Self-observation Principle for Estimating the Other's Internal State[*†] – New Computational Theory on Communication –

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

October 24, 2003

Abstract

We propose a computational theory on estimating the internal states of others, which is the basis of information processing in human communication. To estimate internal states of peers, we have to deal with two considerable difficulties, restricted dimension of estimator parameters and conversion of objective information into subjective. To solve these problems, we propose a new computational theory based on the *self-observation principle*. Learning one's own dynamics provides prior knowledge on the dynamics of others, which reduces restriction of the parameter dimension. On the other hand, learning the association between one's own subjective state and the objective self-observation provides a mechanism for converting objective information obtained from observing others and their subjective information. In this paper, we formalize communication within a framework of dynamics-estimation problems, and explain the two difficulties and our framework. We also discuss the relations our proposal has with evolutional psychology and neuroscience.

Keywords: communication, computational theory, model estimation, mirror neuron, self-observation

1 Introduction

The mechanism behind human communication is still not understood sufficiently well in information science. Not only human beings, but monkeys and dogs communicate with one another. In advanced communication, new strategies emerge, such as tactics, cooperation and cultural elements. However, modern computers cannot communicate. Only in a programmed manner, they communicate with other computers, let alone with humans.

Recent developments in technology are overcoming technical issues, such as the lack of "humanity" in computers, which have prevented computers from communication. There have been rapid improvements in the synthesis and analysis of facial expressions, generation and recognition

[*]This article is a minor revision (based on comments from an English native) of the technical report [MA03]: Makino, T. and Aihara, K. *Self-observation Principle for Estimating the Other's Internal State.* Mathematical Engineering Technical Reports METR 2003-36, the University of Tokyo, October 2003.

[†]This research was 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 of the Japanese Government. This research was also partially supported by Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

of vocalized dialogue, and other areas. Some of the latest studies have been oriented toward human communication through the bodies of humanoid robots [IOIK03, AHP+00]. Nevertheless, we still have found no way of achieving advanced communication that will cause emergence of new strategies.

We expect improvements in brain science will reveal the information-processing mechanism responsive for communication in the brain. However, without a breakthrough theory, even the achievements of brain science will be limited. Even if we investigate all the synaptic connections within the brain, we will still not know how the brain engages in communication. Experiments in cognitive science and neuroscience can only look for evidence in existing theories; they cannot elucidate a wholly-unknown mechanism.

We need a *computational theory* that explains what underlies communication. A computational theory is a basic framework for research on brain information-processing, which states the purpose and requirements of communication and the input/output of information processing for communication [Mar82, Kaw96]. Based on computational theory, we can propose several algorithms that satisfy this purpose and requirements. We can then implement the proposed algorithms on computers, as well as look for evidence of the algorithms in the brain.

As for communication, one existing computational theory states that communication consists of interactions that occur through estimation of other's internal states [KDH01]. However, to the best of our knowledge, no one has yet found an algorithm that conforms to this theory. This suggests some incompleteness in the theory, in other words, the theory may lack some key factors that are essential in solving difficulties within actual algorithms. If this is the case, we need a new computational theory, which contains key factors in addition to the original theory.

In this paper, we investigate the existing computational theory on communication within the framework of dynamics-estimation problem, and point out that it cannot be applied to practice because of two difficulties, limits in the parameter dimensions of the estimator and conversion of objective information to subjective. To solve these problems, we propose a new computational theory based on the *self-observation principle*, where the dynamics model, which is learned through objective observation of the self, is applied to others to estimate their internal states. The first difficulty is reduced by prior knowledge on the dynamics of others provided by learning one's own dynamics, while the second is resolved through learning the association between one's own subjective state and the objective self-observation. We also discuss the proposed theory within the context of related scientific domains, and explain why the self-learning dynamics model is related to human self-consciousness and mirror neurons.

2 Communication in a framework of dynamics estimation

In this section, we formalize the computational theory of communication within the framework of dynamics-estimation problems.

2.1 Existing computational theory of communication

We first define what we mean by communication, to avoid ambiguity. We also introduce an existing computational theory on communication.

