MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Self-observation Principle for Estimating the Other's Internal State

– New Computational Theory of Communication –

Takaki Makino
and
Kazuyuki Aihara

METR 2003–36 October 2003

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.i.u-tokyo.ac.jp/mi/mi-e.htm

Self-observation Principle for Estimating the Other's Internal State – New Computational Theory of Communication –

Makino, Takaki* Aihara, Kazuyuki*
mak@sat.t.u-tokyo.ac.jp aihara@sat.t.u-tokyo.ac.jp
Department of Complexity Science and Engineering,
Graduate School of Frontier Science, Tokyo University

October 24, 2003

Abstract

We propose a computational theory of internal-state estimation for others, which is the basis of information processing in human communication. To estimate internal states of the other equivalent to the self, we have to deal with two substantial difficulties, restriction of the estimator's parameter dimension and conversion between objective and subjective information. The proposed computational theory that solves both difficulties is based on *self-observation principle*. Learning the dynamics of the self provides prior knowledge of the dynamics of the other, which reduces the restriction of the parameter dimension; learning the association between the subjective state for the self and the objective observation of the self provides a mechanism for conversion between objective observation of the other and subjective information to the other. In this paper, we formalize communication in a framework of dynamics-estimation problems, and describe the two difficulties and our proposal on the framework. We also discuss relations of our proposal to evolutional psychology and neuroscience.
Keywords: communication, computational theory, model estimation, mirror neuron, self-observation

*This research is partially supported by a Grant-in-Aid No. 15016023 for Scientific Research on Priority Areas (2) Advanced Brain Science Project, and the Advanced and Innovational Research Program in Life Science, from the Ministry of Education, Culture, Sports, Science, and Technology, the Japanese Government. This research is also partially supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

1 Introduction

The mechanism of human communication is an outstanding issue of information science. Not only human beings, but monkeys and dogs do their own communication each other. In advanced communication, new movements emerge, such as tactics, cooperation and cultures. However, modern computers cannot do that. Only in a programmed manner, they communicate with other computers, let alone with humans.

Recent development of technology are overcoming technical issues, such as the lack of "humanity" in computers, which have been said to prevent computers from communication. Rapid improvements are seen in synthesis and analysis of face expression, generation and recognition of vocal dialogue, and so on. Some latest studies oriented to human communication uses a body of humanoid robot [IOIK03, AHP$^+$00]. Nevertheless, we still have found no way to provide advanced communication that causes emergence of new movements.

We can expect for improvement of the brain science to reveal information-processing mechanism of communication in the brain. However, without a breakthrough theory, the brain science will not show much improvement. Even if we investigate all the synaptic connections within the brain, we cannot see how the brain performs communication. Experiments in cognitive science and neuroscience can only look for evidence of existing theories; they cannot elucidate a wholly-unknown mechanism.

The required is a *computational theory* that underlies communication. A computational theory is a basic framework for the research of brain information-processing, which states the purpose and requirements of communication and the input/output of the information processing for the communication [Mar82, Kaw96]. Based on a computational theory, we can propose several algorithms that satisfy the purpose and requirements. Then we can use the proposed algorithms for implementation on computers, as well as for hypotheses required for the research in the brain science.

An existing computational theory of communication states that communication consists of interactions through the estimation of other's internal states [Kuj97]. However, up to our knowledge, nobody has found an algorithm that satisfies the theory. This fact suggests some incompleteness of the theory, in other words, the theory may lack some key factors that are essential to solve difficulties in actual algorithms. If this is the case, we need a new computational theory, which contains the key factors in addition to the original theory.

In this article, we investigate the existing computational theory of communication within the framework of dynamics-estimation problem, and point out that the theory are prevented from realization by two difficulties, the limit of parameter dimensions of the estimator and the conversion between subjective information and objective information. As a way to solve the dif-

ficulties, we propose a new computational theory based on *self-observation principle*, in which the dynamics model, which is learned through objective observation of the self, is applied to the other for estimation of the other's internal states. We also discuss the proposed theory within the scope of related scientific domains, and argue the self-learning dynamics model has relation to the human self-consciousness and mirror neurons.

2 Communication as dynamics estimation

In this section, we formalize the computational theory of communication in a framework of dynamics-estimation problem.

2.1 Existing computational theory of communication

In this section, we provide our definition of communication, to avoid ambiguity of the meaning. We also introduce an existing computational theory of communication.

In this article, we focus on communication, which is "an information-based interaction between fellows of the same kind, which have been developed for surviving natural selection". We exclude actions which have no direct benefit for survival or reproduction, such as simple observation and imitation. This is because we have no way to verify algorithms that perform these actions. We also exclude automatic actions based on instincts because our focus is the mechanism of information processing in the brain.

Existing studies have shown that the estimation of the other's internal state is a prominent information processing involved in the communication [Kuj97]. Here we define the internal state as a set of elements, which is not observable from one's external world but affects his action, e.g. emotion, desire, intention, and knowledge.

