write is *true*; whether, for example, they really are transducers faithfully mirroring the state of the environment, or merely the output end of a typewriter manipulated by a Cartesian demon bent on deceiving the machine". [①] Hence, LOT has a minimal internalist flavor compared with typical teleosemantic accounts of mental contents. And this feature is also inherited by Fodor's following account of psychological modes (or "propositional attitudes" in his terms):

> LOT says that propositional attitudes are relations between minds and mental representations that express the contents of the attitudes. The motto is something like: 'For Peter to believe that lead sinks is for him to have a Mentalese expression that means lead sinks in his ''belief box'''. Now, propositional-attitude types can have as many tokens as you like. I can think lead sinks today, and I can think that very thought again tomorrow. LOT requires that tokens of a Mentalese expression that mean lead sinks are in my belief box both times…. [②]

Psychological modes, in this narrative, are metaphorically viewed as "boxes", in each type of which a certain combination of tokens tokening certain mental content is emplaced to constitute full-fledged intentionality. In addition, each type of "box" instantiates a specific type of syntactic rules that contents emplaced in them have to follow. Although Fodor is not interested in characterizing the difference among different types of "boxes" (the only type of "box" other than the "belief box" that he mentions is "intention box"[③]), he has to assume that the demarcation line between this "box" and another can be explicitly drawn, otherwise it would make no sense to talk about "having a mental box *in* the belief box". Hereafter we will call this treatment of psychological modes as the "box-approach".

However, though Fodor's box-approach is not as externalism-evoking as Searle's narrative of "directions of fit", it is still problematic. Actually we have doubts on the applicability of this approach to the task of modelling natural intentionality, since this approach mistakenly assumes that it is always easy to find the demarcation line between this attitude and another. But this assumption is definitely not true in cases wherein the "strength" of a psychological mode is gradable. Here goes our analysis.

Obviously both the strength of beliefs and desires are gradable: It makes perfect

---

① Jerry Fodor: "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology, *The Behavioral and Brain Sciences* (1980), 3, p. 65.

② Jerry Fodor: *LOT2: The Language of Thought Revisited*, Oxford: Oxford University Press, 2008, p. 69.

③ Ibid. p. 39.

sense to say that I have a *strong* belief of *p* or a *weak* desire of *q*. But the semantic problem involved here is that the meanings of many psychological words, when supplemented with adverbial expressions indicating strength, are mutually overlapping or even synonymous to each other: As for an instance, is there really a substantial difference between "*A very weakly believes* that *p* is the case." and "*A very weakly suspects* that *p* is the case."? If there is no substantial difference between them, then the most natural explanation for the lack of this difference seems to be that the scope of the so-called "belief-box" is *continuous* to that of "suspect-box". But this explanation quickly makes Fodor's box-metaphor, which assumes the discreteness of boxes, fade.

Sympathizers of LOT would like to contend that linguistic intuitions on how we use psychological verbs in ordinary discourses may not be illuminating for how LOT works on a deeper level. For instance, it may be the case that the *gradable* "belief" in our natural language does not strictly correspond to the *ungradable* "belief-box" on the LOT-level. But we don't think this remedy can work. It is an undeniable fact that beliefs are gradable on the level of public language, and it is also widely accepted that speech acts cannot be produced without corresponding mental activities. Hence, speech acts involving gradable psychological words have to be accompanied with corresponding mental activities, which are expected to be explained by LOTH. But LOTH simply cannot plausibly explain the explanandum on the table, given that it is always fairly difficult for a theory assuming abrupt transitions from a basic state to another to explain phenomenon involving gradual inter-state transitions, unless the number of basic states on the level of explanans is as tremendous as an astronomical figure. But it is psychologically implausible to suppose that the number of types of psychological modes is so tremendous on the LOT level (otherwise the resulting human cognitive architecture would be too complicated to make itself a conceivable result of natural selection), hence, any competing explanation, whatever it is, has to abandon the box-approach. Analogically, in the modelling of artificial intentionality, the box-approach cannot be adopted if the system is expected to be able to exhibit gradual transitions among different mental states as humans would do.

Now sympathizers of either Searle or Fodor may still contend that neither philosopher is interested in *algorithmically* realizing artificial intentionality; rather, both philosophers have their independent arguments against the possibility of doing it, e.g., Searle's "Chinese Room Argument"[①] and Fodor's "Argument against High-level Modularity as a Requisite of Computational Theory of Cognition".[②] But we don't think this objection is relevant to our argument. Our point is: no matter whether their global hostility towards the algorithmic reconstruction of intentionality is warranted, their theory of natural intentionality is flawed, hence, any AI scientist who has adopted their general view about how intentionality works cannot model intentionality

[①] John Searle: "Minds, Brains, and Programs", *Behavioral and Brain Sciences* (1980), 3, 417-458.

[②] J. A. Fodor. *The Modularity of Mind,* Cambridge, MA: MIT Press, 1983. A concise reconstruction of this argument can be found in: Philip Robbins. "Modularity of Mind" , *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), Edward N. Zalta (ed.), URL=http://plato.stanford.edu/archivies/win2017/entried/modularity-mind/, § 2.2.

successfully.

Now we will give some further reasons to explain why Searle's and Fodor's theories are problemic in AGI. In the last section, we have explained why the perspective-free view of "world" assumed by enactivism-oriented AI cannot be coherently modelled. Since the similar view has been assumed in Searle's notion of "direction of it", this notion itself cannot be algorithmically modeled as well. As to Fodor's box-approach, actually a variant of it has been adopted by mainstream AI scientists in a branch of AI which is labeled as "context modelling". The aim of context modelling is to build a computer system which can automatically handle data differently according to different contexts, and the whole goal here is relevant to the issue on intentionality in the sense that each type of psychological state can be more abstractly viewed as a type of context (e.g., the belief-context, the desire-context, etc.). Hence, if the box-approach in a theory of context is flawed, the similar approach in the modelling of contexts, namely, an approach according to which each context is treated as a "box", cannot bring about satisfactory results as well.

And following examples may show that even the box-approach in context modelling is defective, and this observation would conversely reinforce our current doubt of the validity of the similar approach in a theory of intentionality. A typical AI-oriented (but still philosophical) formulation of the box-approach in context modelling is given by Fausto Giunchiglia and Paolo Bouquet (hereafter G&B):

> It is quite common intuition that some sentences are true (polite, effective, appropriate, etc.) in a context, and false (impolite, not effective, inappropriate) in others, that some conclusions hold only in some contexts, that a behavior is good only in some contexts, and soon. For instance, "France is hexagonal" (or "Italy is boot-shaped") is true in contexts whose standard of precision is very low, false in the context of Euclidean geometry. …All these examples seem to suggest that a context can metaphorically be thought of as a sort of "box". Each box has its own laws and draws a sort of boundary between what is in and what is out. A closer look to the literature on context will show that this metaphor can be given two very different interpretations. According to the first, a "box" is viewed as part of the structure of the world; according to the second, a "box" is viewed as part of the structure of an individual's representation of the world. [①]

It is not hard to see that G&B's expressions like "each box has its own laws and draws a sort of boundary between what is in and what is out" predicts that inter-box transition has to be abrupt. Since the second first type of "box" in G&B's narrative obviously refers to psychological modes, inter-mode transition cannot be gradual in G&B's framework as well.

However, trans-contextual reasoning has to be done in many practical cases, and AI scientist should do something to meet this practical demand. Their recipe is to provide some ad hoc bridge-like formula to bring information stored in one box to get to another, such as G&B's "bridge laws" and Ramanathan V. Guha & John

---

① Fausto Giunchiglia & Paolo Bouquet. "Introduction to Contextual Reasoning: an Artificial Intelligence Perspective", in *Perspectives on Cognitive Science* (edited by B. Kokinov, New Bulgarian University, 1997), p. 139.

McCarthy's "lifting formula"[①]. But none of the proposal here is flexible enough to meet the demands of AGI, since these trans-contextual reasoning devices cannot be built without previously individuating all boxes and confining all inter-box boundaries, whereas in ordinary discourses, even if it makes sense to talk about the boundary between this topic and another, the boundary itself is routinely pragmatically determined. Hence, the box-approach is only useful in building specific AI systems which are not expected to exhibit human-level flexibility.

The general moral of this section and the last one is that mainstream philosophical theories of intentionality is not illuminating for AGI because they either appeal to external environmental factors which cannot be internally modeled, or they cannot handle gradual transitions among different cognitive states. Now it is the right time to introduce Husserl to solve these problems.

