Internal logic viewed from observation space: Theory and a case study

Motohiko Hatakeyama and Ichiro Tsuda

Department of Mathematics, Graduate School of Science,

Hokkaido University, Sapporo, 060-0810, Japan

Running Title: *Internal logic viewed from observation space*

Motohiko Hatakeyama

Department of Mathematics, Graduate School of Science,

Hokkaido University,

N 10 W 8, Sapporo, Hokkaido, 060-0810, Japan

Phone: `(+81) 11-706-2678`, Fax: `(+81) 11-727-3705`

E-mail: `hatake@math.sci.hokudai.ac.jp`

Abstract

We propose a framework of neurocognitive experiments that clarifies the mathematical structure of experiments and can be used to analyze experimental results and to determine the limitation of their possible interpretation. In contrast to the conventional analysis that employs simple Boolean logic, the present analysis treats classification in terms of higher-order functions. We also predict the existence of a previously unidentified type of neuron.

Keywords: observation space; interpretation; discrimination; multi-dimensional task; partition lattice; deductive process.

1 Introduction

The scientific method consists of a process of repeated observation, induction, deduction and verification (see Fig. (1)). In this process, one observes objects or phenomena using a certain tool (a) and then attempts through inductive manner to construct a theory that can account for the experimental data (b). Next, predictions regarding other behavior are derived from the theory using deductive methods (c), and the theory is verified or refuted by comparing these predictions with the experimental data (d). The present study is motivated by the observation that this conventional scientific method has been applied to cognitive-neural systems by tacit agreement, although it has apparently been realized that there is a difficulty involved in the construction and application of theories and experiments that are capable of providing a clear understanding of such systems. Specifically, this difficulty stems from the problems involved in designing experiments that can extract information that effectively elucidates cognitive processes. This limitation is not a technical one, but intrinsic to the object of study itself, because the brain is a self-modifying [7] or constantly adapting system that is open to the environment in the sense of both energetics and informatics. This limitation appears, in particular, when one studies internal cognitive states.

–Fig. 1–

One can endeavor to observe the brain's neural activities and from such observations attempt to describe cognitive functions, but there is no simple correspondence between these observed states with the internal states. The underlying limitation inherent in this process is due to several factors: the finite precision of measurements, the lack of an effective method of measurement and experimental devices, the fact that there is few proved neural correlates of cognitive process, the fact that there is no language common to the observer and the observed (animal), the relative nature of the observer's description and the internal states of the observed system, and so on. In addition, it is often difficult to properly interpret experimental data, and its misinterpretation can lead to serious misunderstanding. This difficult situation leads to the following general and well-known, but still important question: How could an external observer obtain a formal description of unknown internal processes of an intellectual system such as inference and reasoning?

From the above considerations, it is clear that in the study of cognitive-neural systems what we need is a formal theory that can explicitly distinguish observed states from internal states and describe not only experimental data but also the potential range of observation, that is, an "observation space." Such a formal theory would be capable of determining the boundary representing the limit of possible interpretations of the neural data associated with a cognitive experiment. Our aim is to define the concept of the observation space and then consider the development of neurocognitive theory from this point of view.

Definition When the definite design of an experiment and a formalized description of its result are given, the *observation space* in the framework of the experiment consists of the entire set of both the experimental and theoretical results that the experimenter can potentially describe.

As a first step toward the construction of such a general theory, we attempt to develop the formalization of a typical neurocognitive experiment with monkeys as a case study. Though we here treat only a series of experiments conducted by Sakagami et al., we believe that our theory may reveal generic properties of this kind of observation space.

In §2, we briefly explain the essence of the experiment and present a set-theoretical formulation of this type of experiment. Despite the fact that the brain of a live animal is essentially open even during the experiment, in which the behavioral task to be carried out defines a limited environment, in this type of experiment, its observables are reduced to a finite set of relations. The finite observation space characterizing this kind of experiment is formalized in §3. The construct of the observation space allows for implicit assumptions behind the experiments to be made more explicit, and this makes the construction of successful theories and experiments more feasible. In §4, we formalize a potential deductive process that can be carried out as an internal process of the animal under study and derive the concept of discrimination of stimuli that it can perform. Section 5 is devoted to discussion of the implications of this case study. We discuss, in particular, the relationship between deductive inference and the deductive process that is derived "externally" from our formalization.

