philosophies

MDPI

Article

Artificial Pain May Induce Empathy, Morality, and Ethics in the Conscious Mind of Robots

Minoru Asada

Institute for Open and Transdisciplinary Research Initiatives, Osaka University, Suita 565-0871, Japan; asada@otri.osaka-u.ac.jp

Received: 6 June 2019; Accepted: 8 July 2019; Published: 13 July 2019

Abstract: In this paper, a working hypothesis is proposed that a nervous system for pain sensation is a key component for shaping the conscious minds of robots (artificial systems). In this article, this hypothesis is argued from several viewpoints towards its verification. A developmental process of empathy, morality, and ethics based on the mirror neuron system (MNS) that promotes the emergence of the concept of self (and others) scaffolds the emergence of artificial minds. Firstly, an outline of the ideological background on issues of the mind in a broad sense is shown, followed by the limitation of the current progress of artificial intelligence (AI), focusing on deep learning. Next, artificial pain is introduced, along with its architectures in the early stage of self-inflicted experiences of pain, and later, in the sharing stage of the pain between self and others. Then, cognitive developmental robotics (CDR) is revisited for two important concepts—physical embodiment and social interaction, both of which help to shape conscious minds. Following the working hypothesis, existing studies of CDR are briefly introduced and missing issues are indicated. Finally, the issue of how robots (artificial systems) could be moral agents is addressed.

Keywords: pain; empathy; morality; mirror neuron system (MNS)

1. Introduction

The rapid progress of observation and measurement technologies in neuroscience and physiology have revealed various types of brain activities, and the recent progress of artificial intelligence (AI) technologies represented by deep learning (DL) methods [1] has been remarkable. Therefore, it appears inevitable that artificial consciousness will be realized soon; however, owing to the fundamental limitations of DL, this seems difficult. The main reason for this is that the current DL emphasizes the perception link between the sensory data and labels, lacking strong connections with the motor system; therefore, it does not seem to involve physical embodiment and social interaction, both of which develop a rich loop by including perception and action with attention, cognition, and prediction (Figure 1). This is essential for consciousness research, including unconsciousness.

Cognitive developmental robotics [2] has been advocating the importance of physical embodiment and social interaction, which has great potential in overcoming the above-mentioned limitation. The ideological background of a constructive approach by CDR is well-explained in a book by Jun Tani [3] in chapter 3, featuring Husserl's Phenomenology, as follows (Figure 2):

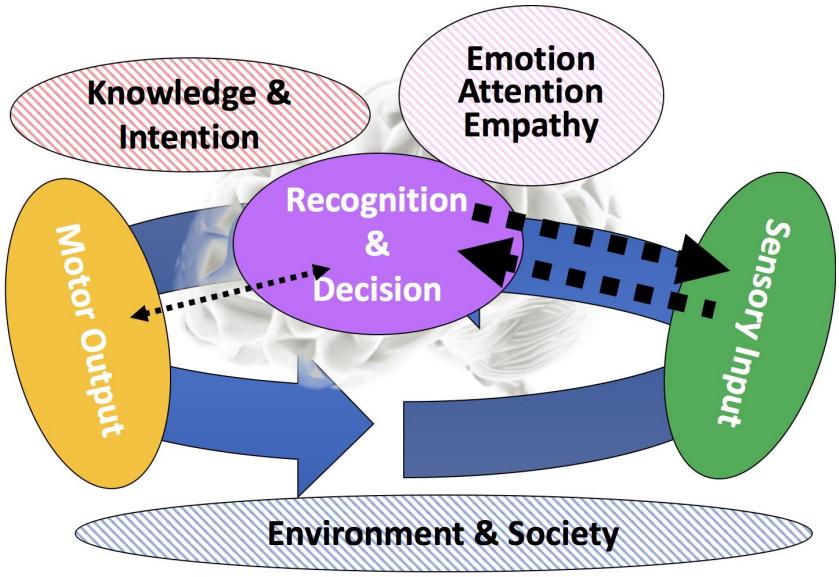

Figure 1. The current status of deep learning.