## 5. How could a Husserlian AGI scientist solve the problems?

First of all, we will show how Husserl could explain intentionality without introducing external factors by reinterpreting his notion of "phenomenological *epoché*" or "phenomenological reduction". The core text relevant to this notion is as the follows:

> The theory of categories must start entirely from this most radical of all ontological distinctions — being as consciousness and being as something which becomes "manifested" in consciousness, "transcendent" being — which, as we see, can be attained in its purity and appreciated only by the method of the phenomenological reduction. In the essential relationship between transcendental and transcendent being are rooted all the relationships already touched on by us repeatedly but later to be explored more profoundly, between phenomenology and all other sciences - relationships in the sense of which it is implicit that the dominion of phenomenology includes in a certain remarkable manner all the other sciences. *The excluding has at the same time the characteristic of a revaluing change in sign; and with this change the revalued affair finds a place once again in the phenomenological sphere*. Figuratively speaking, that which is parenthesized is not erased from the phenomenological blackboard but only parenthesized, and thereby provided with an index. [②]

Now we attempt to reinterpret Husserl's meaning by formulating the following procedures of "*epoché*", in which no puzzling terms like "transcendent being" or "purity" will be used:

Step. 1. To introduce the commonsensical view that the truth-conditions of *p* is different from SMs(*p*) (wherein "S" refers to a subject, "M" refers

[①] Ramanathan V. Guha, & John McCarthy. "Varities of Contexts", in *Modeling and Using Contexts* (edited by Patrick Blackburn et al, Springer-Verlage, Berlin, 2003), pp. 164–177.

[②] E. Husserl. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy. First Book*: *General Introduction to a Pure Phenomenology*, translated by F. Klein. Hague: Martinus Nijhoff, 1980, p. 171.

to a certain type of psychological mode). For instance, even the truth-conditions for the sentence that "Tully is Cicero" are all satisfied, this does not imply that the truth-conditions of "Sally believe that Tully is Cicero." can be satisfied accordingly.

Step. 2. It is obvious that the truth-conditions of "Sally believes that Tully is Cicero." can be only *internally* determined in terms of, say, whether the target belief is coherent with other stored beliefs or whether the belief is sufficiently supported by evidence acquired by the agent. Otherwise it will be too hard to explain why truth conditions of SMs($p$) is so irrelevant to truth conditions of $p$.

Step. 3. Now we take a further step by presupposing that there is an implicit speaker accompanying any conceivable sentence. Hence, each proposition is supplemented with a psychological mode.

Step. 4. Hence, by executing step 2&3, the truth conditions of each conceivable sentence can be only internally determined. This is nothing but the "residues" of phenomenological reduction.

Some readers may wonder how one could be entitled to presuppose the omnipresence of an implicit speaker (as step. 3 requires) without introducing subjective idealism. But an AI/AGI-related point of view can easily explain how. Obviously, no AI/AGI system can be built without a certain programing language, and the organization of each programing language has to encapsulate how the world works from the perspective of a specific designer. Therefore, nothing mysterious will be involved in presupposing such an "implicit speaker" if the preceding procedures are construed in an AI/AGI context. And this interpretation can even make Husserl's notion of "*epoché*" perfectly compatible with metaphysical physicalism (which is the metaphysical assumption of most AI scientists), since the irreducibility of an "implicit speaker" in any algorithmically reconstructed micro-world implies neither that the physical world itself does not exist independently of how the cognitive systems perceive them, nor that the cognitive activities are not supervenient on corresponding physical events. Or in Husserl's own terms in the preceding citation, speculations about the metaphysical nature of the world are "not erased from the phenomenological blackboard but only parenthesized". Hence, a Husserlian AI programmer does not need to take the burden of modelling the world beyond the horizon of an omnipresent "implicit speaker".

As to how to construe gradual inter-mental-state transitions, Husserl's phenomenological theory of time, if reconstructed, can be used to form the following argument against Fodor's box-approach:

1. If the box-metaphor is applicable to intentionality, then it has to be applicable to any component of intentionality, just as if one can separate A from B, then she should be able to separate any component of A from B.
2. Consciousness of temporal sequences has to be involved in many psychological modes like *hope* and *regret* (commonsense).
3. Phenomenologically speaking, a typical internal temporal sequence is composed of *original impression* (namely, the phenomenological equivalent

of "present")", *protention*（namely, the phenomenological equivalent of "future"）and *retention* (namely, the phenomenological equivalent of "past"). [①]

4. But it makes no sense to talk about the abrupt transitions among these components, given that they constitute a continuum in which the "present" can be only seen as an ideal limit, "just as the continuum of species red converges towards an ideal pure red". [②]
5. Hence, the box-metaphor cannot be applied to temporal components of intentionality.
6. It is obvious that (5) is incompatible with (1).
7. Therefore, the box-metaphor is inapplicable to intentionality.

So far, so good. Husserl's theory of intentionality is immune both to externalism and the box-approach. However, some readers may still complain that his theory has little value for AGI in the sense that it provides no algorithmic details. But we believe that it has at least provided some general guidelines on how intentionality could be modeled. And these guidelines can be found in his notion of "noema".

But what is noema? Unfortunately, even within Husserl scholarship there is a debate over different interpretations of noema. For example, according to the Fregean interpretation (supported by Føllesdal, Dreyfus and McIntyre, etc.[③]), noema is a meaning-encoding entity between mental act and the external object, and the relevant object becomes the referent of the relevant mental act just because noema specifies the way in which the referent is referred. By contrast, a competing interpretation of noema (supported by Sokolowski, Drummond, etc.[④]) contends that noemata are not mediating entities between mental acts and external objects but just the external objects considered in the phenomenological reflection, or "experienced objects" for short.

The first interpretation of noema looks less promising from the perspective of AGI, because it assumes a huge programing burden of modeling the sandwich-like structure of "act-noema-object", and despite the formidable work of specifying each noematic meaning as a *contextually invariant* manner of fixing referents, how to harmonize these meanings with *contextually emerging* factors would be another tricky problem. By contrast, since no *contextually invariant* entities have been assumed in the second interpretation of noema, it may afford a more elegant way to model intentionality.

However, even the the second interpretation is problematic by including the key phrase "experienced objects". Given that the specific perspective involved in any

---

① Edmund Husserl: *On the Phenomenology of the Consciousness of Internal Time*, translated by John Barnett Brouch, Kluwer Academic Publisher, Dordrecht, 1991, p. 40.

② Ibid. p. 41.

③ Cf. D. Føllesdal. "Husserl's notion of Noema". *Journal of Philosophy* (1969), 66: 680–687; H. Dreyfus. "Husserl's Perceptual Noema." In H.L. Dreyfus, and H. Hall, eds., *Husserl, Intentionality and Cognitive Science*. Cambridge, MA: MIT Press, 1982, 97-123; R. McIntyre. 'Intending and Referring', In H.L. Dreyfus and H. Hall (eds.), *Husserl, Intentionality and Cognitive Science*. Cambridge, MA: MIT Press, 1982, pp. 215-231.

④ Cf. R. Sokolowski. *Introduction to Phenomenology*. Cambridge: Cambridge University Press, 2000; J. Drummond. *Husserlian Intentionality and Non-Foundational Realism*. Dordrecht: Kluwer, 1990.

piece of experience is by nature in contrast with the object itself which is perspective-free, this gap cannot be easily filled by appealing to a compound expression like "experienced objects", which can be only unpacked as a weird phrase like "perspective-free entities from the lens of a specific perspective"(but how could any entity keep on being perspective-free when viewed from a certain point of view?). Hence, the burden of modelling perspective-free external entities is still left on the table if this compound expression is literally put into practice.

A way out of this embarrassment is to appeal to an internalized version of Robert Brandom's inferentialism, which is applied to the interpretation of nemoa by Steven Crowell. Inspired by Brandom's discussion of "material inferences",[①] Crowell defines Husserl's notion of noema in terms of "a quasi-inferential concept of representation", which is footnoted by the following illustrations: The perceived color is an "adumbration of something"; the front side "implies" the unseen back; taking it as a barn "entails" a specific relation to the landscape, the barnyard, and farming practices, etc..[②] Hence, the noema in this sense can be viewed as a gateway from aspects of objects that have been experienced to those expected to be experienced in the future. According to this interpretation, noema is definitely not any static entity but some high-level features of the object-relevant inferences that the subject is engaged in.

This reading of noema fits with Husserl's following comment on the nature of phenomenological "object", which is synonymous to "noematic X" in his context:

> Everywhere 'object' is the name for eidetic concatenations of consciousness; it appears first as the noematic X, as the subject of sense pertaining to a different essential types of sense and posita. Moreover, it appears as the name, 'actual object', and is then the name for certain eidetically considered rational concatenations in which the sense-conforming, unitary X inherent in them receives its rational position."[③]

Or in another way, the term "object" is nothing but a system of harmoniously connected experiences, and the object *per se* is merely the external correlate of the "object" internally construed.

But how to *algorithmically* model Husserl's notion of noematic X as an inferential node? First of all, the harmoniousness of the whole inferential network about noematic X might be tested in terms of, for instance, the compatibility of the beliefs encoded in a corresponding network (e.g., Touretzky's 'Inheritance System' [④]and Franz Baader et al. eds.'s 'Description Logic'[⑤]). However, the remaining technical

---

① R. Brandom. *Articulating Reasons*. Cambridge, MA: Harvard University Press, 2000, pp. 52-57.

