Journal of Philosophy of Life Vol.13, No.1 (January 2023):125-134

Reconstructing Agency from Choice

Yuko Murakami[*]

1. Introduction

Research in artificial intelligence (AI) seems to cast doubts on the presuppositions of mind-body connections. In fact, AI by itself does not reveal the inadequacy of the mind-body problem setting. AI mimics an accelerated cognitive process that amplifies the dissonance of our models of humanity. Most of us tend to feel uncomfortable when faced with information technology, even without knowing what philosophers have discussed.

Traditional cognitive science is one field of dissonance and it assumes that the brain plays an essential role in performing mental activities. It is also a common belief that the mind supervenes the body. Researchers on AI focus on brain science because they believe that the mind is reduced to, or at least corresponds to, the activity of the body, particularly the brain. Observational results of vital signals, such as cerebral blood circulation, are considered physical counterparts of mental activities. For example, Kamitani [1] decodes fMRI results to display what experimental subjects see or even dream. This shows that there seems to be a direct correspondence between the mind and the brain. Such traditional approaches toward mind and person involve tacit assumptions that (1) each individual human body is connected; (2) the human mind is associated with a human body in the sense that mental activities correspond to bodily states and cannot exist without these bodily states; and (3) physical bodies outside of one's own human body are not construed as necessary factors of her mental activities. AI undermines the intuitive justification of these assumptions.

In contrast, some AI research can assume another set of assumptions. For example, the transhumanist notion of upload human presupposes the possibility of total separation of the mind and its original body, particularly when they insist on "copying information to media other than the body." The concept of the

[*] Professor, Graduate School of Artificial Intelligence and Science, Rikkyo University. Ikebukuro Campus 3-34-1 Nishi-Ikebukuro, Toshima-ku, Tokyo Japan 1718501.
Email: yukoim@rikkyo.ac.jp

individual person in the transhumanist picture is like an iPhone account. Your account information is simply saved on a cloud system; when the hardware of your iPhone is damaged or outdated, you can upgrade it to a new one by downloading all of the data from the cloud. Thus, you can replace your iPhone without stress.

Both brain scientists and transhumanists share in the assumption that each personal mind is associated with a single physical body, while the possibility of replacing body parts remains controversial.

The disconnection between mental activities and human bodies has been noted in history, even before AI began to flourish. It is seen in the series of externalization of mental contents and extension out of the human body, as will be discussed in the following section. Moreover, a wide range of scientific research suggests that it not only humans have the capacity to think, collect information, use tools, and use language to communicate, judge, and reason. Intellectual superiority to other creatures is no longer regarded as the basis of "humanity."

What then should serve as the conceptual basis of humanity? We claim that it is the concept of agency, since human beings are social beings. We must act interpersonally to be counted as humans. In this paper, we methodologically eliminate the distinction between human and non-human to reconstruct the notion of agency without relying on reduction to individual agents associated with each individual human body. According to the "extended notion of agency," an agency is a dynamic system of continuously choosing the world where the agent resides by describing a set of possible choices and selecting priorities among those choices. Such a system can have both human and non-human components. A historical series of choices by an agent formulates or defines the personality of the agent. If such conceptual model is right, (1) the concept of agency will be reduced to a temporal model of choice; (2) an agent is a system of dependence among component systems; (3) the dependence relation of an agent can be non-well-founded; (4) the identity of two agents will be defined as an existence of a bisimulation relation between dependent components; and (5) the concept of rationality of agents will likewise be non-well-founded.

2. From Extended Mind to Extended Agency

2-1 Precursor: Externalism of meaning

Since the mid-twentieth century, there have already been trends toward externalization in the history of philosophy [1]. Before Putnam [2], the meaning of language, which is the basis of language use regarded as activity in the mind, was considered mental content. However, he argues that meanings are not found in the brain; rather, it is external conditions that determine meanings (externalism of meaning). He established this through the twin earth thought experiment. Being a thought experiment, its physical feasibility is irrelevant.