In this paper, we focus on communication, which is "information-based interaction between peers, which has been developed for the purpose of surviving under pressure of natural selection". We have excluded actions which are of no direct benefit to survival or reproduction, such as simple observation and imitation. This is because we have no way of verifying algorithms that can undertake these actions. We have also excluded

reflex actions based on instincts because our focus is on the mechanism responsible for information processing in the brain. In addition, we restrict the partners of communication into peers, i.e., colleagues with equivalent constructions and complexity.

Existing studies have shown that estimating peers' internal state is the main form of processing the information involved in communication [KDH01]. Here, we define the internal state as a set of elements, which is not observable from one's external world but affects actions, e.g. emotions, desire, intent, and knowledge.

If a colleague can help, one can survive better than if alone. To get this assistance, one needs to appeal to peers to act beneficially for oneself, and this is the purpose of communication. To do this, however, one must know what form appeal must take to be effective. In other words, prior to appealing, one needs to predict others' actions under various conditions. This is required if the information for communication is to be processed effectively. Furthermore, since a peer's actions are hardly predictable from observable information, one needs to estimate others' internal states, prior to predicting actinos.

For example, it may be difficult to predict one's future actions based only on external observation of his state (a thin man with unsteady gait). If, however, we are able to estimate his internal states (hungry and looking for food), this helps us to predict his future actions (when he finds food, he will eat it). Such predictions help our actions for survival (we can hide our food) as well as the interactions that changes his internal states (giving the thin man food will alleviate his hunger and make him happy).

In advanced communications, internal states are estimated recursively, i.e. "by estimating 'how he is estimating my internal state' as part of his internal state". For example, chimpanzees exhibit deceiptful actions [PW78]. One chimpanzee pretended that some danger was approaching from a distance, expecting this action would distract another chimpanzee and prevent him from scolding him. Recursive estimates, which underlie such actions, also play a crucial role in human communication.

Despite its significance, a largely unknown is how the human brain estimates peers' internal states. One reason for this is the lack of computational studies. For any information processing done by humans, it is difficult to find the corresponding region in the brain, if we do not know what kind of representation is used or what algorithms can provide the process. Computational studies have tried to construct these representations and algorithms, so that we can use them as hypotheses of the mechanisms in the brain.

To study this further, we need a computational theory, which would form a common basis for the algorithms, such as the purpose of calculation and the structure of input/output. In this study, we define the computational theory for communication as follows.

---

1. Improve one's chances of survival and reproduction through selection of actions. (This includes the one selecting actions that causes action by peers that is beneficial to him).
2. Estimate the peer's actions to achieve 1.
3. Estimate the peer's internal states and the relation between internal states and the environments/actions to achieve 2.

The remainder of this paper is devoted to principles of algorithms that conform to this computational theory.

2.2 Dynamics-estimation framework

Estimating unobservable parameters involves the estimation of several components. We present a general framework of a dynamics-estimation problem to identify the components in estimating human dynamics.

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

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

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

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

Let us assume 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 results, the predictor needs to calculate an estimate for the current state $\hat{\boldsymbol{x}}(t) \in \hat{S}$ as well as estimate of the target dynamics, $\hat{f}$ and $\hat{g}$. This is a dynamics-estimation problem (Fig. 1).

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

However, the problem becomes extremely difficult if f and g are both unknown. Since we have no clues on original state space S, the estimator has to provide reconstructed state space $\hat{S}$, where state estimation $\hat{\boldsymbol{x}}$ is placed. Generally, in a deterministic system with constant input, reconstruction is not guaranteed unless the dimensions of $\hat{S}$ is larger than twice the dimensions of original state space S (embedding theorem [AIYK00]). The increase in the dimensions of $\hat{S}$ multiplies the number of parameters of estimated maps $\hat{f}$ and $\hat{g}$, and as a result, estimation becomes less feasible. Conversely, partial information on f or g may help estimation through reducing the dimensions of $\hat{S}$.

2.3 Communication in a framework of dynamics estimation

We can regard the process of estimating a peer's internal states as the dynamics-estimation framework extended to a non-stationary system. Although the two communicating peers are equal, we will only focus on one of them as the dynamics estimator, who we will call *the self*. The target of the self's estimates is called *the other*.