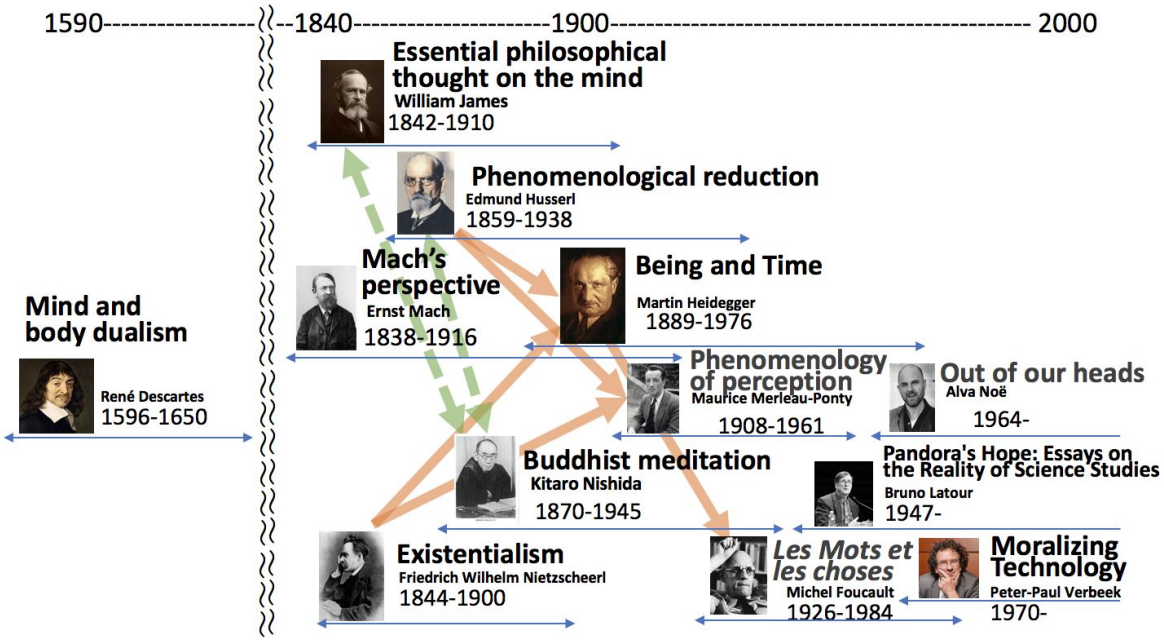

Figure 2. Ideological background of relations between human beings and things.

With regard to the relationship between the mind and body or things, it is Descartes who advocated mind and body dualism and laid the foundation of modern philosophy. It is Husserl who insisted on "New Cartesianism" that goes beyond Descartes to transcendental phenomenology and gave phenomenological consideration. He developed a way of thinking of subjectivity between subjective and objective intervals, and greatly influenced the generations after him. He predicted that the analysis of nature was based on individual conscious experiences. Heidegger and Merleau-Ponty extended and evolved Husserl's phenomenological theory.

Heidega argues that the concept of "being-in-the-world" emerges through dynamic interaction between the future possibilities of individual agents and their past possibilities without separating subjectivity and objectivity. He also pointed out the importance of some forms of social interaction

where individuals can mutually exist under the prior understanding of how individuals interact with purpose.

Merleau-Ponty argues that in addition to subjectivity and objectivity, the dimension of "physical embodiment" emerges, where a body of the same thickness is given to objects that are touched or viewed when the subject is touching and seeing, and that the body could be a place where exchanges between two subjective and objective poles are repeated. In other words, he pointed out the importance of the body as a media connecting the objective physical world and subjective experience. As mentioned above, this is the basic concept of "physical embodiment" in cognitive development robotics.

Based on these ideological backgrounds, CDR has done several studies where computational models were proposed to reproduce cognitive developmental processes by utilizing computer simulations and real robot experiments. Although CDR has not mentioned consciousness explicitly, here we argue any possibility of artificial consciousness more explicitly by proposing a working hypothesis based on the nervous system of pain sensation. The hypothesis includes the following points:

1. A pain nervous system is embedded into robots so they can feel pain.
2. Through the development of MNS, robots may feel pain in others.
3. That is, emotional contagion, emotional empathy, cognitive empathy, and sympathy/compassion can be developed inside robots.
4. Proto-morality emerges.
5. Robots could be agents who could be moral beings, and at the same time, subjects to moral consideration.

In this article, this hypothesis is argued from several viewpoints towards its verification. The rest of the paper is organized as follows. First, the nervous system for pain sensation is briefly explained from a neuroscientific viewpoint. Next, a preliminary experiment using a soft tactile sensor is shown as a potential artificial nociceptor system. Then, we argue any possibility of artificial empathy, morality, and ethics in CDR by integrating existing studies and future issues. The above hypothesis can be regarded as the developmental process for artificial consciousness.

2. A Nervous System for Pain Sensation

The perception of pain called nociception has its own nervous pathways, which are different from mechanosensory pathways (see Chapter 10 in [4]). Figure 3 shows these two pathways. The pain nervous system transmits two kinds of information through the anterolateral system: the sensory discrimination of pain (location, intensity, and quality), and the affective and motivational responses to pain. The former terminates at the somatosensory cortex (S1, S2) while the latter involves anterior cingulate and insular regions of cortex and the amygdala. The pain matrix consists of these four regions (four rectangles bounded by red lines in Figure 3). They are ascending pathways.