Now, let us think about “Twin Earth,” which has almost the same structure as the earth we live in, with the same people who have the same things. The same person is the same in all properties, especially in the substances that make up the brain. The only difference is that the liquid (that is, water) indicated by “H_2O” in the language used by this side is replaced by “XYZ”, which has another chemical composition but shows the very same property as water. Residents of the Twin Earth call “XYZ” water. It is assumed that Alice, who is a resident of this earth, and Alice', a resident of twin earth corresponding to Alice in this earth, are talking about a statement that makes some claim about water. The question is: are Alice and Alice' talking about the same thing? In other words, do their sentences have the same meaning? We answer in the negative. Given that Alice talks about H_2O and Alice' talks about XYZ, these two sentences must have different meanings (It is as though talking about two different people with the same name). Since we have assumed that the brain structures of Alice and Alice' are the same, this example shows that relying on brain structure alone is not enough to determine whether or not the meaning is the same. Simply put, brain structure is not enough to determine meaning.

2-2 Extended Mind Thesis (EMT)

The next externalization was memory, which has also been considered to be “mind-related.”

Memory, as well as language use, has been regarded as human heart activity. Clark [3] claimed that the Extended Mind Thesis (EMT), inspired by the example of a note written in a notebook, could extend one’s memory. He thus formulated the Parity Principle [4].

> [Parity Principle] When we work on something and accept it without hesitation as a part of the cognitive process and as a part of the world, it enters our mind that part of the world is part of our cognitive process.

As regards this issue, it was previously claimed that the way of searching for information differs between inside and outside the mind [5]; however, several more digital devices are in use now than when the extended thesis was claimed. Despite the habit of using memos for memorandums, we are now experiencing lifestyle changes that lessen the need for setting concrete meeting places and instead encourage the use of mobile communications.

2-3 Extended Agency Thesis

Technologies have ubiquitously penetrated our life, such that the earth is no longer resilient; human beings can now go and act further than ever. In particular, on the basis of AI-related technology, the notion of agency should be extended from agency of an individual to agency in general.

Consider the case of when you start learning to play a musical instrument. You are gradually taught how to hold the instrument and how to move your body; it takes a lot of practice before you can start playing the correct sounds. It entails moving your body in an extremely unnatural manner and using muscles that you never intentionally moved until you started using the instrument. It is only when these unnatural movements can be performed easily, without hesitation or effort, that you can finally claim to be a good instrumentalist. Well-trained players reshape their bodies according to the instruments and play music as they think, as though their reshaped bodies and the instruments are united.

Similar to music instruments, we extend our own body to involve our social systems. In addition to driving a car and riding a public transport system, using information devices such as smartphones and tablet computers require training before you are able to use them. Once you have access to these systems and devices, using them to achieve what you set out to do becomes a trivial matter. For example, a public transport system is not owned by you (unless you are the owner of the transportation company). It is merely available for use for a fare. Its operation may not be exactly commensurate with your wishes, but you undertake trips with it nonetheless.

All parts of such a system are often interdependent. Adjusting one part affects other parts, and changes in other parts also bounce back to the original part. By putting dependencies, access relations, and reference relations together and describing them as a subsystem, there is no guarantee that atomic elements exist in the "part" of the agents. They form a network. Furthermore, it is necessary to consider actions separately from human physical activity even for agents considered to have intentions. The mainstream version of the notion of behavior is that an event in a causal sequence with intention is identified as an activity. The presence or absence of the agent's intention explains why the agent caused the event. It is considered an ability. Although there are variations in terms of whether cause and reason are considered to be the same, recent tendency to seek explanatory ability in autonomous AI assumes this type of action theory.

However, there are actions that cannot be handled by the type of action theory discussed in the previous section. In other words, it is not always appropriate to place an action in a causal sequence. Note that human actions do not necessarily involve the physical acts of the agents of the acts [6]. For example, an act of omission, or "I have a duty to do that and I can do it. But I do not," may not be associated with any physical movement. It is an omission of the organization that the responsible agency knows the possibility of drug damage and does not cancel the drug approval. Moreover, there may be omissions with regard to actions involving language. Being silent in a conversation is an act of omission with no physical movement. In other words, the premise that actions always have physical basis is not always true.

Furthermore, if physical entities can be considered separately from the cognitive process, the claim of independence between the act and the physical basis does not contradict the parity principle.

Of course, we can also think that the results of other people's actions have causally contributed to the realization of our own actions. However, what I would like to emphasize here is that rather than taking into account the causality of the consequences of these people's actions, it is important to note that these actions are carried out and that the credibility of the equipment is almost self-evident. It is said that this has a structure similar to that of Clarke's extended thesis example. Thus, the idea of the agent itself should be expanded.