Figure 2 illustrates this framework, where the self is estimating the other's internal dynamics as well as the 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 part of the self's internal state, $\boldsymbol{x}_1$, and be constructed by model learner M in state-transition map f_1. Since the other is also constructing a dynamics model of the self, the whole framework becomes symmetrical. We can also see a fractal structure if the self recursively estimates a dynamics model of the self in the other's internal state.

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

3 Difficulties with parameter dimensions

In this section, we describe one of the two difficulties involved in estimating the other's internal states, i.e., restricted parameter dimensions of the estimator.

3.1 Limits of parameter dimensions

In the previous section, we stated that estimating the other's internal states can be regarded as

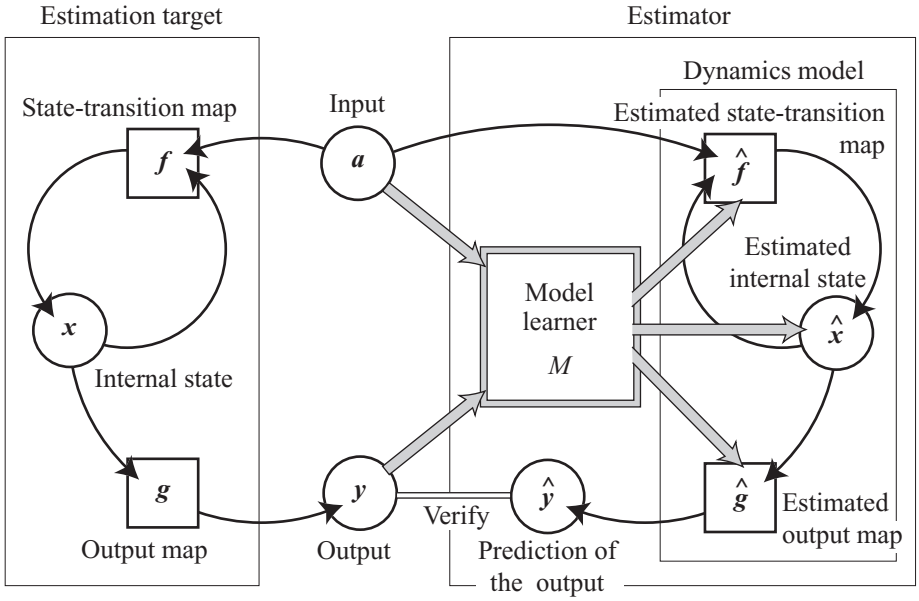


Figure 1: Framework of dynamics-estimation problem

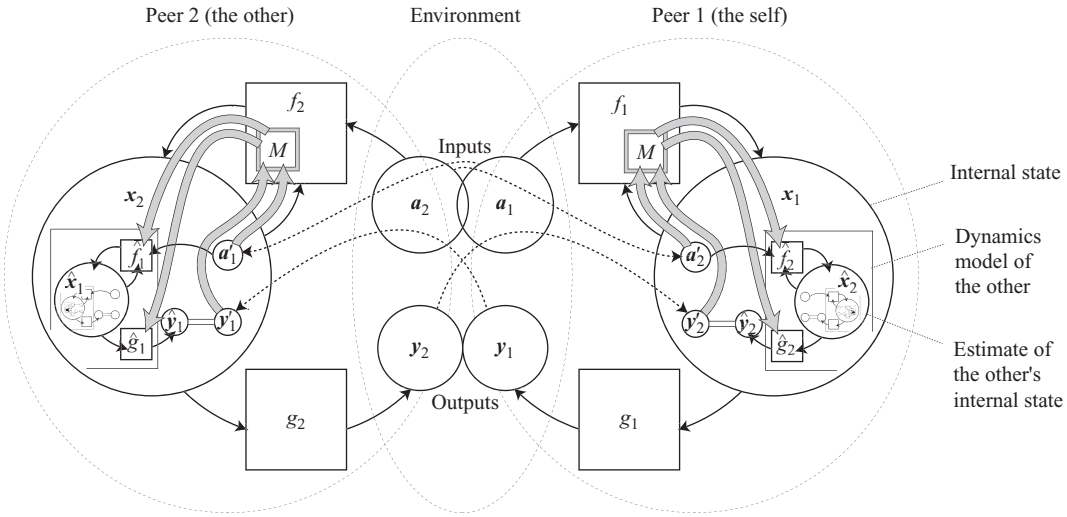


Figure 2: Communication in a framework of dynamics estimation

a problem within the dynamics-estimation framework.