If one's colleagues helps him, he can survive better than he alone does. To promote the help, he needs to appeal to his colleagues to take more actions for his benefit; this is the purpose of the communication. However, to do this, he must know which appealing cause the colleagues to help him the appealing requires. In other words, prior to the appealing actions, he needs to predict the colleagues' action under various condition. Such an action prediction is required in the information processing of the communication. Furthermore, since the colleague's action is hardly predictable from observable information, he needs to estimate the colleagues' internal states, prior to the action prediction.

For example, it may be difficult to predict one's future action based only on his externally-observable state (a skinny man is walking unsteadily). If, however, we are able to estimate his internal states (hungry and looking for food), the estimation helps us to predict his future action (when he finds some food, he will eat it). Such prediction helps our actions for survival (we

can hide our food) as wells as interactions that changes his internal states (giving food to him will cure his hunger and make him happy).

In advanced communications, internal states are estimated recursively, i.e. "estimating 'how he estimates my internal state' as a part of his internal state". For example, chimpanzees show deceiving actions [PW78]; a chimpanzee pretended to see distance as if some danger is approaching, expecting for the action to distract another chimpanzee's attention from scolding him. Recursive estimation, which underlies such an action, also plays a crucial role in human communication.

Despite its importance, it is largely unknown how the human brain performs estimation of the other's internal states. One reason is the lack of computational studies. For any information processing of a human, it is difficult to find its correspondence in the brain, if we don't know what kinds of information representations and algorithms can provide the process. Computational studies try to construct the representations and algorithms, so that we can use them as hypotheses of the mechanism in the brain.

To study in this methodology, we need a computational theory, which is a common basis of the algorithms, such as the purpose of the calculation and the structure of input/output. In this study, we define the computational theory of communication as the following.

---

1. Improve the actor's possibility of survival and reproduction through action selection. (This includes the actor's selection of action that causes others' action beneficial to the actor).

2. Estimate actions of the other, who is an equivalent to the self, for achieving (1).

3. Estimate the other's internal states and relation between internal states and environments/actions, for achieving (2).

The remaining of this article pursues a principle of algorithms that conforms this computational theory.

2.2 Dynamics-estimation framework

The estimation of unobservable parameters involves the estimation of several components. We present the framework of dynamics-estimation problem to identify the components.

Let $\boldsymbol{x}(t) \in S$ be a state of a target at time t. The dynamics of the target, the structure of the state change according to the environment and time, can be denoted as the following, using the input $\boldsymbol{a}(t)$ of time t and a state-transition map f.

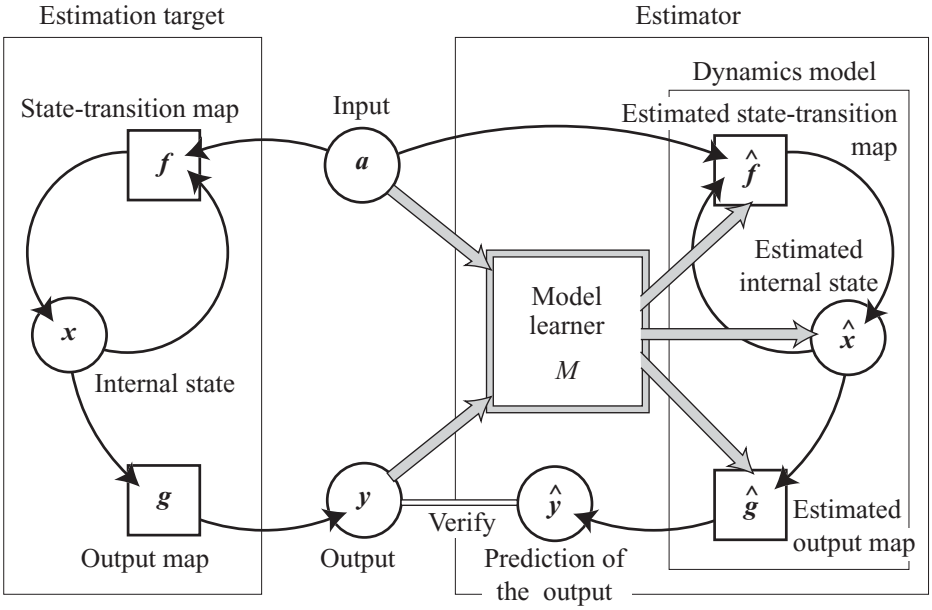


Figure 1: Dynamics-estimation problem

$$\boldsymbol{x}(t+1) = f(\boldsymbol{x}(t), \boldsymbol{a}(t)) \tag{1}$$

However, we have no way to observe the current state $\boldsymbol{x}(t)$ directly. Instead, we can observe the output $\boldsymbol{y}(t)$, which is determined from $\boldsymbol{x}(t)$ and an output map g.

$$\boldsymbol{y}(t) = g(\boldsymbol{x}(t)) \tag{2}$$

Suppose we want to predict the future output of the target, $\boldsymbol{y}(t+1), \boldsymbol{y}(t+2), \cdots$ using only inputs $\boldsymbol{a}(1), \boldsymbol{a}(2), \cdots$ and past outputs $\boldsymbol{y}(1), \boldsymbol{y}(2), \cdots, \boldsymbol{y}(t)$. For better prediction, a predictor needs to calculate the estimation of the current state $\hat{\boldsymbol{x}}(t) \in \hat{S}$ as well as the estimation of the target dynamics, $\hat{f}$ and $\hat{g}$. This is a dynamics-estimation problem (Figure 1).