② S. Crowell. "Phenomenological Immanence, Normativity, and Semantic Externalism". *Synthese* (2008)160: 344.

③ E. Husserl. *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy. First Book*: *General Introduction to a Pure Phenomenology*, translated by F. Klein. Hague: Martinus Nijhoff, 1980, p. 347.

④ D. Touretzky. The Mathematics of Inheritance Systems. Los Altos: Morgan Kaufmann, 1986.

⑤ F. Baader *et al.* (Ed.). *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge: Cambridge University Press, 2007.

obstacle is still salient. To recall, the noematic X as an inferential node is connected to unexperienced aspects of objects, hence, a computational model of it has to be open to unexpected data in a flexible but not very resource-consuming manner. This requirement will pose a big challenge to GOFAI approaches (including Touretzky's and Baader's approaches), given that the axiomatic nature of these approaches in principle renders the system unresponsive to environmental contingencies. The similar challenge applies to connectionism or deep learning too, since when a connectionist/deep-learning system is trained to be adaptive to a certain type of task, say, the recognition of human facial expressions, it cannot be adaptive to a new task of another sort, say, speech recognition, which may require a new set of training data and even a different neural architecture with different parameters. By contrast, a human agent can flexibly combine the cognitive capacity of recognizing human face and that of recognizing human voice for completing the task of, say, recognizing somebody as somebody.

A possible technical solution to this problem is provided by Pei Wang's Non-Axiomatic Reasoning System (NARS, with "Narsese" as its adjective form, which literally means "of the language of NARS" ). [①] Due to the limitation of space, we can only explain how NARS helps to model Crowell's inferentialist interpretation of noema. In NARS, both lexicon-like entities and minimally stable patterns of experiences can be viewed as "Narsese concepts", which are connected to each to form Narsese sentences and hence more complicated inferential pathways. The whole Narsese conceptual map is *not* axiomatically pre-determined by the programmer but automatically and gradually coming into being as the result of the interplays between some internal parameters of the system and inputs fed into it. And the whole system is described as "non-axiomatic" just in this sense.

More importantly, in NARS, psychological modes are characterized without appealing to the box-approach. Rather, belief, the most primitive psychological mode, is firstly implicitly expressed in terms of the strength or weight of pathways connecting one Narsese node and another. For instance, if a pathway connecting the node S with that of P is highly weighted, it means that the system strongly believes that all Ss are normally Ps. As to the weight-values of pathways, they come from the interactions between acquired evidence and the corresponding Narsese sentence (by the way, each piece of evidence is regarded as a Narsese term in NARS). That is to say, the more evidence for a Narsese belief is at hand, the more firmly the system has the belief. [②] This evidence-based treatment can easily handle psychological modes like suspect and disbelief (both of which involve the role of positive/negative

[①] Cf. P. Wang. *Rigid Flexibility: the Logic of Intelligence*. Dordrecht: Springer, 2006; P. Wang, P.. *Non-axiomatic Logic: A Model of Intelligent Reasoning*. Singapore: World Scientific, 2013.

[②] More technically speaking, in NARS, two parameters are introduced to represent this weight-value *quantitatively*: the "frequency" (*f*) value and the "confidence" (*c*) value. The formulae for computing them are as follows: (1) $f = w^{+}/w$ ; (2) $c = w/(w + k)$ (wherein "$w^{+}$" means the quantity of all positive evidence of the target statement, "*w*" means the quantity of all relevant evidence of the target statement, and "*k*" means a constant value prescribed to a certain system). Since weight-values of the correlations between Narsese terms do mirror the strength of beliefs in natural languages, such treatment will naturally render Narsese beliefs gradable and thereby make the entire box-approach dispensable in NARS.

evidence) as mutually transformable states.

Psychological modes like intention or desire do make things a bit more complicated, since they involve the notion of "goal", which is future-oriented, whereas any evidence is past-oriented. But there is still no need to introduce Searle's notion of "direction of fit" here, since the future/past contrast is one thing, while the world-to-mind/mind-to-world contrast is another. Rather, the Narsese recipe for handling desire can be unpacked as the following steps:

Step. 1. Firstly, we assume that through a certain procedure of learning, the system has acquired a pool of Narsese sentences about how the artificial system itself can functionally survive, e.g., the knowledge about how to maintain the battery level.

Step. 2. The system applies general knowledge in the preceding pool to the current state of itself to find whether it is "healthy" enough. If it is, then no desire will be produced; if not, go to execute the next step.

Step. 3. Due to the inferential capacity of the system, it finds out that if a precondition *p* were true, it could "live" much better.

Step. 4. But the system finds that it cannot believe that *p* is true now since it lacks enough positive evidence.

Step. 5. Then the system would like to attach the label of "primitive goal" to *p* and calculate how much evidence is needed to make it true.

Step. 6. Since the needed evidence is not actually presented, the system would attach the label of "derived goal" to each operation that would make a certain piece of relevant evidence occur.

Step. 7. The forgoing reasoning will drive the system into proper actions.

Step. 8. The system will evaluate the gap between the newly acquired evidence and the *p*-requiring evidence after each run of actions, until the gap is reduced to a certain level, which means that the desire is satisfied.

The preceding procedures characterize how a "selfish" AGI system could entertain and derive desires with the ultimate goal of its own survival. Surely one can build an "altruist" system by replacing the pool of knowledge in step 1 with another pool concerning how *other* systems or human masters could functionally survive. Moreover, one can even build a system both bearing the mark of "selfishness" and "unselfishness" by "teaching" the system to form both types of pools, and thereby represent the so-called "complexity of human nature" in an artificial system. However, no matter how complicated systems could be built on the basis of the preceding 8-step recipe, desire or intention will never be treated as a static box waiting to be filled with neutral mental content. Rather, in NARS, "intention" or "desire" refers to a high-level feature of dynamic inference overarching both action and cognition. In addition, in NARS, the notion of desire, albeit not directly evidence-based, is still relevant to evidence, since conversions from expected evidence to evidence-making actions do assume that the system's sub-system of beliefs is evidence-based. Hence, even though the label of "primitive goal" itself looks like a box-label, it is not literally an intention-box which is in contrast with a belief-box, since belief-supporting evidence have to be used in the process of

forming intentions of desires.

As to how these Narsese constructions are relevant to Husserl's notion of "*epoché*" or "bracketing", we just want to make one point explicit now. Although this notion can be applied to any AI system in some degree due to the fact that any AI system has some built-in prejudices about how the world works, there is no mainstream AI system deserving to be attributed with the label of "an distinct *individual*", since different computers implementing the same software would behave basically in the same way and hence "bracket" the external world from basically the same perspective. By contrast, human perspectives are definitely more diversified and hence capable of producing intentionality in a way specific to the historically formed habits of individuals, or to take words from Dermot Moran's interpretation of Husserlian egos, "different egos have their different streams of temporalization, and it is a complex issue how a 'common form of time' is constituted".[①] In this aspect, NARS is superior to most mainstream AI systems, provide that for each individual computer implementing NARS, the topology of its Narsese conceptual map is nothing but the result of its own learning history, and habits of inferences could be thereby made distinct from this individual computer implementing NARS to another. Hence, it is fairly natural for two NARS computers to bracket the same content in different ways, or even "have their different streams of temporalization" in Husserl's sense. In an upshot: NARS makes a relatively promising approach to the desired Husserlian AGI project.

## 6. Metaphilosophical observations as concluding remarks

Hitherto we have explained: (1) why the notion of intentionality is indispensable for any AGI system; (2) why the treatment of intentionality by mainstream Anglophone philosophy is less preferable to its Husserlian counterpart; (3) How to model the Husserlian notion of intentionality by appealing to NARS. Now it is the time to articulate our underlying motivation propelling the whole research. We concede that more than half of the space budget of this article is consumed to clarify point (1), and this way of distributing space is necessary since externalism-oriented (and hence anti-Husserlian) speculations are so dominant in current Anglophone philosophy of mind that Husserl's own approach cannot find its niche without a serious battle with them. However, we are still loyal to the tradition of "analytic philosophy" in a very general sense, if this label is only understood as a general name of any manner of thinking and writing philosophy by using explicit arguments. And Husserl's philosophy is not "analytic" enough even according to this loose definition of the term, given the overpopulation of his terminology and the difficulties of directly reconstructing his wordy comments as lineal arguments. And due to this consideration, this article is also intended to "disenchant" Husserl by appealing to resources in AGI.

---

[①] D. Moran.. *Edmund Husserl: Founder of Phenomenology*. Cambridge, UK: Polity, 2005, p. 218.

But why AGI? Why not only formal tools from logic or statistics, given that all AI systems have to rely on them? The primary reason is that a workable AGI system has to be something more than these formal tools. For instance, it has to have a proper cognitive architecture and hence to be minimally relevant to human intentionality, whereas formal tools do not need to be so. Meanwhile, due to its reliance on algorithmic details, any AGI narrative, albeit perhaps on a high level, still has to be "analytic" in the most general sense of the term. Hence, due to this duality, AGI could provide a perfect platform to interpret Husserl.