However, the brains of animals have been performed dynamics estimation since ancient days, irrelevant of communication. Feedback-based control of muscle movement and environment map construction are good examples of how changes in the external world are predicted. The brain estimates non-observable states and transition/output maps to resolve the complex relations between input and output. Researchers are still trying to reproduce this form of estimation in a computational model [WK98].

Can estimating the other's internal states, which is the basis of communication, be accomplished through an extended, complicated version of dynamics estimation? It seems that, in the area of computational communication studies, there is an implicit consensus that the answer to this is 'yes'.

We, however, claim the reverse. This is because the other, who is equivalent 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, the estimator requires extra dimensions to reconstruct the state space. It is very difficult to estimate all of the target's dynamics — state-transition map $\hat{f}_2$, output map $\hat{g}_2$, and the current state $\hat{\mathbf{x}}_2$ — solely by observing the target from the outside.

Theoretically, it is still possible to estimate the dynamics through a massive amount of observation data. However, the estimator's parameter dimensions need to be far larger than those of the target's internal state. In other words, if the estimator's number of parameter dimensions is fixed, s/he can only estimate a very simple target, whose number of dimensions is substantially smaller than the estimator's. A more complex target beyond that can only be approximated vaguely.

For more advanced estimates, such as the recursive ones, it is necessary to estimate the internal state of the target, who is just as complex as the estimator. It is obvious that the target's parameter dimensions exceed any preconceived limits and cannot be estimated solely through external observation.

Even a human being has a very limited capacity to predict problems involving the estimation of dynamics with hidden parameters. It is very unlikely that a capacity beyond this is necessary for communication.

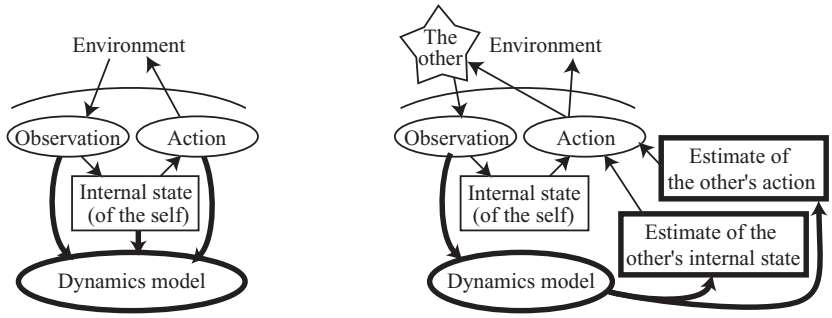
3.2 Clues for estimating dynamics

The problem with limited parameter dimensions is caused by the presumption that all dynamics are to be estimated only from observation. As discussed in Section 2.2, we may be able to solve this if we have some partial information on the dynamics, such as the state space, transition map, and output map.

One naive solution involves partial information that resides innately in the brain. That is, the human brain has *a priori* mechanism of basic information processing that extracts one's internal states from external observation. We know that the brain holds innate knowledge for some primitive elements, including that for facial expressions and tones of voice.

However, this solution is insufficient in describing human communication. We learn most of what we know, including self knowledge, after birth. It is unlikely that a human being innately knows all about the other, and undergoes complex processing such as recursive estimation. In fact, psychologists have evidence of age-dependent development in their 'theory of mind' in human children [Kuj97].

We need some description of how dynamics are learned. In other words, the brain needs some source for learning data, in addition to observing others.



(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 self

3.3 Solution using self

We propose that the estimator's brain learns the other's dynamics from his/her own dynamics (Fig. 3). That is, the self assumes that the other obeys similar dynamics to the itself. If this is true, the brain can roughly gauge the other's internal states. Slight difference can be acquired after this.

Three advantages can be gained from this assumption.