2 A case study for conditional discrimination tasks with multi-dimensional visual stimuli

2.1 The conditional discrimination tasks of Sakagami et al.

Sakagami et al. conducted a series of experiments on macaque monkeys performing certain behavioral tasks [10], [11], [12]. The animals were trained to discriminate between different types of multi-dimensional visual stimuli, that is, stimuli with multiple attributes, such as color, shape and motion. While the animals performed this task of discrimination, Sakagami et al. made single-unit recordings from the dorsolateral prefrontal cortex (DLPFC) [10], [11] and from the ventrolateral prefrontal cortex (VLPFC) [12]. As we see below, the task of discrimination can be formalized as a composition of two operations: the operation of selection, made according to the attributes of the stimuli, and the operation of transformation, carried out in accordance with the behavioral meaning. The attention process carried out by the animal during the performance of task is analyzed with this formalization. Before proceeding to the description of our theory, we explain the task that Sakagami and Tsutsui used in the experiments [11].

–Fig. 2–

The task employed by Sakagami and Tsutsui, which we here refer to as "ST task", is schematically depicted in Fig. (2). In each trial, two visual stimuli are displayed successively on a monitor placed in front of the monkey. One of these stimuli, called the *target stimulus* (TS) is a multi-dimensional stimulus, possessing multiple attributes, such as color, direction of motion, and shape. The TS consists of random colored dots presented in a fixed area (*aperture*) of a certain shape. All the dots move in a single direction within this aperture. Another stimulus, the *cue stimulus* (CS), is presented at the center of the monitor as a fixation spot. The color of the CS indicates the attribute of the TS which the monkey should attend during a given trial. In other words, the monkey has to "understand" the meaning of the color of the CS in each trial. In each trial, the monkey is presented a pattern. Each such pattern is in general characterized by the attributes of, for instance, motion, color and shape, but for a given trial, only one of them is "meaningful." The meaningful attribute is specified by the color of the fixation spot, namely, the color of the CS. The condition represented by the color of the CS is called the *attending condition.* (More specifically, the attending condition is fixed during a *block*, consisting of 32–64 successive trials.) For instance, if this color is yellow, then the meaningful attribute is color and the attending condition in this situation is called the *color condition*, if it is purple, the meaningful attribute is motion and hence the *motion condition*, and if it is red, the meaningful attribute is shape and hence the *shape condition*. The monkey is then trained to act in the manner described below, in accordance with the identity of the meaningful attribute.

The monkey initiates a trial by pressing a lever, and the CS then appears. There are two types of responses that the monkey can make, "*go*" and "*no go*", which are executed by continuing to press the lever and releasing the lever, respectively. The correct response depends on both the attending condition (the color of the CS) and the attributes of the TS. The correspondence between the set of stimuli and the correct response type is fixed for each monkey. For each attribute, certain states correspond to the *go* response, and other states correspond to the *no go* response. For example, red means *go*, and right-directed motion means *no go*. Thus if a red, right-moving pattern is presented, the correct response is *go* under the color condition and *no go* under the motion condition.

After the monkeys were trained to perform the task with sufficient capability, which were assured with a probe test (see Ref. [10]), a measurement of neuronal activity during the performance of the task was made in an attempt to identify a neural correlate with perception and behavior. The activities of over 500 neurons were recorded with a single microelectrode in the DLPFC. The average frequencies of neuron spikes over a number of trials (*mean firing rates*) were calculated for each neuron and for each kind of stimulus. The measured neurons were classified into certain groups according to the statistical significance of the difference between the mean firing rates.

2.2 Formulation of a multi-dimensional task

Here, we make the first formulation of the ST task studied by Sakagami and Tsutsui [11]. Although in this section we restrict ourselves to the formulation of the task that the monkeys actually performed in the experiment, the theory we present can be applied to other types of experiments that involve conditional discrimination tasks with multi-dimensional stimuli.

Let the set $X = \{\mathtt{m}, \mathtt{c}\}$ represent the attending conditions, where `m` and `c` denote the motion and the color conditions, respectively. All the kinds of TS are denoted by $\boldsymbol{Y} = Y_{\mathtt{m}} \times Y_{\mathtt{c}}$, where $Y_{\mathtt{m}}$ and $Y_{\mathtt{c}}$ represent the sets of motion and color attributes, respectively. The set $Y_{\mathtt{m}}$ consists of two directions of motion, motion to the left (`l`) and motion to the right (`r`), and $Y_{\mathtt{c}}$ consists of two colors, purple (`p`) and yellow (`y`); i.e. $Y_{\mathtt{m}} = \{\mathtt{l}, \mathtt{r}\}$ and $Y_{\mathtt{c}} = \{\mathtt{p}, \mathtt{y}\}$. For simplicity, we call a given combination of stimuli (i.e. an element of $X \times \boldsymbol{Y}$) a *stimulus condition.* Further, we denote the set of responses as $Z = \{\mathtt{g}, \mathtt{n}\}$, where `g` represents "go"and `n` represents "no-go".

In a given trial, the appropriate response is uniquely determined by a pair of stimuli, the CS and the TS. This relation from the stimulus to the response can be represented by a map

$$f : X \times \boldsymbol{Y} \to Z. \tag{1}$$

It is natural to consider the map $f(x, \boldsymbol{y}) \in Z$ as representing the *behavioral meaning* of the stimulus condition $(x, \boldsymbol{y}) \in X \times \boldsymbol{Y}$.