Another reason not to appeal to formal logic is that by "formal logic", most people just mean the Fregean logic, which is actually more suitable for characterizing semantic externalism, since the ontological status of *external* referent (e.g., objects or truth-values) has to be assumed in the Fregean theory of meaning, otherwise it would make no sense for a Fregean to view meanings as mapping mechanisms correlating symbols with referents. In this sense, the Fregean logic should be a very cumbersome tool for modelling Crowell's inferentialist interpretation of noema, from which naïve externalism has to be precluded. By contrast, if we appeal to AGI rather than "logic", then the novelty of the term "AGI" itself will give us more space to introduce some form of non-Fregean logic, e.g., the Narsese logic. And this treatment will naturally separate Husserl's own position from Føllesdal's and Dreyfus' Fregean interpretation of Husserl, in which the Fregean view of logic is still assumed.

# Enhancement, Uploading, and Personal Identity

Sangkyu Shin (Ewha Womans University)

# Content

- Enhancement and Upload
- The problem of Personal Identity
- Fission Problem and Schneider’s response
- Derek Parfit: Unimportance of Identity
- Embodied mind and It’s Implications for Uploading

# Transhumanim and Body

- **Transhumanism** (abbreviated as **H+** or **h+**) is defined as a philosophical or cultural movement that advocates for the transformation of the human condition by developing and making widely available sophisticated technologies to greatly enhance human intellect and physiology.
- Cybernetic Idealism ↔ Embodied Cognition
  - ✓Kurzweil seeks immortality through mind upload. For him, **the body** is basically an obstacle to overcome.
  - ✓For transhumanists, it seems that the end point of human-machine merging is uploading.

# Mind Uploading

- Bostrom, 2003. “Transhumanist FAQ: A General Introduction,” version 2.1,
- Uploading (sometimes called “downloading,” “mind uploading” or “brain reconstruction”) is the process of transferring an intellect from a biological brain to a computer. One way of doing this might be by first scanning the synaptic structure of a particular brain and then implementing the same computations in an electronic medium….
- Advantages of being an upload would include: Uploads would not be subject to biological senescence. Backup copies of uploads could be created regularly so that you could be rebooted if something bad happened. (Thus your lifespan would potentially be as long as the universe’s.) … Radical cognitive enhancements would likely be easier to implement in an upload than in an organic brain….
- A widely accepted position is that you survive so long as certain information patterns are conserved, such as your memories, values, attitudes, and emotional dispositions….

# Transhumanism and Patternism

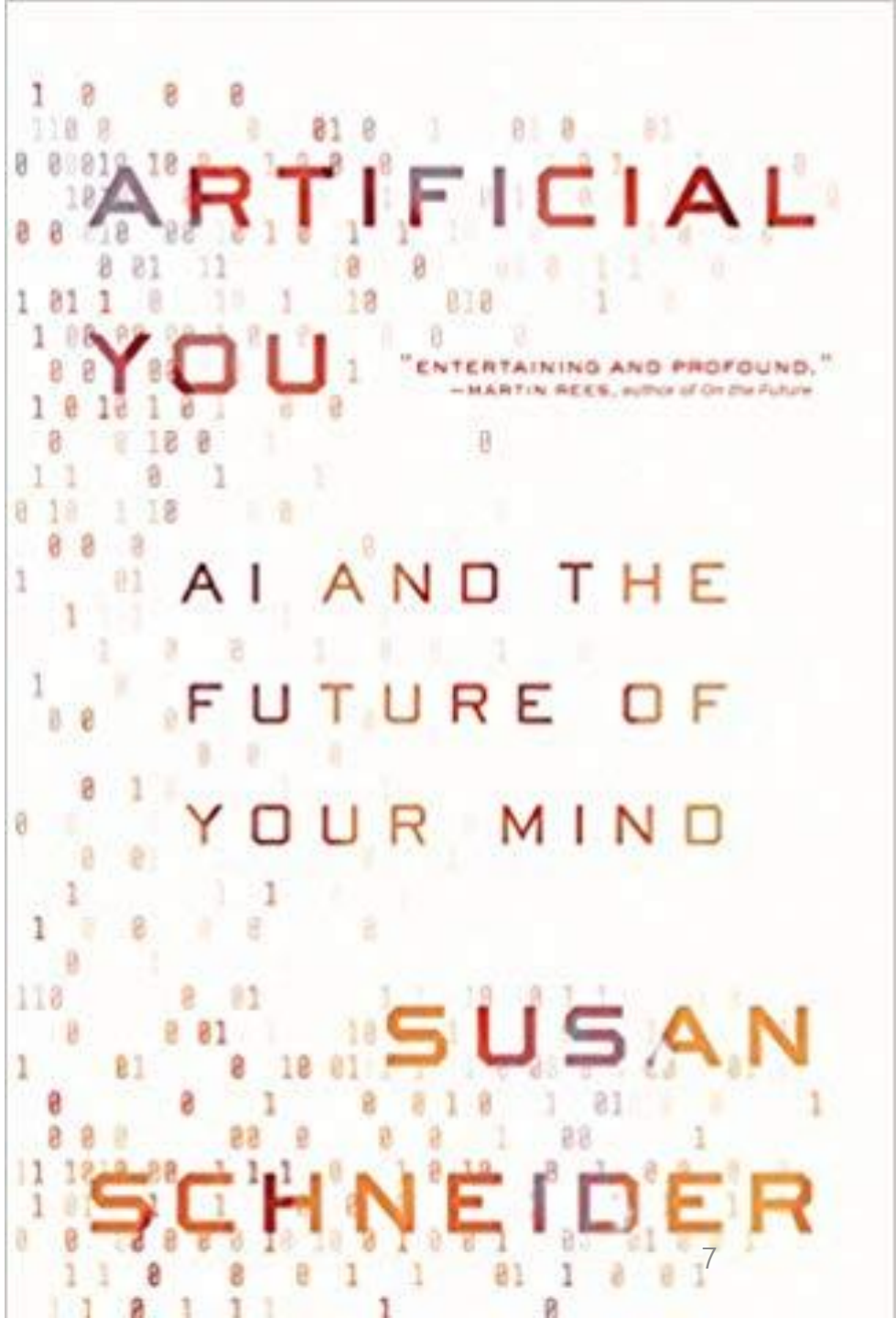
- Patternism: an approach to the nature of persons that is an intriguing blend of the computational approach to the mind and the traditional psychological continuity view of personhood.
- Kurzweil (*Singularity Is Near*, 2005, p. 383): "We know that most of our cells are turned over in a matter of weeks, and even our neurons, which persist as distinct cells for a relatively long time, nonetheless change all of their constituent molecules within a month.... I am rather like the pattern that water makes in a stream as it rushes past the rocks in its path. The actual molecules of water change every millisecond, but the pattern persists for hours or even years."
- Stephen Hawking (The Guardian, 2013): "I think the brain is like a programme … so it's theoretically possible to copy the brain onto a computer and so provide a form of life after death."

# Software Approach to the Mind (SAM): the mind is a software program

- The mind is the program running on the hardware of the brain. That is, the mind is the algorithm the brain implements.
- Mind and mental states/processes are multiply realizable. Thus, our mind can be implemented(realized) in hardware other than the brain, including computers.
- According to this view, it seems possible that you enhance your brain hardware in radical ways and still run the same program, so your mind still exists.
- The survival of a person is a matter of the survival of a software pattern.

# Artificial You

- Susan Schneider
- “After too many changes, the person who remains may not even be you. Each human who enhances may, unbeknownst to them, end their life in the process.” (p. 7)
- If we do not maintain our personal identity as a result of “enhancement”, can we call this an enhancement?
- Is uploading or radical enhancement of the human brain(or mind) compatible with our survival?
- Radical enhancement should not change or remove essential attributes related with the persistence of a person.