- It complies with the computational theory of communication, i.e., it enables the internal state of a peer to be estimated, as s/he is equivalent to the self and so are his/her dynamics. This is the most natural way of satisfying the constraint of equivalency.
- The self can obtain much more information from its own dynamics than the other's. In fact, a human being can observe his/her own internal states, such as emotions and desires, as well as his/her externally non-observable input, such as pain and touch. A dynamics model constructed from this wealth of information makes it much easier to estimate the internal states of others.
- We can also reduce state space dimensions. The brain does not need to reconstruct internal state space from scratch if it uses information on internal states and transition maps from the self. In addition, knowledge on the structure of internal state space simplifies the learning of the output map.

The following example illustrates this within the context of a human communication scenario. A person (the self) sees another person (the other) accidentally bumping his hand on a table and bleeding. The self has already learned, from his own experience, associations between input and the state change (bumping hand causes pain), and the associations between this observation and the internal state (bleeding hand is painful). Using the learned knowledge, the self estimates the internal state of the other from observing him/her and his/her environment (S/he bumped his hand and is bleeding, so this is painful). Combined with a decision on what action to take, the self can assist the other (apply ointment to the wound), and change his/her internal state (relieve the pain and make him/her happy).

Note that one's dynamics model does not necessarily reflect himself/herself perfectly. We know that the framework for observing the self involves substantial difficulties with representing its complexity within itself [Rös98]. In this study,

however, self-learning is only used to provide the 'seed' for the dynamics, enabling the other's internal state to be estimated. Thus, an approximate dynamics model to assess the self's dynamics is sufficient for this purpose.

4 Difficulties with converting objectivity to subjectivity

This section describes another difficulty, converting objective information to subjective information. The other's internal state cannot be estimated merely through the dynamics model on the self, because the input for the model, the other's subjective information, needs to be reconstructed from observed information. However, reconstructing subjective input is in the same class of difficulty as estimating internal states. In this section, we point out that a naive self-application principle cannot escape from the necessity of reconstructing subjective input, and propose a solution, a self-observation principle.

4.1 Self-application principle

One method, which naively achieves the proposal in the previous section, involves applying the self's action rule to predicting the other's actions. We name this as *self-application principle* (Fig. 4). Here, one's brain projects (copies) his/her own state-transition map f_1 and output map g_1 to the corresponding maps in the dynamics model, $\hat{f}$ and $\hat{g}$. After this, the brain uses the dynamics model to estimate the other's internal state $\hat{\boldsymbol{x}}$ and predict the other's output $\hat{\boldsymbol{y}}$.

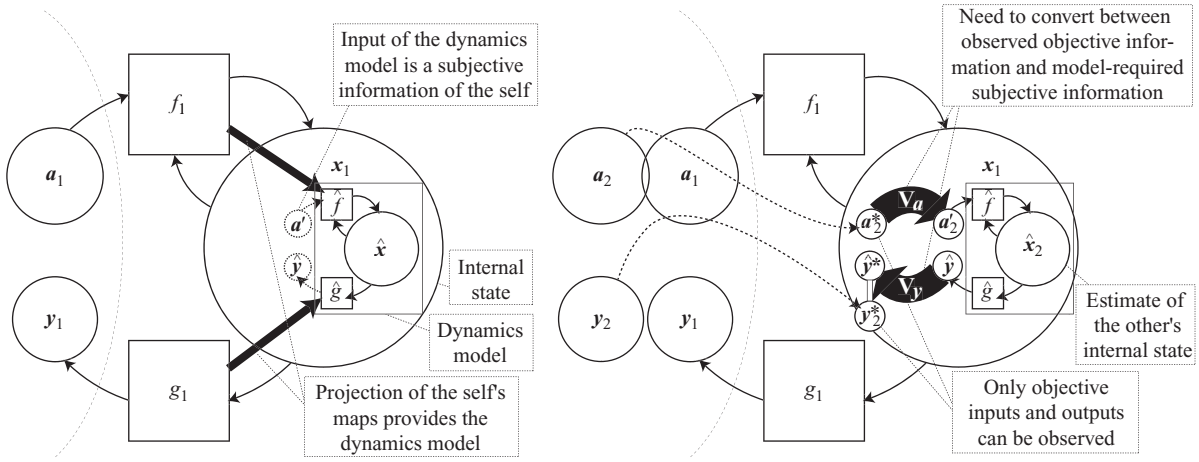
The problem with this principle lies in the different stances between the self and the other. One can only observe objective information from the other. (objective information is denoted by an attached asterisk, e.g. $\boldsymbol{a}_2^*$). This information is incompatible with the input of map $\hat{f}$, which has been copied from the state-transition map of the self. The compatible and required input $\boldsymbol{a}_2'$ is the input that the self would have received if s/he had had the other's stance, i.e., information on subjective input. To bridge this gap, the brain requires a conversion process $V_{\boldsymbol{a}}$, which converts $\boldsymbol{a}_2^*$ into $\boldsymbol{a}_2'$.