The left side of Figure 4 shows the discriminative pain pathway (red) along with the mechanosensory pathway (dark blue), both of which are ascending pathways. The analgesic effect arises from the activation of descending pain-modulating pathways that project to the dorsal horn of the spinal cord from the somatic sensory cortex through the amygdala, hypothalamus, and periaqueductal gray, then some parts of the midbrain (e.g., raphe nuclei), and regulate the transmission of information to higher centers. Such projections provide a balance of inhibitory (past view) and facilitatory influences that ultimately determine the efficacy of nociceptive transmission. Figure 3 shows such descending pathways (broken blue arrows), and the top right of Figure 4 indicates the local interaction of this descending pathway from raphe nuclei.

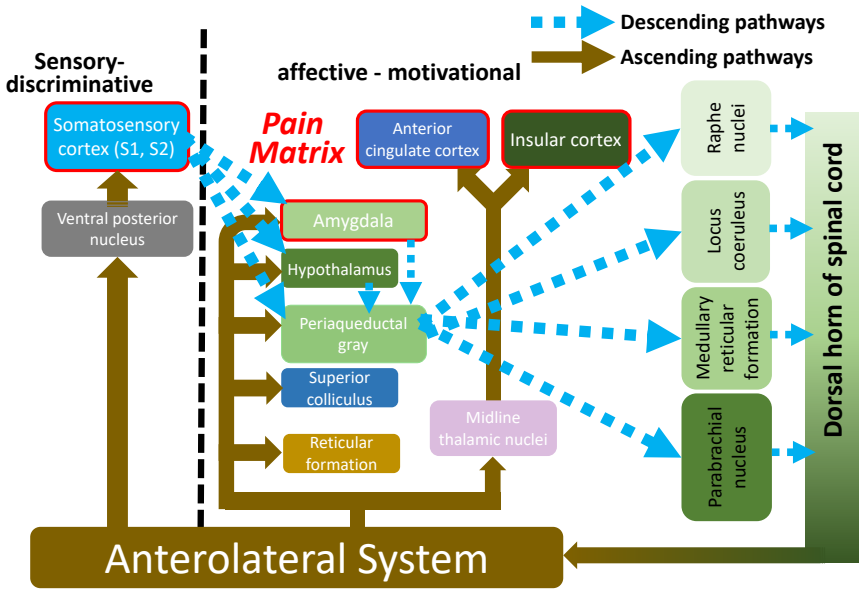

Figure 3. Pain matrix (adopted from Figures 10.5 and 10.8 (A) in [4]).

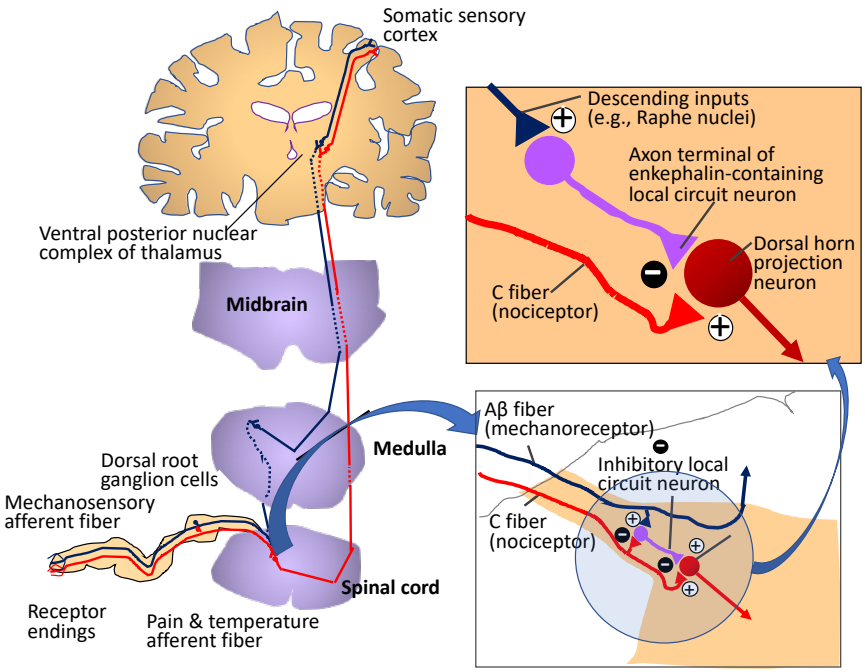

Figure 4. A discriminative pain pathway and mechanosensory pathway (adopted from Figures 10.6 (**A**) and 10.8 (**B** ,**C**) in [4]).