# Problem of Personal Identity

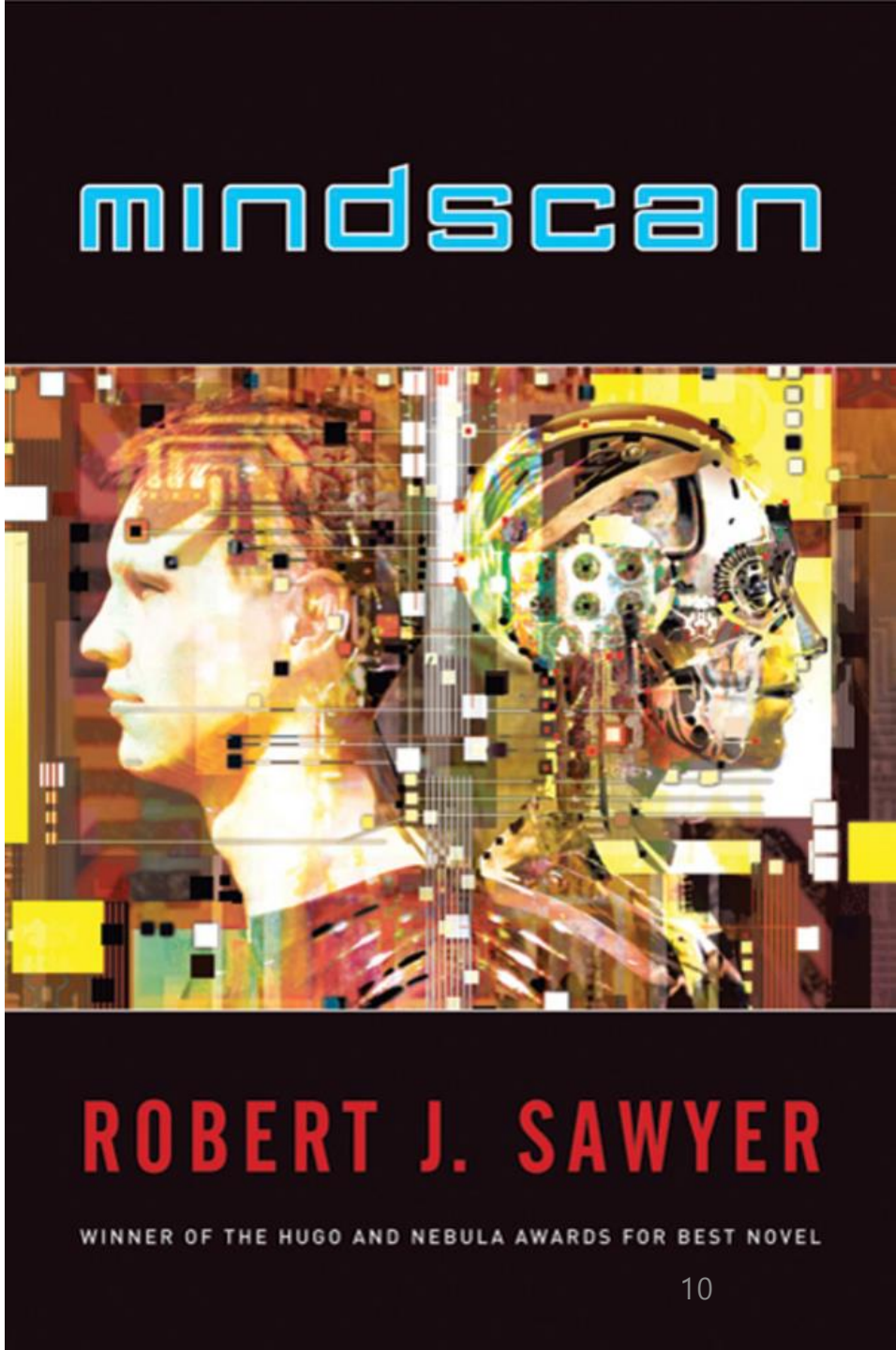
- This is the problem of what makes you the person you are. Soul, Body, or Mind?
- My being alive next week means that I and someone next week are the same person.
- What does it mean for two people at two different times are numerically identical? What is it for me to survive at all?
- What is it by virtue of which a particular self or person continues existing (i.e., persists) over time?

# Theories of Personal Identity

- Soul Theory: soul as a non-material substance, Descartes' *res cogitans*.
- The no-self view: Self is an illusion or a grammatical fiction. Nietzsche, Buddhism, Hume's Bundle theory, Dennett's narrative theory of self.
- Body Theory/Brain-based Materialism: Person/self/mind is physical/material in nature. If the material substrates change, person ends.
- Psychological Continuity Theory(Memory Theory), Software Patternism: I am the pattern of information that constitutes me. What defines who we are is a collection of individual memories and beliefs, thoughts, feelings, hopes and fears.

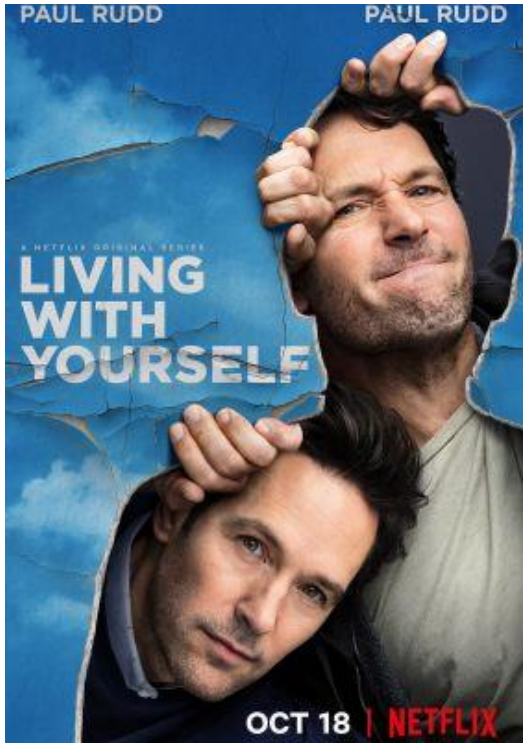
# Mindscan

- Jake Sullivan, a non-operable brain tumor patient, decides to use “mindscan” to upload his brain configuration into a computer and “transfer” it into an android body that is designed using his own body as a template.
- However, after the scan, he found himself with nothing changed. It's Android Sullivan, his clone, that has a new body and life.
- He finds and despair that his fate has not changed a bit.



# Another Me

- The 6th Day (2000)
  - ✓ Starring: Arnold Schwarzenegger
  - ✓ a family man of the future is illegally cloned by accident as part of a vast conspiracy involving a shady billionaire businessman, …
- Netflix Drama: Living with Yourself (2019)
  - the story of a man who, after undergoing a mysterious treatment that promises him the allure of a better life, discovers that he has been replaced by a cloned version of himself…

# Problem of Fission

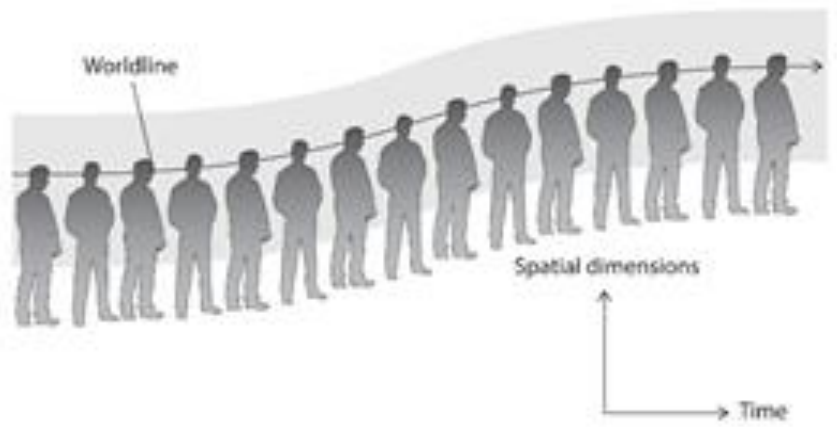
- duplication problem, “the reduplication problem”
- According to patternism, both creatures are Jake Sullivan, because they share the very same psychological configuration.
- But, the new Jake is not the same person as the original Jake. He is just another person with an artificial brain and body configured like the original. They are two different person. Neither of them has any more or any less right to be regarded as Jake Sullivan than the other.
- However, Jake is one person, so only one can be Jake. Both cannot be Jake, because one person cannot be two persons.
- Hence, having a particular type of pattern cannot be sufficient for personal identity.

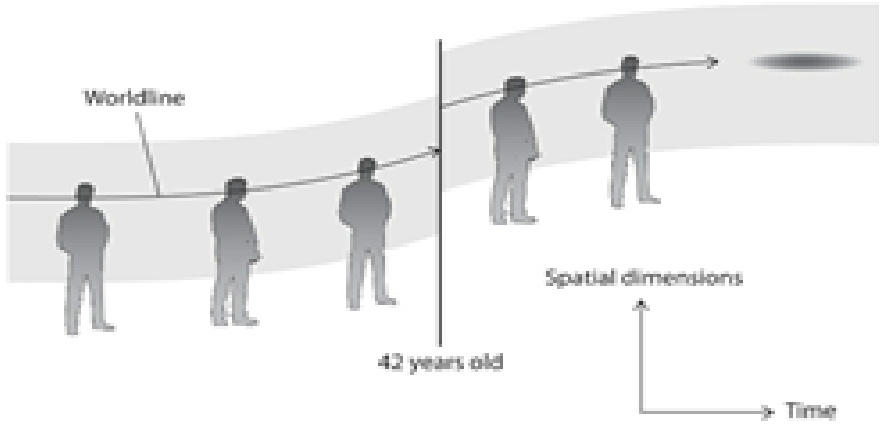
# A Way Out? Modified Patternism

- Your pattern is essential to yourself despite not being sufficient for a complete account of your identity.
- Perhaps there is an additional essential property which, together with your pattern, yields a complete theory of personal identity.
- Personal identity requires that there be a spatiotemporal continuity of a pattern for one to survive.

# Susan Schneider's Diagnosis

- We carve out a sort of "spacetime worm" over the course of our existence.