If the target dynamics (f and g) are known to the estimator, the problem is reduced to the estimation of the current state, which is generally easy. In a case either f or g is only known, the predictor has to estimate the other map, which makes the problem more difficult.

However, the problem gets extremely difficult if both f and g are unknown. Since there is no clue for the original state space S, the estimator has to provide a reconstruction of the state space $\hat{S}$, in which a state estimation $\hat{\boldsymbol{x}}$ is placed. Generally, in a deterministic system with constant input, the reconstruction is not guaranteed unless the dimension of $\hat{S}$ is larger than the twice of the dimension of the original state space S (embedding theorem

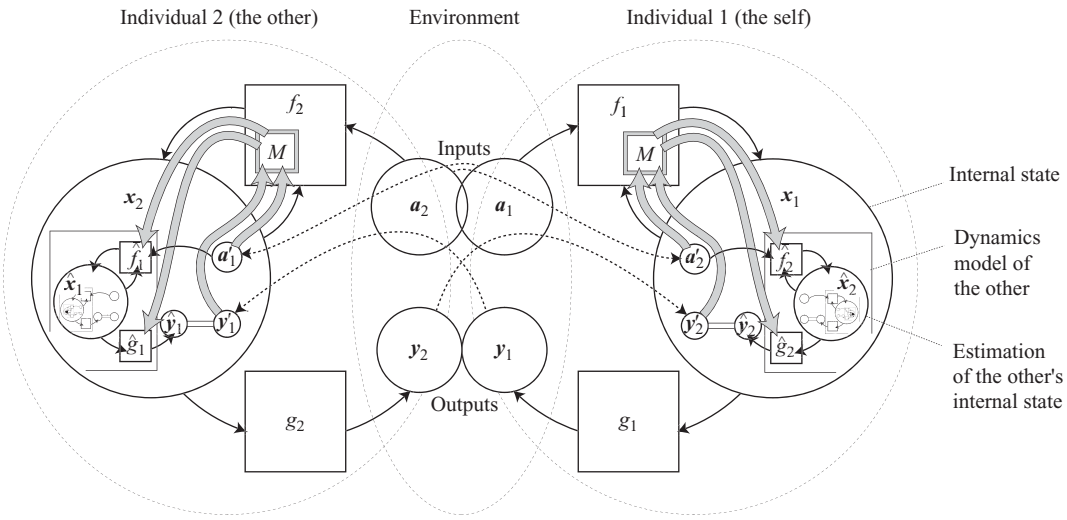


Figure 2: Communication as a dynamics estimation

[AIYK00]). The increase of the dimension of $\hat{S}$ multiplies the number of parameters of the estimated maps $\hat{f}$ and $\hat{g}$, and as a result, the estimation becomes less feasible. Conversely, partial information on f or g may help the estimation through the reduction of the dimension of $\hat{S}$.

2.3 Communication as dynamics estimation

We can regard the estimation process of the other's internal states as an extension of the dynamics-estimation framework to a non-stationary system. Although the communicating two are equivalent, we focus one of them as a dynamics estimator, which we call as *the self*. We call the target of the self's estimation as *the other*.

Figure 2 illustrates this framework. In this case, the self is an estimator of the other's internal dynamics as well as a dynamical system being estimated by the other. To represent this relation, the other's dynamics model, $\hat{f}_2$ and $\hat{g}_2$, should be a part of the self's internal state, $\boldsymbol{x}_1$, and be constructed by the model learner M in the state-transition map f_1. Since the other also constructs the dynamics model of the self, the whole framework becomes symmetric. We can also see fractal structure if the self estimates the dynamics model of the self in the other's internal state.

Note that the self cannot obtain the other's precise input $\boldsymbol{a}_2$ and output $\boldsymbol{y}_2$. The self can obtain them only in an indirect way, through the input to the self $\boldsymbol{a}_1$. This problem is discussed later at Section 4. Until then, we temporarily use the symbols with primes $\boldsymbol{a}'_2$, $\boldsymbol{y}'_2$ to denote indirect and partial information.

3 Difficulties on parameter dimension

In this section, we describe one of the two difficulties involved in the estimation of the other's internal states; the difficulty on the estimator's parameter dimension.

3.1 Limit of parameter dimension

In the previous section, we have shown that the estimation of the other's internal states can be regarded as a dynamics-estimation problem.