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

We claim that these processes, which convert objective information to subjective and vice versa, is the second difficulty with the computational theory of communication. Contrary to intuition, these processes are not trivial, because the two types of information are in totally different modes of representation. In terms of vision, the other's viewpoint needs to be obtained from the self's through a complex calculation that involves viewpoint conversion. Almost all senses of touch and pain needs to be reconstructed from other senses, such as visual and auditory sources.

For example, suppose that the self saw the other accidentally bumping his hand against a table. If the self puts himself/herself in the other's position, s/he could imagine the bump would create surprise in his emotion. This imagination, 'putting oneself in the other's position' is an estimation by the self-application principle. However, to do that, the self has to be able to convert his/her viewpoint into the other's subjective seeing of the table and the hand, and reconstruct the sense of pain from that; otherwise, s/he cannot put himself/herself in the position of the other. Such complex conversion and reconstruction is unlikely to be acquired from only observing others, let alone be innate knowledge in human beings.

In other words, estimating the other's subjective input is as difficult as estimating the other's internal state. The other's internal state is private, as is the other's subjective input; knowledge



(a) Self-projection stage (b) Self-application stage

Figure 4: Self-application principle

of private processes, $\hat{f}$ and $\hat{g}$, does not create a way of extracting private from public information. We claim that the self-application principle does nothing more than displace the problem of estimating internal state into that of converting objectivity to subjectivity, and does not solve the actual problem.

4.2 Self-observation principle

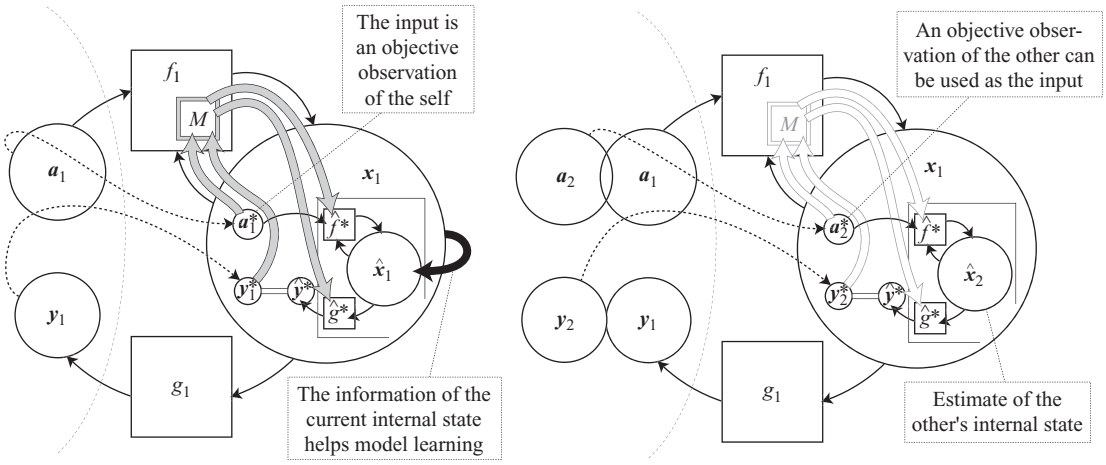
We propose here a *self-observation principle*, which resolves the actual problem. This is done by constructing a dynamics model from the objective observation of the self, and then applying it to estimate the other's actions.

Figure 5 illustrates this. The state of the self $\boldsymbol{x}$ 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 information that has been objectively observed.