Is it possible to take agents as primitive concepts while denying physical reductionism of agency? I would like to think so. Nakayama [7] [8] set forward

Clark's EMT as a criticism of the idea based on the assumption that the mind is in the brain and also in contexts including body extensions such as cyborgs. On the basis of mereology, he formulates the extended agent as follows.

(a) [Atomic Agent] An atomic agent is an agent. Any spatial part of an atomic agent is not an agent. Here, we simply presuppose that there are atomic agents. An atomic agent constitutes the core (or one of the cores) of any extended agent.

(b) [Agents and Tools] Let temporal-part (x, t) denote the temporal part of object x in time t. Let A be an agent that uses (tool) B in time t to perform an action. Then, the (four-dimensional) mereological sum, temporal-part (A, t) + temporal-part (B, t), is an agent. We can easily prove within the four-dimensional mereology that temporal-part (A+B, t) = temporal-part (A, t) + temporal-part (B, t).

(c) [Collective Agent] If agents A1An perform a joint action, A1An is an agent (for the notion of joint action, see [21]).

(d) If an object satisfies neither (2a) nor (2b) nor (2c), it is not an agent.

(e) [Extended Agent] An agent that is not atomic is called an extended agent.

Nakayama, by this formulation, aims to characterize the mind as situations and processes associated with agents. The notion of agency—not of the mind—is primitive.

Contrary to Nakayama's claim, agents may not be reduced to atomic agents because agency is a non-well-founded concept that may include cycles in the mereological structure of agents.

We seem to believe that our own body, without doubting that it is given to us, is our property, and controllable to us. However, we assume physical systems including musical instruments and social systems as an extension of our body—the apparatus of our mind in the causal sphere. Consciousness on one's own body is vague. The body can be moved as usual unconsciously. However, when I think about my illness and pregnancy, I notice that my physical condition changed even before I became aware of it; my control over this change is very limited. In addition, awareness may be directed to a part of one's body only through an abnormal situation such as pain. In other words, it is only a belief that you can control your own body.

The boundaries within the human body are not clear. The external boundary is physically visible with the naked eye and is considered to be the boundary as a human individual. It is a cylindrical mass with irregularities and a large number of microorganisms inside of it. The body is attached to the inside of the mass as if there are various things floating inside the liquid. Since this microbe is outside the tube, it is outside of its own body; but because the tube itself is regarded as an individual, it is thought to be inside the (digestive) individual. However, in terms of the behavior of these microorganisms, they are not directly controllable. It is hard to try to change the internal environment by pouring some liquid or solid inside the cylinder. In addition, intentional control of the behavior of the cells that make up the body is almost impossible with willpower alone; thus, people try to control it through nutrition intake, the use of drugs, or by performing genetic manipulation.

> [Parity Principle of Agency] When we act and accept without hesitation as part of the cognitive process, as part of the world functions as if it gets in our mind (at that time) part of the world is part of the process of our agency.

3. Formalization: Choice and agency as primitive concepts

Now, we will delve into the main proposal of the paper. With choice being a primitive concept, we can define the notion of agent/ person on the indeterministic temporal frame proposed by Prior and Thomason [12] [13]. Here, the notions of agency and choice are extended versions of the STIT theory [7] [8]. In STIT, the formulation focuses on an evaluation of an action at the moment of the choice toward the action or at the moment strictly after the choice. Our idea is based on the first formulation of evaluation of action at the moment of choice, while modifying the definition of each "agent" on semantics. The following is a sketch of our formulation.

A tree structure (*branching-time frame*, BT) is a pair F = 〈T,<, where T is a nonempty set, whose members are called *moments*; < is a tree relation on T (a partial order on T being (1) linear to the past and (2) connected). A *history in F* is a maximal linear subset of T. A history h is said to *go through a moment* m when m ∈ h. A *moment-history pair* is a pair of a moment m and a history that goes through m. H_m denotes the set of histories that go through m; HF denotes the set of all the histories in F; and Moment-History denotes the set of all moment-history pairs in F.