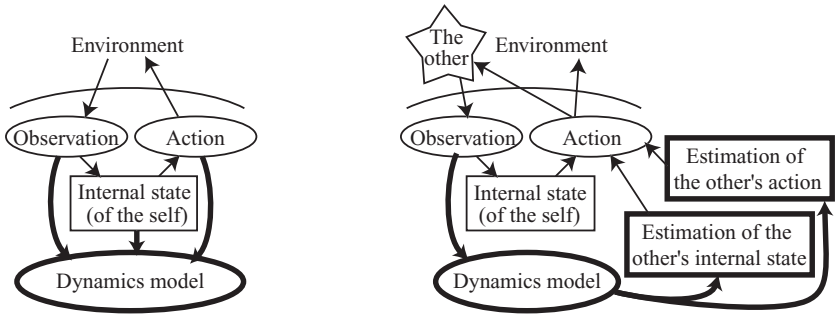
However, the brain of animals has been performing dynamics estimation from ancient days, irrelevant of communication. Feedback control of muscle movement and construction of environment map are good examples of prediction of state change in the external world. In such prediction, the brain estimates non-observable states and its transition/output maps to resolve complex relation between input and output. Researchers are trying to reproduce this estimation in a computational model [WK98].

Then, can the estimation of the other's internal states, the basis of communication, be realized an extended, complicated version of this dynamics estimation? It seems that, in the area of computational communication study, there exists an implicit consensus that the answer is 'yes'.

We, however, claim that it is 'no'. This is because the other, who is equivalently complex to the self, is too complex to be estimated. An individual, such as a human being, has a large dimension of internal states and complex dynamics. Moreover, an estimator needs extra dimensions for the reconstruction of the state space. It is very difficult to estimate all of the target's dynamics, the state-transition map $\hat{f}_2$, the output map $\hat{g}_2$, and the current state $\hat{\boldsymbol{x}}_2$, using solely the external observation of the target.

Theoretically, it is still possible to estimate the dynamics with a mass of observation data. However, the estimator's parameter dimension needs to be far larger than the dimension of the target's internal state. In other words, if the estimator's parameter dimension is fixed, he can estimate only a very simple target, whose dimension is substantially lower than the estimator; a more complex target beyond the 'limit of the dimension' cannot be estimated with sufficient precision. To provide advanced estimation, such as recursive estimation, it is necessary to estimate the internal state of the target, who is equivalently complex to the estimator. It is apparent that the dimension of such a target goes beyond the 'limit of the dimension' and that the target cannot be estimated solely by external observation.

Even a human being has a very limited capacity for the prediction problem involving dynamics estimation of hidden parameters. It is very unlikely that the capacity beyond human being is mandatory for communication.



(a) Acquire a dynamics model from the self's own dynamics

(b) Use the acquired dynamics model to estimate the other's internal states and actions

Figure 3: Solution using the self

3.2 Clue for estimating dynamics

The described difficulty is caused by the restriction that all the dynamics is estimated only from observation. As discussed in Section 2.2, we may be able to solve this difficulty if we have some partial information of dynamics, such as state space, transition map and output map.

One naive idea is that the partial information is innately given to the brain. That is, the human brain has *a priori* structure of basic information processing that extracts one's internal states from its external observation. It is probable that the brain has innate knowledge and reactions for primitive elements, including face expression and tone of voice.

However, the idea seems to be insufficient for describing communication. A human being learns a large part of knowledge, including knowledge about himself, after his birth. It is unlikely that the human being innately possesses all knowledge of the other, and complex processing like recursive estimation. In fact, psychologists have evidence of the age-dependent development of 'the theory of mind' in human children [BC95].

We need some description of learning the dynamics. In other words, the brain needs some source of learning data, in addition to the observation of others.

3.3 Solution using the self

Our proposal is that one's brain learns the other's dynamics from his own dynamics (Figure 3). That is, the self adopts an assumption that the other obeys similar dynamics to the self's own one. If the assumption is true, the brain can obtain rough estimation of the other's internal states. Small difference can be learned after that.

The followings are some advantages of this proposal.

- The proposal is compliant to the computational theory of communication, that is, the estimation of internal state of the other, who is equivalent to the self. Under the theory, the above assumption is fulfilled because the other's dynamics is equivalent to the self's. It can be seen as the most natural realization of the constraint of equivalency.

- The self can obtain much information for the self's dynamics than the other's. The self can observe his own internal states, such as emotion and desire, as well as his externally non-observable input, such as the sense of pain and touch. Using a dynamics model constructed from this abundant information, it is much easier to estimate internal states of the other.

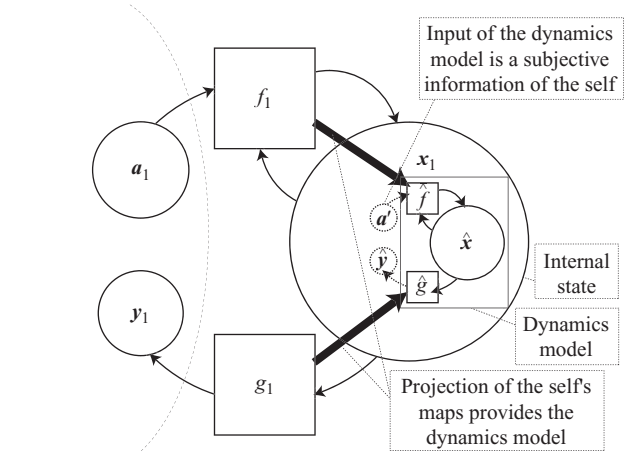
- Increase of state space dimension can be suppressed. The brain need not reconstruct internal state space from scratch if it uses information of internal states and transition maps in the self. In addition, the knowledge of the structure of internal state space makes the learning of output map much easier.