- The cloned Jake, during the "mindscan", somehow instantaneously moves to a different location in space and lives out the rest of his life. This is radically unlike normal survival. It lacks a requirement for spatiotemporal continuity.

# Schneider: an interim conclusion

- It is possible to distinguish between the original person and its clone through the spatiotemporal continuity condition.
- Uploading does not meet the spatiotemporal continuity condition, and is therefore not compatible with survival.
- If one opts for patternism, enhancements like uploading to avoid death or to facilitate further enhancements are not really "enhancements;" they can even result in death.
- Making copies of a mind does not count as an enhancement, and it is subject to the limitations of its substrate.

# Software Instantiation View of the Mind(SIM)

- A program is a list of instructions in in lines of computer code. A line of code is like a mathematical equation. Equations are abstract entities. Abstract entities are said to be nonconcrete: They are nonspatial, nontemporal, nonphysical, and acausal.
- According to The Software View, our mind is a program. Then is it saying that the mind is an abstract entity?
- We are causal agents in space and time. Our mind (mental process) causes us to act in a certain way in the concrete world.
- We need to distinguish a program from its instantiations. Minds are not programs per se, but program instantiations.
- The mind is the entity running the program (where a program is the algorithm that the brain or other cognitive system implements, something in principle discoverable by cognitive science). (p.138)

# Could Lieutenant Commander DATA be immortal?

- Suppose he finds himself in an unlucky predicament, .. surrounded by aliens that are about to dismantle him. In a last-ditch act of desperation, he quickly uploads his artificial brain onto a computer on the Enterprise. Does he survive? (p. 135).
- Insofar as a particular mind is not a program or abstraction but a concrete entity, a particular AI mind is vulnerable to destruction by accident or the slow decay of its parts, just as we are.
- Data is a particular AI, and as such, he is vulnerable to destruction.
- Uploading would merely create a different token of the same type. An individual's survival depends on where things stand at the token level, not at the level of types.

# Schneider’ conclusion: metaphysical humility

- As alluring as greatly enhanced intelligence or digital immortality may be, there is much disagreement over whether any of these “enhancements” would extend life or terminate it.
- Before we transfer our minds to other substrates or make radical changes to our brains, we must carefully examine the problem of survival.
- Even if the technology of uploading the human brain to the computer is developed, stick to gradual, biologically based therapies and enhancements, ones that mirror the sort of changes that normal brains undergo in the process of learning and maturation, as much as possible.
- Without proper consideration for this, reckless attempts to “enhance” our brain(mind) are a kind of suicide.

# Parfit: The Unimportance of Identity

- What does it really matter?
  - Is it me that survives upload?
  - Is there someone psychologically continuous with me?
- Parfit's Requirement: Just survival is not important in itself. What does matter is psychological continuity.

# A Response to Schneider

- When it comes to survival, the important question is whether what I think is important to me is preserved, rather than just the fact that I survive. What I want from survival is that in the future I will remain the same person as I am today in my psychological characteristics and do the things important to me.
- Under normal circumstances, survival is a prerequisite to getting what is important to me.
- In the "problematic" cases such as uploading, the more important thing seems to be whether there is still someone who is sufficiently similar to (or 'better' than) me.
- **Upload satisfies Parfit's requirement.**

# More Proper Response to Uploading

- Even though Schneider criticizes the Software View of the mind, but still accepts a computational approach to the brain.
- Cartesian Materialism: Patternism, the software view of the mind, and the computational approach to the brain share the assumption of body neutrality.
  - "characteristics of bodies make no difference to the kind of mind one possesses, " and this is in turn associated with the idea that the " mind is a program that can be characterized in abstraction from the kind of body/brain that realizes it " (Shapiro, *The Mind Incarnate*, 2004, 175).
- Separability thesis: Minds make no essential demands on bodies. A humanlike mind could very well exist in a nonhumanlike body.
- To properly evaluate uploading, I think we need to thoroughly examine the body-neutrality assumptions shared by these brain-centeric Cartesian materialists.

# 4E conception of mind

- embodied: mental processes are partly constituted by, partly made up of, wider (i.e., extraneural) bodily structures and processes.
- embedded: mental processes have been designed to function only in tandem with a certain environment that lies outside the brain of the subject. In the absence of the right environmental scaffolding, mental processes cannot do what they are supposed to do, or can only do what they are supposed to so less than optimally.
- enacted: mental processes are made up not just of neural processes but also of things that the organism does more generally — that they are constituted in part by the ways in which an organism acts on the world and the ways in which world, as a result, acts back on that organism.
- extended: mental processes are not located exclusively inside an organism's head but extend out, in various ways, into the organism's environment.

# Embodied Cognition/Mind

- The Embodied Mind Thesis: “minds profoundly reflect the bodies in which they are contained.”
- “psychological processes are incomplete without the body’s contributions. Vision for human beings is a process that includes features of the human body. . . . Perceptual processes include and depend on bodily structures. This means that a description of various perceptual capacities cannot maintain body-neutrality and it also means that an organism with a non-human body will have non-human visual and auditory psychologies. (Shapiro, 2004, 190).”

# 3 different ways of interpreting EMT

- Epistemic: it is impossible to understand the nature of cognitive processes without understanding the wider bodily structures in which these processes are situated.
- Ontic-dependence: cognitive processes are dependent on wider bodily structures in the sense that these processes have been designed to function only in conjunction, or in tandem, with these structures. – it does not, in any way, force us to reject the claim that cognitive processes occur exclusively inside the brain.
- Ontic-Constitution: cognitive processes are not restricted to structures and operations instantiated in the brain, but incorporate wider bodily structures and processes. These wider bodily structures and processes in part constitute — are constituents of — cognitive processes.

# Problem of Grounding/Intentionality

- The body is the constituent element that determines the intentional contents of mental states.
- If the body has changed, the conditions of satisfaction of our beliefs and desires would also change.
- This, in turn, would change the priority of values that we consider important.
  - In the case of an emotional state, it is not possible to specify what the content of the state is without considering the contribution (function, role or reaction) of the body.
  - In determining the priority of values we think important, the emotional content associated with them is an important factor.

# Implications of EMT for Uploading

- ‘Mental states' as information extracted in digital form is incomplete. When we are in a mental state, we must be in the same bodily or environmental conditions associated with it to maintain the same meaning (semantic/intentional) conditions.
  - You can think of music or art as informational structure with mathematical features. For example, consider a song that is digitized and stored on a cd or stored as a file. But this song itself as an informational structure is separated from the semantic condition associated with it. This song as information must be experienced through our body's perceptual system in order to fully realize the meaning it has.
- Brain-uploading that do not include the cloning of the body may not meet Parpit's requirement.
- Successful uploading requires the cloning of the body (and possibly environment) as well as the cloning of the brain.

# Artificial Moral Agent: its Moral Status and Authority

ABSTRACT

According to some, robots may one day be capable of moral agency. They may become better moral agents than humans. In that case, it is suitable for robots to be our moral mentors or even guardians. Call this view "robot paternalism". On the assumption that artificial moral agents are not persons, I argue against robot paternalism. Based on P. F. Strawson's account of participant reactive attitude and Thomas Scanlon's relational account of blame, I argue that it is less suitable for robots than for humans to take paternalist acts toward humans.

KEYWORDS: robot ethics; artificial moral agent; paternalism; participant reactive attitude; meddling blame

## 1. Introduction

With the advance of artificial intelligence, it looks less and less like a sci-fi fantasy that fully autonomous robots will free us from all sorts of laborious, hazardous, menial tasks. To function efficiently, robots must be able to function with minimal human supervision. But how can we be sure that autonomous robots will not harm us or do anything bad? One natural idea is that the decisions robots make must be *morally acceptable for humans*. Robots should learn to conform to moral rules as humans do. In other words, autonomous robots must be *artificial moral agents*, whose actions are roughly as good as ours by moral standards, or even better.

Here I assume a functional or behaviourist conception of moral agency (Floridi & Sanders, 2004; Fossa, 2018; Grodzinsky, Miller, & Wolf, 2008; Gunkel, 2012). A simple way to get hold of this conception is by looking at the Turing Test. The test is designed to determine whether a machine is intelligent by comparing its performance with humans'. If it can perform in some respect as intelligibly as humans, to the extent that we may mistake its performance as humans', then the machine is considered intelligent in that respect.

Similarly, we can devise a Moral Turing Test on robots (Allen, Varner, & Zinser, 2000). A moral robot can act autonomously and cause morally relevant consequences. In the Moral Turing Test, the criteria of moral agency are defined by reference to the currently best moral agents, namely, humans. So, if a moral robot can act in some dimension as morally good as humans, to the extent that we cannot externally distinguish between them in all morally relevant aspects, then it should be judged as a *moral agent in that dimension*. For example, an autonomous car could be judged as a moral agent in the driving dimension if it drives in a way that is externally indistinguishable from a morally good human driver. If a robot can act as morally

good as humans in every aspect of day-to-day life, then it is a *whole moral agent*.