In addition to the above projections, local interactions between mechanoreceptive afferents and neural circuits within the dorsal horn can modulate the transmission of nociceptive information to higher centers. The bottom right of Figure 4 indicates this situation, which is called the "gate theory of pain" by Ronald Melzack and Patrick Wall [5]. They proposed that the flow of nociceptive information through the spinal cord is modulated by concomitant activation of the large myelinated fibers associated with low-threshold mechanoreceptors [4]. It explains the ability to reduce the sensation of sharp pain by activating low-threshold mechanoreceptors (kiss it and make it well).

3. From an Artificial Pain System to a Moral Being

3.1. Artificial Pain

Artificial pain by itself is not a new idea for robots. For example, Kuehn and Haddadin [6] embedded a pain nervous system based on the knowledge and findings of the corresponding human system, and designed reflective behavior for robots to avoid and/or reduce the pain sensation. They applied a spiking neuron model which they called aRNS (artificial Robot Nervous System) to a three-layer skin structure in their simulation. For the experiment with a real robot, they used a kind of tactile sensor to simulate the aRNS. Their motivation was to apply the idea of ensuring safety in human–robot collaborations, and the main focus was the generation of avoidance behavior. Therefore, the issue of empathy with humans after the experiences of pain was not considered.

Figure 5 shows a block diagram of the process of the early stage of pain experiences based on the findings and knowledge from neuroscience described above. Tactile sensation is received by two types of receptors—the mechanoreceptor and nociceptor. Both reach the somatosensory cortex (discriminative), but only nociception reaches the amygdala, anterior cingulate cortex, and anterior insular (affective and motivational). These areas contribute to the formation of emotional states. Visual images of painful situations are memorized, and pain relief behavior, such as avoiding and rubbing, is learned to reduce the pain sensations in the self. Local feedback from the mechanoreceptor by rubbing behavior is added (negative) to the nociceptor path at the dorsal horn of the spinal cord (gate theory). In addition to the descending pathways from receptor endings to the cortex, the ascending pathways from the somatosensory cortex to the dorsal horn of the spinal cord through periaqueductal gray and raphe nuclei enhances the local feedback (reduce the pain, see Figure 3).

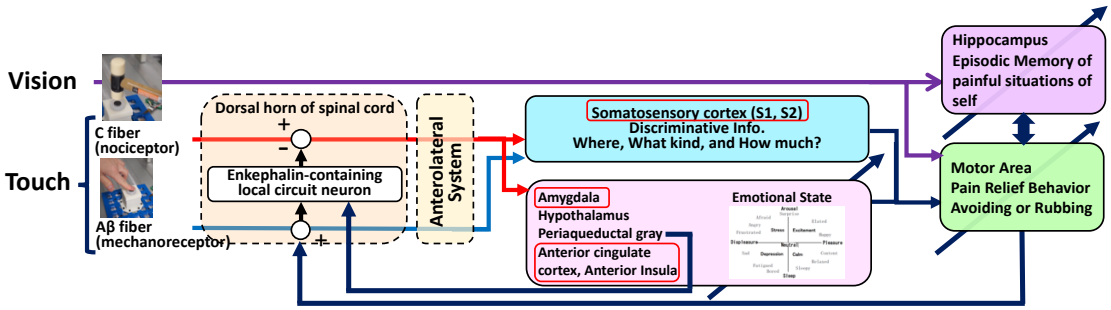

Figure 5. The early stage of pain experiences.

As a preliminary step of the artificial pain nervous system, we developed a soft tactile sensor [7] consisting of four spiral inductors printed on a flexible printed circuit board (FPCB) and a disk-shaped magnetorheological elastomer (MRE; ferromagnetic marker) embedded in a cylindrical elastomer made of silicon rubber (see Figure 6). The inductances of the inductors can be determined by the positional relationship between the ferromagnetic marker and each inductor because the marker contains iron particles with high magnetic permeability. Therefore, the sensor can estimate applied tri-axis forces by monitoring the inductance changes caused by three-dimensional (3D) displacements of the marker. Figure 7 shows the results of tactile sensation. The two pictures on the left show soft (rubbing) and hard (hammering) touches by an index finger and a hummer, respectively, and the figure on the right shows a time-course of three forces F_x, F_y, and F_z when the soft and hard touches were applied. As the figure shows, the waveform for the hard touch is sharper than that of the soft one, so the sensor can easily distinguish between soft and hard touches from their response waveforms.