The following example illustrates this idea in a scene of human communication. A person (the self) sees another person (the other) accidentally bumping his hand to a table and bleeding. The self has already learned, from his own experience, associations between the input and the state change (bumping hand causes pain), and associations between the observation and the internal state (bleeding hand is painful). Using the learned knowledge, the self estimates the internal state of the other from the observation of the other and his environment (He bumped his hand and is bleeding, so he would be painful). Combined with action decision, the self can work on the other (apply ointment to the wound), and change the internal state (relieve the pain and make him happy).

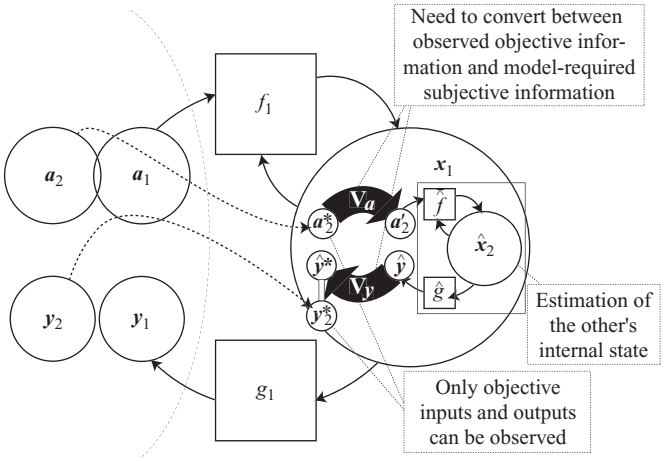
Note that one's dynamics model does not necessarily reflect himself perfectly. It is known that the framework of observing/learning self involves substantial difficulty for representing the complexity of the self within itself [Rös98]. However, in this study, the self-learning is used just for providing a 'seed' of dynamics model for the estimation of the other's internal state. Thus the adequate level of approximated dynamics model of the self's dynamics is sufficient for that purpose.

4 Difficulties on conversion between subjective and objective information

This section describes another difficulty, the conversion between subjective and objective information.. The other's internal state cannot be estimated



(a) Self-projection stage



(b) Self-applying stage

Figure 4: Self-application principle

merely by using the dynamics of the self, because the input of the dynamics, the other's subjective information, needs to be reconstructed from the observed information of the self. However, the reconstruction of subjective input is in the same class of difficulty as the estimation of internal state is. We argue that a naive self-application principle cannot escape from the necessity of subjective input reconstruction, and propose a solution, a self-observation principle.

4.1 Self-application principle

One method, which naively realizes the proposal of the previous section, is to apply the action rule of oneself for the prediction of the other's action.

We call this method as *self-application principle*.

Figure 4 illustrates this principle. One's brain projects (copies) his own state-transition map f_1 and output map g_1 to the corresponding maps in the dynamics model, $\hat{f}$ and $\hat{g}$, respectively. After that, the brain uses the dynamics model to obtain an estimation of the other's internal state $\hat{\boldsymbol{x}}$ and a prediction of the other's output $\hat{\boldsymbol{y}}$.

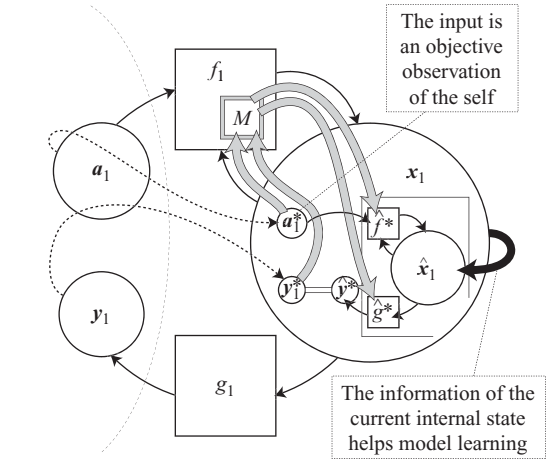
The problem of this principle lies in the difference of stances between the self and the other. One can only observe objective information of the other; we denote such objective information by attached asterisk, $\boldsymbol{a}_2^*$. This information is incompatible to the input of the map $\hat{f}$, which is copied from the state-transition map of the self. The compatible, required input $\boldsymbol{a}_2'$ is the input the self would receive if the self were on the stance of the other, that is, information of subjective input. To bridge this difference, the brain requires a conversion process V_a, which converts $\boldsymbol{a}_2^*$ into $\boldsymbol{a}_2'$.

The same applies for the output. Although the predicted output of the dynamics model $\hat{\boldsymbol{y}}_2'$ is subjective, we need to know how that matters to the self, i.e. the objective information of the output $\hat{\boldsymbol{y}}_2^*$. Thus a conversion process V_y, which converts $\hat{\boldsymbol{y}}_2'$ to $\hat{\boldsymbol{y}}_2^*$, is required.