In the model-learning stage (Fig. 5 (a)), model learner M learns the dynamics of $\boldsymbol{a}_1^*$ and $\boldsymbol{y}_1^*$, which are objective observations of the self's input and output. Here, *objective observation of the self* means observing the self through external environment, where a similar observation is applicable to peers, e.g. sight of hand moving, bumping sound, and allo-centric (objective space) arrangements of the table and hands. In this stage, the learning is easy because actual internal state $\boldsymbol{x}_1$ of the self can be used as teacher data for internal state $\hat{\boldsymbol{x}}$ of the model.

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

For example, the self first learns that "bumping the hand causes pain" from the his/her own experience. The self's dynamics model learns from objective observation of himself/herself (such as the sight of the hand moving, a bumping sound, and allocentric arrangements of the table and hand) and their relation to the pain in the self's internal state. After observing the other bumping his/her hand on the table, the self estimates that s/he feels pain from noticing such things as the other's hand moving, a bumping sound, and the



(a) Model-learning stage (b) Model-application stage

Figure 5: Self-observation principle

allocentric locations of the table and the hand.

Note that these observations have no direct benefit in determining the self's actions. To avoid bumping his/her hand again, the self only needs to learn the relations between ego-centric (subjective space) arrangements of the table, muscle movements of the arm, and the resulting pain. However, such subjective information is not enough to estimate the other's internal states. To do this, the self needs to obtain objective observations and relate them to his/her own 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 reduced elements to be calculated due to eliminating the conversion from objectivity to subjectivity and vice versa, $V_{\boldsymbol{a}}$ and $V_{\boldsymbol{y}}$.

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 the dynamics model is provided as model learner M in the self-observation principle, unlike simple projection in the self-application principle. From this point of view, the self-observation principle can be regarded as a variant of the self-application principle, which solves the problem of reconstructing 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 in the self's dynamics. it is not straightforward to design the learner in the self-application principle; when the prediction is wrong, the learner needs to choose whether map $V_{\boldsymbol{a}}$ or $\hat{f}$ is to be corrected. That is not the case with the self-observation principle, where the two maps are joined into $\hat{f}^*$.

4.4 New computational theory

The self-observation principle, which we propose, solves the two difficulties with estimating the other's internal states, i.e., the limited dimensions of estimator parameters and the conversion from objectivity to subjectivity. However, the principle is so general that a number of algorithms can satisfy the principle's requirements.

Of course, they have to solve many other practical difficulties.

To spur studies on these algorithms, we can integrate the proposed principle into a new computational theory. Although the existing computational theory (Section 2.1) was so incomplete that we could hardly find any algorithm that could deal with the difficulties, our new computational theory enables us to study new algorithms by clarifying difficulties and approaches. These new algorithms can lead to the development of a new artificial intelligence as well as hypotheses on the function of the brain.

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

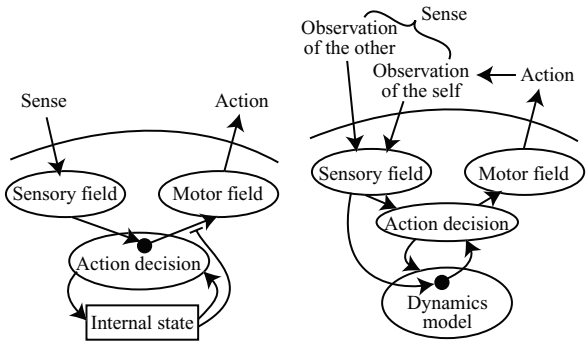
---

4. Construct a dynamics model, which is applicable to the other, through objective observation of the estimator himself/herself, to achieve (3) (estimation of the other's internal states).

Of course, we do not deny the possibility of an alternate computational theory of communication. There may be another way to solve the limits of estimator dimensions, and the original computational theory 'to estimate the other's internal state', which we have used as a basis, may need some modifications. We hope that our proposed principle will lead us to a deeper understanding of the computational studies of communication, including alternate theories.

5 Related studies

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



(a) Rizzolatti's hypothesis [RA98]

(b) Hypothesis derived from the self-observation principle

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

5.1 Psychology

The idea of "self-observation for a tool to estimate the other's internal states" was proposed in 1970s by an evolutional psychologist, Nicholas Humphrey [Hum78, Hum84]. However, he made a purely theoretical hypothesis, which described the origin of self-consciousness through a Darwinian process of evolution. His specifications for which part of the self should be observed were somewhat vague.