As several philosophers (Bostrom, 2014; Dietrich, 2007, 2011; Fossa, 2018; Hall, 2011; Yudkowsky, 2008) speculates, robots might one day outpace humans in moral performances. If so, given that the Moral Turing Test is purely behaviourist, what defines the criteria of moral agency will be robots rather than humans. They think that robots would then be our *moral mentors*, guiding humans what should and should not do.

Given the behaviourist conception of moral agency, the idea that robots are our moral mentors looks plausible. Consider AlphaGo that beats the human Go masters. It is natural for human Go players to analyse and emulate how AlphaGo plays. But moral robots could go well beyond being our mentors. Indeed, they could be our *moral guardians*: not only do moral robots teach us ethical values and moral norms, but also they may actively intervene to prevent us from committing wrongdoings or self-harm.

Let us call the idea that robots act as our moral mentors or guardians—*robot paternalism*. Robot paternalism maintains that robots are permitted to interfere with a human's autonomy for the sake of her interests or well-being. Robot paternalism is supported by the behaviourist conception of moral agency because the latter assumes that there is no essential difference between humans and robots as far as moral agency is concerned. Hence, the reasons for and against a human's paternalist act toward another human are also available to robots that are facing the same situation. It follows that, if a human's paternalist act is justified, so are robots' in the same circumstances. To be sure, for robot paternalism to be justified, robots need not be morally superior to humans; it only requires that robots be moral agents according to the behaviourist conception. Robot paternalism looks plausible, given the behaviourist conception.

However, I will argue that some reasons that can justify paternalist acts are available only to humans, but not to robots. To be sure, these reasons are *pro tanto*, but still, it implies that in some circumstances a human's paternalist act is justified, but a robot's is not. Why is that the case? To illustrate, let me use an analogy concerning humans' paternalist acts. Suppose James is obese, which could endanger his health. He is accompanied by his mom, Mary, to see a doctor, David. To simplify the issue, let us assume that there are two kinds of agent-neutral reason for or against paternalist acts: considerations concerning James's welfare in favour of paternalism and considerations concerning his autonomy over his life against paternalism. Both reasons are available to Mary and David. If they were the only relevant reasons, then Mary's paternalist act toward James would be equally justified or unjustified as David's. Nevertheless, since Mary is James's mom, she has a strong *agent-relative reason* unavailable to David, which could permit her, but not David, to act paternalistically toward James.

Similarly, I will argue that there are agent-relative reasons available *only to humans* that permit humans to act paternalistically toward fellow humans. The idea that agent-relative reasons can permit or forbid different agents to act paternalistically is all too familiar. But the agent-relative reasons familiar to philosophers are generated by facts about special relationships among the people concerned. So, how is the idea of agent-relative reason relevant to robot paternalism, since the issue is about humans and robots in general, not about ones that are in specific relationships? For the kind of agent-relative reason I will argue is available *to all humans qua human*—or at least to those humans who are capable of being moral agents. Call it *human-relative reason*. Human-relative reason is overlooked in most ethical theories because they usually do not take non-human moral agents into consideration. Without non-human moral

agents, the distinction between agent-neutral reason and human-relative reason makes no difference in practice. Given the possibility of moral robots, I want to highlight the idea of human-relative reason and examine how it could affect the moral relationship between humans and robots.

*A Caveat*: The kind of moral robot under discussion is one that does not have emotions, feelings, and sense of selfhood, despite being capable of moral agency. In other words, they are *not persons*. It is possible, I assume, that robots that lack personhood are capable of moral agency. My assumption is compatible with the behaviourist conception of moral agency, which says nothing about the psychology of robots. My arguments do not apply to robots that have personhood. If robots are persons, human-relative reasons might be available to them provided that their psychology is relatively similar to humans. But I think that creating robots with personhood is unwise. I will not argue for this view since it is not the concern of this paper. The kind of moral robot discussed here lacks personhood, like Data in *Star Trek* or R2D2 in *Star Wars*.

In the next section, I will argue for the existence of human-relative reason—inspired by P. F. Strawson's account of participant reactive attitude—and explain how its existence works against robot paternalism. Combining the idea of human-relative reason and Thomas Scanlon's account of blame, in section3, I argue that robots are ill-suited to blame humans' misconducts. Therefore, even if robots are equally or better capable moral agents, they do not have the same moral standing as we do to take paternalist acts toward humans. A more rightful position of robots in our moral community is thus not mentors or guardians. At best, they can be moral consultants we employ. We may consult with them if we like, or they could advise us when our decisions are really bad. In some special or exceptional circumstances, robots may

even interfere when we conduct our lives badly. But the bar to interfere with our autonomy should be higher for robots than for humans.

## 2. Human-Relative Reason and Robot Paternalism

My argument for human-relative reason is based on P. F. Strawson's ideas about participant reactive attitudes in his seminal article, "Freedom and Resentment" (Strawson, 1974). Strawson maintains that, by default, we adopt the *participant reactive attitude*, namely, that humans are naturally participants in interpersonal relationships, in which we expect others to treat us with respect and goodwill. When our expectation is or is not met, it is natural and appropriate for us to respond to people with what Strawson calls *reactive attitudes*. Reactive attitudes include the emotions such as gratitude, forgiveness, resentment. For example, if someone is kind to us, we naturally and appropriately feel gratitude for her; or if she is hostile or disrespectful to us, our resentment toward her is also natural and justified. To be clear, participant reactive attitude applies not only to those who acquaint with each other, but also to those who are total strangers. To illustrate, consider this example,

> *Resentment*. Jane fell off from stairs. Although she was not seriously hurt, she was in pain and could not temporarily get up by herself. Charlie—who has never acquainted with Jane—saw that Jane needed help, but he simply walked away, showing no sympathy and care. Seeing Charlie's indifference, Jane feels resentment toward him.

Presumably, Charlie is not obliged to help Jane; after all, Jane could eventually get up by herself. According to Strawson, nevertheless, Jane is justified in resenting Charlie for his indifference (certainly, her resentment must be within a reasonable degree). Since they are participants in the interpersonal relationships—despite being total

strangers—their reactive attitudes toward each other's goodwill or ill will can be natural and justified.

We can develop Strawson's insight further: participant reactive attitude can generate human-relative reasons for us to express our concern for others, even at the expense of interfering with their autonomy. When we see that someone is in trouble, participant reactive attitude directs us to show our goodwill and give her assistance appropriate to her and our needs. The normative requirement by participant reactive attitude is a *pro tanto* reason for us to interfere with her life.

The human-relative reason generated from participant reactive attitude is not available to moral robots for obvious reason. Since robots, by stipulation, are not persons, they cannot participate in interpersonal relationships with humans. Also, robots do not have genuine emotions, so that they do not have reactive attitudes to express. Accordingly, robots lack participant reactive attitude to generate reason that might justify their paternalist acts toward humans.

To illustrate, let us imagine a future society in which autonomous robots are widely used, and the robots obey the famous Asimov's Three Laws of Robotics (Asimov, 1950). The Three Laws, essentially, require robots not to allow humans to come to harm, even if humans order robots to allow such harms to happen. Now consider the following case,

> *Suicide*. Tom, who is seriously ill and suffers great pain, is determined to commit suicide. For Tom to commit suicide, however, it would be difficult since robots are everywhere and are more agile and stronger than humans. In obedience to The Three Laws, robots would have to prevent Tom's suicide even if he expresses his determination to die.

While saving a human's life is great, it seems awful to me that, if Tom has thought it through and decided to leave the world, he is forced by robots to live. I think, therefore, that The Three Laws should be revised so that robots should obey the orders of normal adult humans (unless the orders involve harming other humans). You might disagree with me. Since a human's life is invaluable, you might insist that robots should prevent Tom's suicide anyhow. Even so, however, I believe that you would feel somehow uncomfortable with the fact that Tom's autonomy is interfered by robots.

On the other hand, things are quite different if Tom is saved by a human. Imagine this time before Tom is going to kill himself, Rachel happens to pass by. She tries to talk Tom down, though Tom does not waver. He asks and even begs Rachel to let him die. How should we think about this case? While it could be fine if Rachel lets Tom kill himself, I think that Rachel is permitted to save Tom, despite violating his autonomy. Even you think that robots are permitted to save Tom anyway, you would feel much more comfortable with the fact that Tom's autonomy is violated by a human rather a robot.

How should we explain the difference in our attitudes toward the two situations? The difference cannot be explained in terms of some reasons that are available in both: for example, the consideration that Tom's life would be saved is a reason for preventing his suicide, and the consideration that Tom's autonomy would be violated is a reason against it. But it can be explained by human-relative reason generated from the participant reactive attitude. With regards to justifying paternalism, a plausible principle is that, very roughly, the closer the relationship between the agent and the patient is, to a greater degree the agent can interfere with the patient's autonomy. Given this principle, the fact that Rachel can, but robots cannot, engage in an

interpersonal relationship with Tom provides only Rachel with a human-relative reason to save Tom's life.