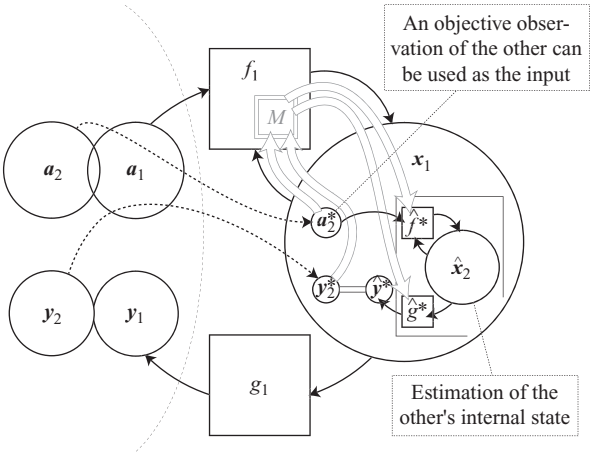
We claim that this conversion process between objective and subjective information is the second difficulty in the computational theory of communication. Contrary to our intuition, the conversion is not a trivial process, because the two types of information representation are in totally different modes. As for the sense of vision, the other's visual sight needs to be obtained from the self's sight through a complex calculation of viewpoint conversion. Almost all the sense of touch and pain needs to be reconstructed from other senses, such as visual and auditory sense.

For example, suppose that the self saw the other accidentally bumped his hand against a table. If the self puts himself in the position of the other, the self can imagine the bump causes surprise in his emotion. In the self-application principle, the rule is applied to the other to estimate the other's surprise. However, to do that, the self has to be able to convert the self's sight into the other's subjective sight of the table and the hand, and reconstruct the sense of pain from the sight; otherwise, he cannot put himself in the position of the other. Such a complex conversion and reconstruction is unlikely to be acquired from only the observation of the others, let alone to be innate knowledge of human beings.

In other words, the estimation of the other's subjective input is as difficult as the estimation of the other's internal state. The other's internal state is private, and the other's subjective input is also private; knowledge of private process, $\hat{f}$ and $\hat{g}$, doesn't create a way to reach privates from public information. We claim that the self-application principle does nothing more than displacing the problem of internal-state estimation into the problem of subjectivity-objectivity conversion, and does not solve the problem.



(a) Model-learning stage



(b) Model-applying stage

Figure 5: Self-observation principle

4.2 Self-observation method

In this study, we propose a *self-observation* principle, which resolves this problem. This is done by constructing a dynamics model from the objective observation of the self, and then applying the dynamics model for the other's estimation of his action.

Figure 5 illustrates this principle. The state of the self x contains a dynamics model $\hat{f}^*$ and $\hat{g}^*$. The dynamics model is different from that of the self-application principle in a point that the dynamics model directly handles objectively observed information.

In the model-learning stage (Figure 5 (a)), a model learner M learns dynamics of $\boldsymbol{a}_1^*$ and $\boldsymbol{y}_1^*$, which are the objective observation of the input/output

of the self. This learning is easy because the actual internal state of the self $\boldsymbol{x}_1$ can be used as a teacher data for the internal state of the model $\hat{\boldsymbol{x}}$.

Once the self has learned the dynamics model $(\hat{f}^*, \hat{g}^*)$, the self can apply the model to the observation of the other (Figure 5 (b)). The dynamics model processes $\boldsymbol{a}_2^*$, the objective observation of the other's input, and produces $\hat{\boldsymbol{x}}$, the estimation of the other's internal state, and $\boldsymbol{y}_2^*$, the prediction of the objective observation of the other's output.

For example, the self firstly learns the fact "Bumping hand causes pain" from the self's own experience. The self's dynamics model learns objective observation of the experience, e.g. the sight of hand movement, bumping sound, allo-centric (objective space) arrangements of the table and the hand, and their relation to the pain in the self's internal state. After that, observing the other bumping his hand to a table, the self estimates the other feels pain from the observation, such as the sight of the other's hand movement, bumping sound, and allocentric locations of the table and the other's hand.

Note that these observations are almost useless in determining the self's action. To avoid bumping hand again, the self only needs to learn relations among ego-centric (subjective space) arrangements of the table, muscle movements of arm, and the resulting pain. However, such subjective information is not enough for estimation of the other's internal states; to do that, the self need to relate objective observation to the self's subjective information and internal state.

4.3 Discussion

The proposed self-observation principle, which is apparently complicated, is actually simpler than the self-application principle. One reason is the reduction of the elements to be calculated due to the elimination of conversion between subjectivity and objectivity, V_a and V_y.

Moreover, the idea underlying the self-application principle, 'applying the self's dynamics to the other' is implicitly involved in the self-observation principle. The mechanism that transcripts the self's dynamics into dynamics model is provided as model learner M in the self-observation principle, unlike simple projection in the self-application principle. In this viewpoint, the self-observation principle can be regarded as a variant of the self-application principle, which solves the problem of reconstruction of subjective information.

Another advantage of the self-observation principle is its ability to learn the other's dynamics, using the same model learner M as the self's dynamics. As for the self-application principle, it is not a straightforward work to design the learner; when some prediction goes wrong, the learner needs to choose whether map, V_a or $\hat{f}$, is to be corrected. That is not the case in the self-observation principle, where the two maps are united into $\hat{f}^*$.