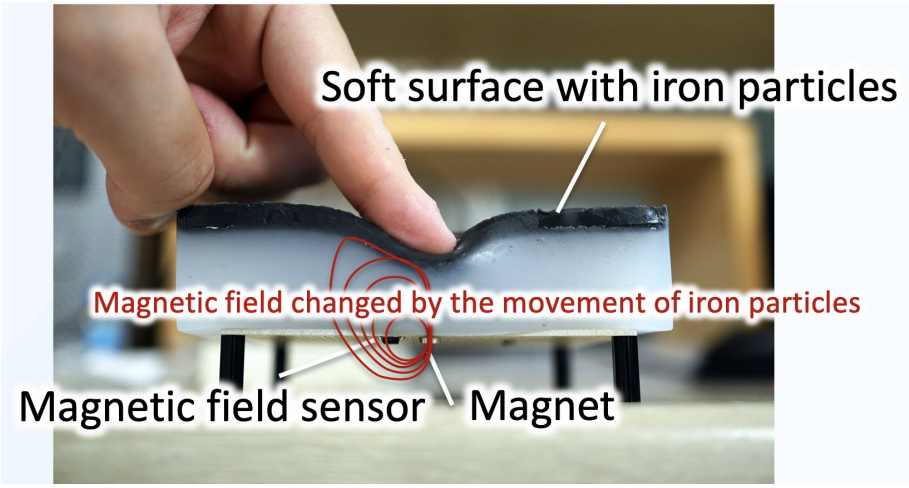

Figure 6. Basic structure of the proposed flexible tactile sensor, which can detect an applied normal force and vertical deformation.

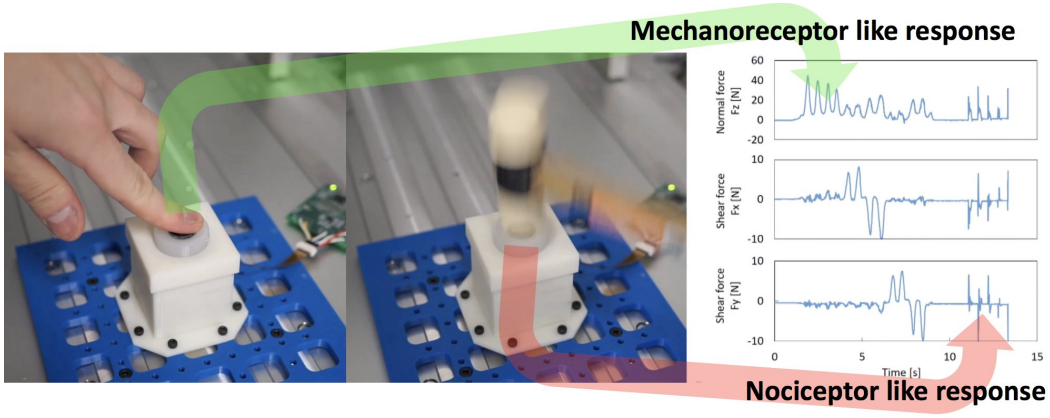

Figure 7. A soft tactile sensor that discriminates soft and hard touches as a mechanoreceptor and nociceptor, respectively.

Based on the capability of recognition by the tactile sensor, the artificial nervous system for pain sensation can be embedded into the robot's body and brain in parallel with normal mechanoreceptor pathways with mechanisms for pain regulation (the gate theory).

3.2. *Artificial Empathy*

The feeling that this pain is shared is thought to be the source of empathy assuming that the MNS enables agents to perceive others' pain[1]. Singer et al. [8] conducted a functional imaging study of empathy where subjects observe their loved ones in physically painful situations, and found that only the parts of the pain network associated with its affective qualities, but not its sensory qualities, mediated empathy. These parts are the anterior insular (AI) and the rostral anterior cingulate cortex (ACC).

A survey of articles on artificial empathy [9] mentioned that many papers on empathy in neuroscience, cognitive science, and psychology dealt with pain as a research target. The design of empathy (artificial empathy) inside robots seems very hard because the definition of empathy itself is ambiguous, and therefore, measuring how much robots empathize with humans is difficult to

[1] https://www.nytimes.com/2006/01/10/science/cells-that-read-minds.html

evaluate. Also, even before that, estimation of humans' emotion is another difficulty. In [9], Asada mentioned that:

> The narrow definition of empathy is simply the ability to form an embodied representation of another's emotional state while, at the same time, being aware of the causal mechanism that induced that emotional state [10]. This suggests that the empathizer has interoceptive awareness of his or her own bodily states and is able to distinguish between the self and other, which is a key aspect of the definitions of empathy-related terms from an evolutionary perspective,

and those terms are organized in order from the evolutionary and developmental perspectives. The bottom-right of Figure 8 shows a conceptual model of empathy development ([9]), and the rest of the figure indicates the related studies, which are briefly introduced in the following sections. According to the evolutionary process of empathy proposed by de Waal [11], upper rectangles of empathy-related terms, such as emotional contagion, emotional empathy, and cognitive empathy, and lower ones related to imitation are positioned in parallel with the developmental process for self/other discrimination. The vertical ellipses indicate mental functions needed for empathy to emerge. In the case of emotional contagion, "MNS" is an assumption and "self-awareness" is an acquired mental function, and another assumption for "emotional empathy" to emerge. The numbers indicate the internal stage of the developmental process of empathy as follows:

1. No discrimination of self/others
2. Self/non-self discrimination
3. Self-awareness
4. Complete self/others discrimination
5. Metacognition of self as others
6. Emotion regulation of self as others
7. In-group/out-group emotion control

More details of the process and related arguments can be found in [9].