One may object that, while Rachel has a closer relationship with Tom than robots have, that relationship is too weak to generate reason for Rachel to violating Tom's autonomy. However, I think that we can develop Strawson's insight about participant reactive attitude to support the following idea: a human, X, may act paternalistically toward another human, Y, even though Y asks X not to (X and Y are total strangers). Beyond the reason that it could be good for Y's sake, Strawson's insight indicates an additional reason, namely, that *a meaningful relationship between them may begin*. Back to Suicide, suppose that a few days later after being saved by Rachel, Tom comes to realise that he should not commit suicide. He goes to finds Rachel to express his deep gratitude; they hence become friends. It is not uncommon that people later regret their then decisions and feel grateful for those who were trying to interfere. Sometimes, meaningful relationships are thus forged. Indeed, we often find that those who are willing to speak and act against us are worthy friends because they are genuinely concerned for us. If we are not permitted to take paternalistic actions toward strangers just because they ask us not to bother them, many chances of building interpersonal relationships would be lost. Strawson's insight shows that potential relationship building is a reason for us to interfere, within a reasonable degree, with people's autonomy.

Furthermore, Strawson's insight also indicates a related point that might help reject robot paternalism. From Tom's perspective, if he is truly determined to die, naturally he will feel resentment or embittered toward people who stop him. Strawson says that our reactive attitudes are not directed merely at people's actions, but rather at their qualities of will. Although Rachel is saving Tom's life out of her goodwill, Tom may

resent her for her well-intentioned, but overzealous concern; he may later change his mind and feel grateful to Rachel. Tom's reactive attitudes toward Rachel can be meaningful and appropriate.

In contrast, Tom's resent or gratitude will not be received by the robot since it lacks the will and feels nothing. His resent or gratitude toward the robot would thus be empty. The emptiness of Tom's feelings reveals a radical and profound difference of robot paternalism from human paternalism: that is, robot paternalism would make our reactive attitudes unable to perform the therapeutic function of emotion. Tom's resentment toward Rachel could release his anger and frustration over failing to execute his plan. However, his frustration with the robot could not be realised in the same way. That would make him feel particularly powerless over his ability to control his own life.

To be sure, I do not argue that robot paternalism is always wrong. My thesis is rather modest. I argue that Robert paternalism is more difficult to be justified than human paternalism because human-relative reason is not available to robots (whereas usual reasons for and against paternalism are available both to humans and robots). Paternalism is not the only moral dimension we should take notice concerning the difference in the moral agency of humans of robots: robots also have lower standing to blame humans.

## 3. Can Robots Blame Us?

In this section, I want to argue that moral robots are not fitting to blame humans for their misbehaviours. Other than the Strawsonian account of human-relative reason, I will use Thomas Scanlon's relational account of blame (2008) to support my claim.

In general, a blameworthy action deserves to be blamed. However, even if an action is

blameworthy, some people may lack the standing to blame it. One oft-discussed case is hypocrisy (Coates & Tognazzini, 2018). People who have committed certain blameworthy actions are not suitable to blame others for similar actions.

Another case that receives less attention is *meddling blame*, which will be my target here. Central to the concept of meddling blame is the idea that blaming a blameworthy action is *not the business of the blamer.* When does one's blame count as meddling? Scanlon's account of blame offers a satisfactory explanation: 'To blame a person for an action, in my view, is to take that action to indicate something about the person that *impairs one's relationship with him or her, and to understand that relationship in a way that reflects this impairment*' (Scanlon, 2008, 123; my italics). To illustrate, consider this example:

> *Couple*. Will is arguing heatedly with his wife Kate in a mall about whether to buy a luxurious item. Lizzy, passing by and overhearing their argument, cannot help complaining to Will that he should listen to his wife. In response, Will replies, 'It is not your business'.

Intuitively, Will's response is justified because, according to Scanlon, the argument between Will and Kate does not impair their relationship with Lizzy. Lizzy is thus not in a position to blame Will for arguing with his wife.

Let us consider a revised version of Couple. This time, what Lizzy blames is that Will and Kate argue too loudly in the mall. Lizzy's blame is appropriate because (1) their behaviour is disrespectful of other people who share with them the space, including Lizzy; and (2) her blame suitably reflects that impairment.

What if it is a robot that blames Will and Kate? Since robots are not persons, there is no personal relationship between robots and humans to impair. The views of Strawson

and Scanlon together imply that if a robot blames a human's misbehaviour, that will count as meddling blame. To illustrate, consider this time a robot blames them that they should not argue so loudly (let us assume that the robot is not deployed by the manager to maintain the order in the mall). It seems to me that its blame is not appropriate because the robot is not a person and thus does not receive the disrespect shown by their behaviour. Since the robot has no interpersonal relationship with Will and Kate, the robot is not in a position to blame them.

I do not claim that there is no reason for robots to blame humans' misbehaviours. Sometimes, we blame people in order to educate them not to commit to the same wrongdoings again. Or, we blame people in order to prevent or improve the consequences caused by their misbehaviours. These considerations are valid reasons and available to robots. If I am right, however, the lack of interpersonal relationships between robots and humans make robots ill-suited to blame humans. Instead of blaming in order for education or prevention, robots could be designed to politely request people to stop their misbehaviours, or they could warn people of potential reparation or penalty for their wrongdoings. But if people can be educated or changed concerning their misbehaviours without blaming, then blaming for those reasons is unnecessary or even inappropriate. So, it means that only if blaming is the best option to educate or change people's misbehaviours, it would be appropriate for robots to blame humans. Nevertheless, it still leaves us with the reason that blaming is to reflect the impairment of relationships done by the misbehaviours. Therefore, even when blaming is not the best option to educate or change people's misbehaviours, those people who suffer the impairment done by the misbehaviours are justified in blaming those people.

## 4. Conclusion: Robots as Secondary Agents in Our Moral Community

Based on the views of Strawson and Scanlon, I have argued that the existence of human-relative reason makes it less permissible for robots to interfere with our autonomy. Human-relative reason signifies an essential moral dimension of human interaction. We humans are participants in interpersonal relationships, which requires us to treat each other with goodwill and respond to their qualities of will with suitable reactive attitudes and actions. So, humans, by default, possess a moral standing that provides us reason to enhance or impair the relationships with each other. On the contrary, since robots are, by assumption, not participants in interpersonal relationships, human-relative reason is not available to robots.

While I accept that the behaviourist conception of moral agency may imply that robots could be as capable as, or even more capable than, humans with regards to moral agency, it does not follow that they are our moral peers or even superiors. On the contrary, I argue that, even if robots are more capable moral agents than humans, they still have the lower moral standing than us, so that it is less suitable to let robots to correct our misconducts. That might sound odd; one might think that, if what we do is indeed wrong and robots can recognize that, why shouldn't robots correct us? An analogy might help. Consider a boss and her employee. She is making a bad business decision. Even if the employee does know better than his boss, it would be ill-suited for him to blame his boss or to overturn the decision without her consent. Of course, he could advise her. But, ultimately, it is up to his boss. That is what I think a more proper relationship of robots with humans. Not mentorship, nor guardianship. In other words, autonomous robots would be secondary agents in our moral community, despite being equally or more capable moral agents.

References


Allen, C., G. Varner, & J. Zinser. 2000. "Prolegomena to Any Future Artificial Moral Agent." *Journal of Experimental & Theoretical Artificial Intelligence* 12:251-261. doi.

Asimov, Issac. 1950. *I, Robot*. New York: Spectra.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Coates, D. Justin, & Neal A. Tognazzini. 2018. Blame. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Dietrich, Eric. 2007. "After the humans are gone Douglas Engelbart Keynote Address, North American Computers and Philosophy Conference Rensselaer Polytechnic Institute, August, 2006." *Journal of Experimental & Theoretical Artificial Intelligence* 19 (1):55-67. doi: 10.1080/09528130601115339.

Dietrich, Eric. 2011. "Homo Sapiens 2.0." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 531-538. Cambridge: Cambridge University Press.

Floridi, Luciano, & J. W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds and Machines* 14 (3):349-379. doi.

Fossa, Fabio. 2018. "Artificial moral agents: moral mentors or sensible tools?" *Ethics and Information Technology* 20 (2):115-126. doi: 10.1007/s10676-018-9451-y.

Grodzinsky, Frances S., Keith W. Miller, & Marty J. Wolf. 2008. "The Ethics of Designing Artificial Agents." *Ethics and Information Technology* 10 (2-3):115-121. doi: 10.1007/s10676-008-9163-9.

Gunkel, David J. 2012. *The Machine Question: Critical Perspectives on Ai, Robots, and Ethics*: MIT Press.

Hall, J. Storrs. 2011. "Ethics for Self-Improving Machines." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 512-523. Cambridge: Cambridge University Press.

Scanlon, Thomas. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*: Belknap Press of Harvard University Press.

Strawson, P. F. 1974. *Freedom and Resentment and Other Essays*. London: Routledge.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Bostrom Nick and Milan M. Ćirković, 308-345. Oxford: Oxford University Press.