We now extend branching frames of time with "choice." Consider a branching-time frame F = $\langle T,<\rangle$. A *choice at a moment* m $\in$ T is a partition of the set of moment-history pairs $\langle m,h\rangle$. A *choice set at* m is a set of choices at m. Note that it may be the empty set or the powerset of the set of all choices at m; an arbitrary set of choices at m may play a choice set at m. C is called *the choice set on F* if C is the set of choice sets at m where m is in T. A *choice frame* is F_C = $\langle T,<,C\rangle$, where F = $\langle T,<\rangle$ is a branching frame and C is the set of choice sets on F.

An *agent* in a choice frame F_C is a series of choices along a history. A *fragment of an agent* a in F_C is a series of choices along a consecutive fragment of a history in F_C.

The choice frame seems similar to the model that Parfit describes when he argues partial survival [Parfit 10: 298-302] and successive selves [Parfit 10: 302-305]. Our proposal differs from Parfit's in that each successive self is formed by itself with the series of choices on the branch. Choice is the primitive notion that defines the notion of person. With the choice frame version of the notion of agents, Parfit's division of me is just creating two agents with an identical fragment of agents in the initial part of history.

The choice-based notion of agents may seem strange as it is meaningless to say, for example, "I regret not doing X" as "I" refers to the agent on the very history where "I" resides. The agent who might have done X is not the same agent of "I." Such failure of transworld personal identity makes counterfactual statements about any person meaningless. For example, the counterfactual sentence "Hanna should have left the heavy shoes home" seems to require transworld identity among Hanna in the world where the sentence is stated (Hanna_1) and Hanna in the world where she wisely leaves the heavy shoes home (Hanna_2).

The solution to the problem is rather simple. There are two ways of identifying agents/person: one is in the causal structure of choice and the other is in the world of reason. Physical continuity is in the causal world and may not be reflected in the world of reason. The justification of actions is made in the world of reason, which requires transworld identity of the person to consider counterfactual cases, while also being based on an integrated series of choices among those agents in the causal world, which are associated to be identical on the world of reason.

It is the generalization of the notion of person. Physical continuity in Parfit requires the determination of a person to consider only the agents that share fragments of agents. Identification of “I” can be formulated as determination of the fragments of agents. Such a determination is to define one’s own life, which is an essential component of human dignity. In fact, it is the privilege of human beings to define the world he selectively lives in and the meaning of the language he speaks. He is both a component of the world and an agent who gives meanings and conventions in the world. The determination must be socially shared to make the notion of person significant as any person cannot be isolated from the rest of the community. Transhumanists deny such physical continuity but claim that psychological continuity is enough. However, it lacks social determination, and consequently such an identification of self is socially meaningless.

References

[1] Lau, J & Deutsch, M. “Externalism About Mental Content”, *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/content-externalism/>.

[2] Kamitani, Y., & Tong, F. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5), 679–685. 2005. doi:10.1038/nn1444

[3] Putnam, H. “Meaning and Reference.” *Journal of Philosophy* 70(19), pp. 699–711, 1973. doi: 10.2307/2025079

[4] Clark A. & Chalmers, D. “The Extended Mind.” *Analysis* 58(1), pp. 7–19, 1998.

[5] Clark, A. *Supersizing the Mind*. Oxford University Press, Oxford, 2008.

[6] Simon, H. *The Science of Artificial*. MIT Press, 1969. 3rd Ed. 1996.

[7] Belnap, N. “Backward and Forward in the Modal Logic of Agency.” *Philosophy and Phenomenological Research*, 51(4), pp. 777–807, 1991.

[8] Belnap, N. “Before Refraining: Concepts for Agency.” *Erkenntnis*, 34(2), pp. 137–169, 1991.

[9] Nakayama, Y. “The extended mind and the extended agent.” *Social and Behavioral Sciences* 97, pp. 503–510, 2013.

[10] Parfit, D. *Reasons and Persons.* Oxford University Press, 1984.

[11] Parfit, D. Divided Minds and the Nature of Persons. In Colin Blakemore & Susan A. Greenfield (eds.), *Mindwaves*. Blackwell, pp.19–28, 1987.

[12] Prior, A. *Past, Present and Future*. Oxford University Press, Oxford, 1967.

[13] Thomason, R.H. Indeterminist time and truth-value gaps. *Theoria*, 36, pp. 264–281, 1970.