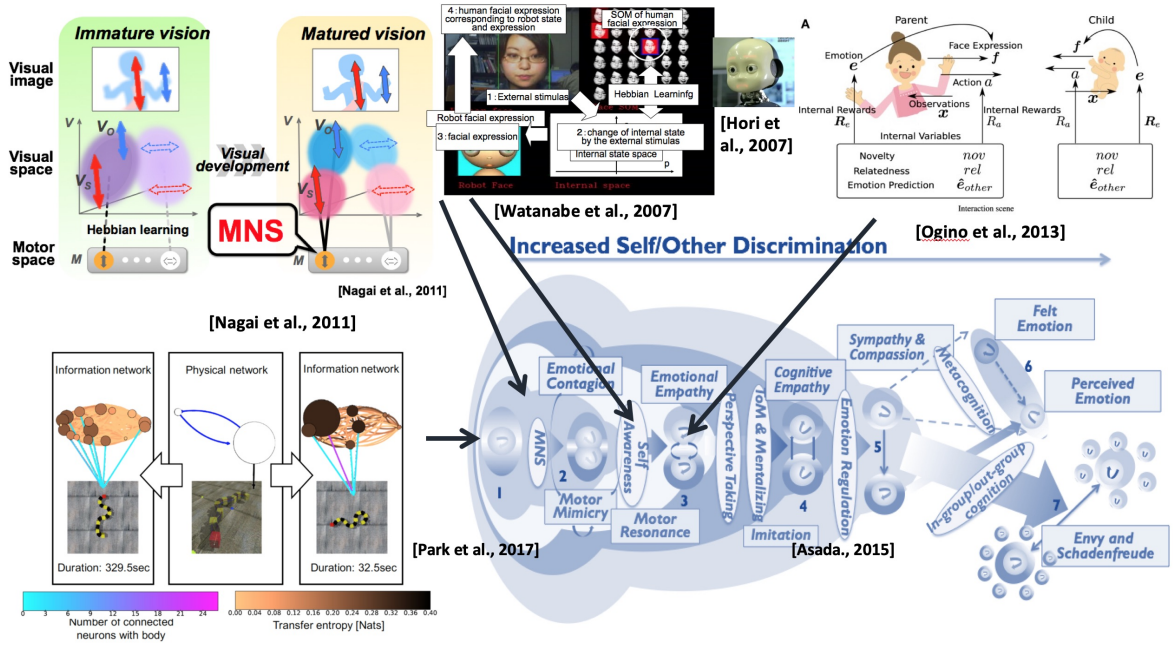

Figure 8. A conceptual model of empathy development and related studies.

3.2.1. Dynamic Coupling Between Body and Brain with Neural Oscillator Networks

Park et al. [12] showed that dynamic coupling between body and brain with neural oscillator networks generated two kinds of subnetwork structures, apart from the anatomical network one—the former consists of many small subnetworks loosely connected to each other and corresponds to a stable motion, while the latter mainly consists of one big sub-network strongly connected with sensory-motor neurons and corresponds to an unstable motion that connects the stable motions. The bottom-left of Figure 8 indicates the summary of this work, and the following two points are essential for the main topic of the paper.

1. Two kinds of motions and network structures behind may correspond to very primitive levels of unconscious (stable motion) and conscious (unstable motion) states, respectively. More plausibly, stable motions could be attractors, and unstable motions appear to transit between the attractors in the phase space.
2. The separation of two kinds of subnetworks can be regarded as functional differentiation that is a basic mechanism for the emergence of new functions [13].

3.2.2. Emergence of MNS and Emotion Sharing

Nagai et al. [14] proposed a computational model for the emergence of an MNS based on the hypothesis that the immature vision leads to self–other correspondence. At the beginning, infants (robots) cannot discriminate between self motion and that of others' due to their immature vision. Gradually, they become able to discriminate, owing to their visual development. However, early connections between action observation and action execution are left unchanged. As a result, the observation of both self-induced and others' motion evoke the motor system—that is, a function of the MNS.