Our study, on the other hand, can be regarded as an improvement to Humphrey's theory for the following reasons. Through formalizing the communication process, we stated the effectiveness of the self-observing principle in solving the difficulties with communication. We also found that observation needed to create the 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 neuroscience and artificial intelligence.

Baron-Cohen, in his book on autism and the theory of mind [BC95], claimed that Humphrey's hypothesis corresponds to the self-application

principle, which contradicts the dissociation of symptoms in autism patients. He also briefly suggested that a different principle based on introspection (which resembles the self-observation principle) will match this dissociation. Although Humphrey's hypothesis covers both self-application and self-observation principles, it is interesting to find a study of autism that supports the advantage of the self-observation principle over the self-application principle.

5.2 Computational neuroscience

Most communication studies on computational neuroscience have not matched with the self-observation principle. Kawato et al. discussed the communication model as an extension of their model of dynamics interaction [KDH01], but their model implicitly assumed identical representation for both subjective and objective information.

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

Within a different context from communication, Tani proposed a constructivist approach for the study of the self and consciousness [Tan98]. However, in this study, an individual was placed alone in an environment, and it did not consider interactions between individuals, let alone estimates of the other's internal state.

5.3 Experimental neuroscience

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

Rizzolatti et al. proposed a hypothesis where a mirror neuron serves as a commander for an action as well as a recognizer of that action (Fig. 6(a)). However, the mechanism of association is unknown. In addition, the response of a mirror neuron to action being observed is claimed to be usually suppressed, without any reasonable discussions on suppression mechanisms.

In contrast, the self-observation principle suggests the role of mirror neuron outlined in Fig. 6 (b). The process in Section 4.2 described the development of mirror neurons very well. A neuron in the dynamics model learns the self's actions, and after this, applies them to the other's actions; as a result, the neuron begins to act as a mirror neuron. Moreover, such a neuron does not require any suppression mechanism because it does not directly trigger any action.

Oztop and Arbib [OA02] offered a hypothesis that the basic functionality of a grasping mirror system is to elaborate the appropriate feedback for opposition-space-based control when an object is manually grasped. They claim that, given this functionality, understanding of action 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 suggest that such a feedback mechanism is also the evolutionary origin of the self-observation mechanism in human communication.

5.4 Artificial intelligence

Existing strategic algorithms, e.g. chess-playing programs, are based on being able to the predict the opponent's actions. Most existing algorithms,

such as the alpha-beta algorithm [Ish89], assume that the opponent will evaluate 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 to this sort of strategic algorithm because the rule symmetry makes it easy to design the viewpoint translator. However, it is possible to design a new strategic algorithm based on the self-observation principle. A program based on such an algorithm would make it possible to learn the opponent's characteristics through successive games and adapt strategy.

6 Conclusion

We investigated existing computational theory on communication, i.e. the estimates of peers' internal states. We found that estimates are posed with two difficulties, the limits of the estimator's parameter dimension and conversion from objectivity to subjectivity. Our proposal for solving these difficulties was the *self-observation principle*: one observes oneself objectively to establish a dynamics model, which is then applied to others. Since the dynamics model learns the association between one's internal state and objective self observation, one can use the model to estimate the internal states of others by observing them objectively. Through clarifying the target and purpose of the learning process and integrating the self-observation principle into computational theory, this study has opened the way in enabling communication to be studied constructively. We also discussed the relation our proposal has with 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.

[Hum78] Nicholas Humphrey. Nature's psychologists. *New Scientist*, pages 900–904, June 1978.

[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 Soft-*

ware 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.

[MA03] Takaki Makino and Kazuyuki Aihara. Self-observation principle for estimating the other's internal state. Mathematical Engineering Technical Reports METR 2003–36, Department of Mathematical Informatics, Graduate School of Information Science and Technology, the University of Tokyo, October 2003.

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

[OA02] E. Oztop and M. A. Arbib. Schema design and implementation of the grasp-related mirror neuron system. *Biological Cybernetics*, 87:116–140, 2002.

[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.

[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.