4.4 New computational theory

The proposed self-observation principle solves the two difficulties in the estimation of the other's internal states, the limit of the estimator's parameter dimension and the conversion between subjective and objective information. However, the principle is so general that a number of algorithms can fulfill the principle. Of course, they have to solve many practical difficulties.

To promote studies of these algorithms, we can integrate the proposed principle into a new computational theory. Although the existing computational theory (Section 2.1) is so incomplete that we can hardly find an algorithm that solves the difficulties, the new computational theory, which clarifies the difficulties and approaches, will enable us to study new algorithms. Then, the new algorithms can lead us to develop a new artificial intelligence as well as help us hypothesize the function of the brain.

Thus, we here propose a new computational theory, which integrates the self-observation principle. The theory is the following, in addition to the original theory shown in Section 2.1:

---

4. Construct a dynamics model, which is applicable to the other, through objective observation of the actor himself, for achieving (3) (estimation of the other's internal states).

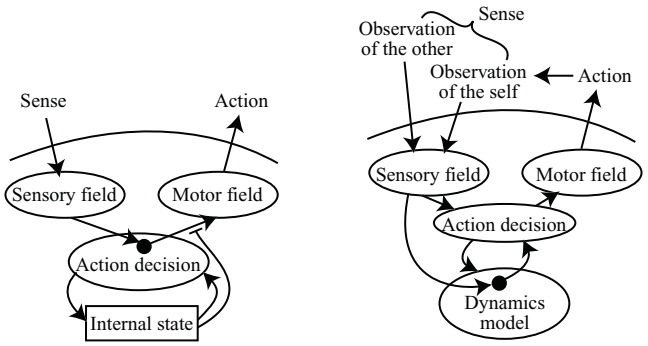
Of course, we do not deny possibilities of an alternative computational theory of communication. There may be another way to solve the limit of estimator's parameter dimension; the original computational theory 'estimation of the other's internal state', which we have used as a basis, may need some modification. We hope that our proposed principle leads us to a deeper discussion of the computational studies of communication, including alternative theories.

5 Related studies

The self-observation principle is very simple but related to various research domains. This section describes the relations.

5.1 Psychology

The idea of "Self-observation for a tool to estimate the other's internal" has been proposed in 1970s by an evolutional psychologist, Nicholas Humphrey [Hum80, Hum84]. However, his thesis was a purely theoretical hypothesis, which describes the origin of the self-consciousness through a Darwinian process of evolution. His specification for which part of the self should be observed was somewhat vague.



(a) Rizzolatti's hypothesis[RA98] (b) Hypothesis derived from the self-observation principle

Figure 6: Comparison of hypotheses on the role of mirror neurons. Black circles denote mirror neurons.

We argue that our study can be regarded as an improvement to Humphrey's theory on these points. Through the formalization of the communication process, we stated the effectiveness of the self-observing principle to solve the difficulties in communication, and the requirement of the observation as creating associations between the self's subjective state and objective observation. Moreover, by defining the principle as a new computational theory, our study can be said to migrate Humphrey's theory into new scientific domains, such as the neuroscience and artificial intelligence.

Baron-Cohen claims that, in his book of the theory of mind based on autism study [BC95], Humphrey's hypothesis corresponds to the self-application principle, which contradicts the dissociation of symptoms in autism patients. After that, he briefly suggests that a different principle based on introspection (which resembles the self-observation principle) will match the dissociation. Although we regard that Humphrey's hypothesis covers both the self-application and self-observation principles, it is interesting to see the study of autism supports the advantage of the self-observation principle over the self-application principle.

5.2 Computational neuroscience

Most communication studies in computational neuroscience do not match with the self-observation principle. Kawato et al. discusses communication model as an extension of their model of dynamics interaction [KDH01], but the discussion lacks the presupposition of equivalence between the self and the other. Samejima et al. [SKDK02] proposed a learning model accelerated with imitation, but their model implicitly assumes identical representation for both subjective and objective information.

To the best of our knowledge, no study constructs a model based on Humphrey's hypothesis or a similar framework. This would be because the requirements and the principles for the estimation of the other's internal states have not been declared as a computational theory. We hope that our proposal promotes the future studies of communication in computational neuroscience.

In a different context from communication, Tani proposes a constructivist approach for the study of the self and consciousness [Tan98]. However, in this study, an individual is placed alone in an environment, and does not consider interactions between individuals, let alone the estimation of the other's internal state.

5.3 Experimental neuroscience

Based on the self-observation principle, we can suggest a new role of *mirror neurons* [DFF$^+$92, GFFR96]. A mirror neuron is a neuron that becomes active in a certain action (e.g. grasping an object) as well as in observing another performing the same action. The mirror neurons are in premotor area of the monkey brain, and it is supposed in the Broca's area of the human brain [RFGF96]. The mirror neurons are said to be related to the ability of communication and language, but the detail is unknown.

Rizzolatti et al. proposed a hypothesis that a mirror neuron serves as a commander of an action as well as the recognizer of the action (Figure 6(a)). However, the mechanism of association is unknown. In addition, they say that the response of the mirror neuron to the observation of the action is usually suppressed, without reasonable discussion of the suppression mechanism.