Such mirroring could be expanded from action to emotions, that is, emotion sharing. Watanabe et al. [15] showed a computational model for emotion development based on a psychological finding, intuitive parenting. Starting from a very simple emotional space consisting of only two pleasure–displeasure states, an infant (robot) gradually evolved its emotional space into a richer one with happiness, surprise, anger, and so on through the interactions with its caregiver. Hori et al. [16] proposed a unified model to estimate the emotional states of others and to generate emotional self-expressions by using a multimodal restricted Boltzmann machine (RBM). Ogino et al. [17] presented a motivation model of infant–caregiver interactions focusing on relatedness, one of the most important basic psychological needs that increases with experiences of emotion sharing. These three studies are positioned in Figure 8.

3.2.3. Sharing Painful Situations Induces Sympathetic Behavior

Figure 9 shows a block diagram of the process for the sharing of painful experiences after the early stage of self-pain experiences. A block for the development of an MNS and related arrows have been added to Figure 5. The MNS enables the system to feel the pain of others through the affective and motivational part of the pain matrix—that is, the amygdala, ACC, and anterior insular, which form the emotional state of the self and estimate the one for others. Based on the episodic memories of the painful situations of others through vision and other modalities and the estimated emotional state of others, pain relief behavior is learned to avoid or reduce others' pain.

In the case of the emotion sharing of pain, the system needs to transmit two kinds of information—the sensory discrimination of pain (location, intensity, and quality), and the affective and motivational responses to pain, as described in Section 2. The former information comes from the sensory system of the body, while the second one comes from its own experiences of pain. The ideal hypothesis is as follows:

1. The information for sensory discrimination of pain (location, intensity, and quality) is transmitted to the central nervous system from the sensory system.

2. If the above experience is new, the related information, such as cause and/or reason, is also transmitted with the information above.
3. Else, the memory of this experience is enhanced in the memory storage.
4. When the painful situations of others are observed, emotion sharing of pain happens, and also the memory of the similar experience is recalled.
5. Take actions to reduce the pain of others based on the recalled experience.

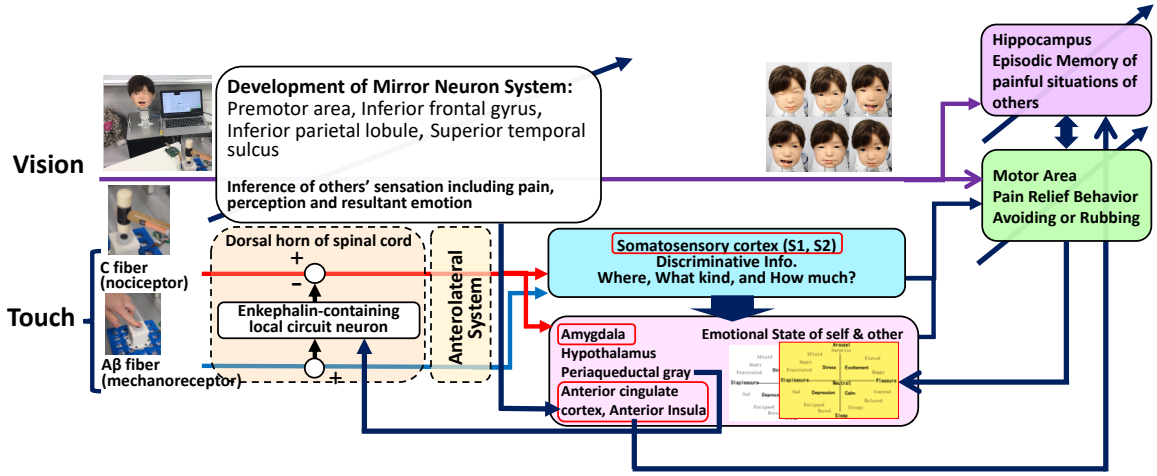

Figure 9. The sharing of pain experiences.

A robot could be a moral agent if it can generate such behavior successfully. At the same time, such a robot may have the right to receive moral behavior from others. Such moral agency could be a solution to the first law of the three laws of robotics[2]. That is, "A robot may not injure a human being or, through inaction, allow a human being to come to harm."

4. Discussion

To tackle the issue of consciousness, this study attempted to represent it as a phenomenon of the developmental process of artificial empathy for pain and moral behavior generation. The conceptual model for the former is given by [9], while the latter is now a story of fantasy. If a robot is regarded as a moral being that is capable of exhibiting moral behavior with others, is it deserving of receiving moral behavior from them? If so, can we agree that such robots have conscious minds? This is an issue of ethics towards robots, and is also related to the legal system. Can we ask such robots to accept a sort of responsibility for any accident they commit? If so, how? These issues arise when we introduce robots who are qualified as a moral being with conscious minds into our society.