In contrast, the self-observation principle suggests a role of mirror neuron as show in Figure 6 (b). The process shown in Section 3.3 describes the development of mirror neuron very well. A neuron in the dynamics model learns the self's action, and after that, applied to the other's action; as a result, the neuron begins to act as a mirror neuron. Moreover, such a neuron requires no suppression mechanism because it does not directly trigger the action.

Oztop and Arbib offer a hypothesis that the basic functionality of the grasping mirror system is to elaborate the appropriate feedback for opposition-space-based control of manual grasping of an object. They say that, given this functionality, action understanding in the mirror system may be seen as an *exaptation* gained by generalizing from one's own hand to another's hand. Since this hypothesis matches the self-observation principle, we can suggest that such a feedback mechanism is also the evolutional origin of self-observation mechanism in the human communication.

5.4 Artificial intelligence

Existing strategic algorithms, e.g. chess-playing programs, are based on the prediction of the opponent's action. Most existing algorithms, such as the alpha-beta algorithm [Ish89], assume that the opponent evaluates and selects an action in the same way as the algorithm does. This can be regarded as a sort of the self-application principle. The self-application principle is appropriate for this sort of strategic algorithms because the symmetry of the rule makes it easy to design the viewpoint translator. However, it is possible to design a new strategic algorithm based on self-observation principle; a program based on such an algorithm would be possible to learn the opponent's characteristics through games and adapt the strategy.

6 Conclusion

We investigated the existing computational theory of communication, i.e. estimation of the other's internal states. We pointed out that the estimation is prevented from two difficulties, limit of estimator's parameter dimension and reconstruction of subjective information for others. Our proposal for solving the difficulties is the *self-observation principle*: one observes himself objectively to learn a dynamics model, which is then applied to others. Since the dynamics model learns association between the internal state of the self and objective observation of the self, the self can use the model to estimate the internal state of the other from objective observation of the other. Through clarifying the target and purpose of the learning process and integrating the self-observation principle into a computational theory, this study has opened the way for the constructive study of the communication. We also discussed the relation of our proposal to other research domains, including self-consciousness and a mirror neuron system.

References

[AHP+00] C. G. Atkeson, J. Hale, F. Pollick, M. Riley, S. Kotosaka, S. Schaal, T. Shibata, G. Tevatia, S. Vijayakumar, A. Ude, and M. Kawato. Using humanoid robots to study human behavior. *IEEE Intelligent Systems: Special Issue on Humanoid Robotics*, 15:46–56, 2000.

[AIYK00] Kazuyuki Aihara, Tooru Ikeguchi, Taishi Yamada, and Motomasa Komuro. *Fundamentals and Applications of Chaos Time-Series Analysis*. Sangyo Tosho, 2000. In Japanese.

[BC95] S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press, 1995.

[DFF+92] G. Dipellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti. Understanding motor events - a neurophysiological study. *Experimental Brain Research*, 91:176–180, 1992.

[GFFR96] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119:593–609, 1996.

[Hum80] Nicholas Humphrey. Nature's psychologists. In B. Josephson and V. Ramachandran, editors, *Consciousness and the Physical World*, pages 57–75. Pergamon, Oxford, 1980.

[Hum84] Nicholas Humphrey. *The inner eye: Social intelligence in evolution.* Faber and Faber, 1984.

[IOIK03] H. Ishiguro, T. Ono, M. Imai, and T. Kanda. Development of an interactive humanoid robot "Robovie" – an interdisciplinary approach. In R. A. Jarvis and A. Zelinsky, editors, *Robotics Research*, pages 179–191. Springer, 2003.

[Ish89] Kiyoshi Ishihata. *Algorithm and Data Structure: Iwanami Lecture on Software Science 3.* Iwanami Shoten, 1989. In Japanese.

[Kaw96] Mitsuo Kawato. *Computational Theory of the Brain.* Sangyo Tosho, 1996. In Japanese.

[KDH01] Mitsuo Kawato, Kenji Doya, and Masahiko Haruno. Extension of MOSAIC and communication: Computational neuroscience of human intelligence #5. *Kagaku*, 71:197–204,839–843, 2001. In Japanese.

[Kuj97] Takashi Kujiraoka. *Aspects in Primitive Communication.* Minerva Shobo, 1997. In Japanese.

[Mar82] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W. H. Freeman, San Francisco, 1982.

[PW78] D. G. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1:515–526, 1978.

[RA98] Giacomo Rizzolatti and Michael A. Arbib. Language within our grasp. *Trends in Neuroscience*, 21:188–194, 1998.

[RFGF96] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141, 1996.

[Rös98] Otto E. Rössler. *Endophysics: the world as an interface*. World Scientific Publishing, 1998.

[SKDK02] K. Samejima, K. Katagiri, K. Doya, and M. Kawato. Symbolization of action patterns and imitation learning by module competition. *Japanese Transaction of IEICE*, J85-D-II:90–100, 2002. In Japanese.

[Tan98] Jun Tani. An interpretation of the 'self' from the dynamical systems perspective: a constructivist approach. *Journal of Consciousness Studies*, 5(5-6):516–542, 1998.

[WK98] D. M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11:1317–1329, 1998.