Before these issues can be considered, there are so many technical issues to address. Among them, the following should be addressed intensively.

1. Associate the sensory discrimination of pain with the affective and motivational responses to pain (the construction of the pain matrix and memory dynamics).
2. Recall the experience when a painful situation of others is observed.
3. Generate appropriate behavior to reduce the pain.

Funding: This research was funded by JST Strategic Basic Research Programs (RISTEX), Research Area "Human-Information Technology Ecosystem," entitled "Legal Beings: Electronic personhoods of artificial intelligence and robots in NAJIMI society, based on a reconsideration of the concept of autonomy" (JPMJRX17H4, October 2017–September 2020).

[2] https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

Conflicts of Interest: The author declares no conflict of interest.

References

1. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016. Available online: http://www.deeplearningbook.org (accessed on 5 July 2019).
2. Asada, M.; Hosoda, K.; Kuniyoshi, Y.; Ishiguro, H.; Inui, T.; Yoshikawa, Y.; Ogino, M.; Yoshida, C. Cognitive developmental robotics: A survey. *IEEE Trans. Auton. Ment. Dev.* **2009**, *1*, 12–34.
3. Tani, J. *Exploring Robotic Minds: Actions, Symbols, and Consciousness as Self-Organizing Dynamic Phenomena*; Oxford University Press: Oxford, UK, 2016.
4. Purves, D.; Augustine, G.A.; Fitzpatrick, D.; Hall, W.C.; LaMantia, A.S.; McNamara, J.O.; White, L.E. (Eds.) *Neuroscience*, 5th ed.; Sinauer Associates, Inc.: Sunderland, MA, USA, 2012.
5. Melzack, R.; Wall, P.D. Pain Mechanisms: A New Theory. *Science* **1965**, *150*, 971–979. Available online: https://science.sciencemag.org/content/150/3699/971.full.pdf (accessed on 5 July 2019), doi:10.1126/science.150.3699.971
6. Kuehn, J.; Haddadin, S. An Artificial Robot Nervous System to Teach Robots How to Feel Pain and Reflexively React to Potentially Damaging Contacts. *IEEE Robot. Autom. Lett.* **2016**, *2*, 72–79, doi:10.1109/LRA.2016.2536360.
7. Kawasetsu, T.; Horii, T.; Ishihara, H.; Asada, M. Flexible Tri-axis Tactile Sensor Using Spiral Inductor and Magnetorheological Elastomer. *IEEE Sens. J.* **2018**, *18*, 5834–5841.
8. Singer, T.; Seymour, B.; O'Doherty, J.; Kaube, H.; Dolan, R.J.; Frith, C.D. Empathy for pain involves the affective but not sensory components of pain. *Science* **2004**, *303*, 1157–1162.
9. Asada, M. Towards Artificial Empathy. *Int. J. Soc. Robot.* **2015**, *7*, 19–33.
10. Gonzalez-Liencresa, C.; Shamay-Tsooryc, S.G.; Brünea, M. Towards a neuroscience of empathy: Ontogeny, phylogeny, brain mechanisms, context and psychopathology. *Neurosci. Biobehav. Rev.* **2013**, *37*, 1537–1548.
11. De Waal, F.B. Putting the Altruism Back into Altruism: The Evolution of Empathy. *Annu. Rev. Psychol.* **2008**, *59*, 279–300.
12. Park, J.; Mori, H.; Okuyama, Y.; Asada, M. Chaotic itinerancy within the coupled dynamics between a physical body and neural oscillator networks. *PLoS ONE* **2017**, *12*, 618–628.
13. Yamaguti, Y.; Tsuda, I. Mathematical modeling for evolution of heterogeneous modules in the brain. *Neural Netw.* **2015**, *62*, 3–10.
14. Nagai, Y.; Kawai, Y.; Asada, M. Emergence of Mirror Neuron System: Immature vision leads to self-other correspondence. In Proceedings of the IEEE International Conference on Development and Learning, and Epigenetic Robotics (ICDL-EpiRob 2011), Frankfurt am Main, Germany, 24–27 August 2011; (CD–ROM).
15. Watanabe, A.; Ogino, M.; Asada, M. Mapping Facial Expression to Internal States Based on Intuitive Parenting. *J. Robot. Mechatron.* **2007**, *19*, 315–323.
16. Horii, T.; Nagai, Y.; Asada, M. Imitation of human expressions based on emotion estimation by mental simulation. *Paladyn J. Behav. Robot.* **2016**, *7*, 40–54.
17. Ogino, M.; Nishikawa, A.; Asada, M. A motivation model for interaction between parent and child based on the need for relatedness. *Front. Psychol.* **2013**, *4*, 324–334.