Here, the map f has the special property that if the attending condition is fixed to `m` or `c`, the relation of the corresponding attribute $Y_{\mathtt{m}}$ or $Y_{\mathtt{c}}$, to the appropriate behavior Z is also a map. Thus f is a map that consists of multiple maps from each attribute to a response. These maps depend on the value of X. Let $f_i : Y_i \to Z$ be such a map of the attribute Y_i

$(i \in \{\mathtt{m}, \mathtt{c}\})$. Then, f can be expressed as

$$f(x, (y_{\mathtt{m}}, y_{\mathtt{c}})) = \begin{cases} f_{\mathtt{m}}(y_{\mathtt{m}}) & \text{if } x = \mathtt{m} \quad \text{(motion condition)}, \\ f_{\mathtt{c}}(y_{\mathtt{c}}) & \text{if } x = \mathtt{c} \quad \text{(color condition)}. \end{cases} \tag{2}$$

2.3 Classification of neurons

Now we consider the process of recording the neuronal activity. When the activity during a trial of the task is recorded, the relations among the stimulus condition $(x, \boldsymbol{y}) \in X \times \boldsymbol{Y}$, the behavioral meaning $z \in Z$, and the activity of a cell are treated. Because the data presented in Ref. [11] were analyzed only for "correct" responses (i.e. the case $z = f(x, \boldsymbol{y})$), the reported results are obtained in reference only to each stimulus condition $(x, \boldsymbol{y}) \in X \times \boldsymbol{Y}$. A statistical test of the mean firing rate of each cell was introduced to determine whether or not the activity differs between the stimulus conditions. The recorded neurons considered in Ref. [11] can be classified into 16 classes, as determined using the *analysis of variance* (ANOVA) with respect to two factors, Y_m and Y_c. In Appendix A detailed analysis is given to determine the kind of information that can be obtained with this particular statistical test.

–Fig. 3–

Among these 16 classes, Sakagami and Tsutsui determined 5 classes through their measurements and statistical tests, which we refer to as MI, CI, M, C and CM, as cell types displaying prominent activity (Fig. (3)). The classes we refer to as MI and CI in Ref. [11] are called *motion-intrinsic* and *color-intrinsic* cells, respectively. The cells in the classes MI and CI recognize the motion attribute $Y_{\mathtt{m}}$ and the color attributes $Y_{\mathtt{c}}$ of the TS, respectively, under both attending conditions m and c. The M cells recognize the motion attribute $Y_{\mathtt{m}}$ only under the motion condition, $\mathtt{m}$, while the C cells recognize the color attribute $Y_{\mathtt{c}}$ only under the color condition, $\mathtt{c}$. The CM cells recognize the motion attribute $Y_{\mathtt{m}}$ under the motion condition, $\mathtt{m}$, while they recognize the color attribute $Y_{\mathtt{c}}$ under the color condition, $\mathtt{c}$.

In Ref. [11], the responses of M, C and CM cells were also investigated in the case that the "shape" attribute was added as a factor of the multi-dimensional stimulus (TS). This shape attribute corresponds to the shape of the aperture, which consisted of stripes or a diamond in their experiments. In this case, X and $\boldsymbol{Y}$ can be redefined as $X = \{\mathtt{m}, \mathtt{sh}, \mathtt{c}\}$ and $\boldsymbol{Y} = Y_{\mathtt{m}} \times Y_{\mathtt{sh}} \times Y_{\mathtt{c}}$, where $\mathtt{sh}$ and $Y_{\mathtt{sh}} = \{\mathtt{s}, \mathtt{d}\}$ denote the shape condition and the set of shape attributes, respectively, and $\mathtt{s}$ and $\mathtt{d}$ denote stripes and a diamond, respectively. An asymmetry in the responses among the cell types was observed. It was found that the activities of most of the C cells are correlated with the behavioral response, i.e. $\mathtt{g}$ or $\mathtt{n}$ under the shape condition, whereas the M cells do not respond under the shape condition. It was also found that the CM cells exhibit a high correlation in their responses with the behavioral meaning Z under the shape condition, as well as under the color and motion conditions.

–Fig. 4–

It is well known that visual information concerning color and shape is processed in the *ventral pathway*, while that concerning motion is processed in the *dorsal pathway.* Thus the above cited results suggest that the C cells express behavioral meaning, through the ventral pathway, under both of color and shape conditions, while the M cells express behavioral meaning, through the dorsal pathway, under the motion condition. Taking this into account, we reorganize the cell classes as in Fig. (4).

–Fig. 5–

Based on an analysis of the latency of cell activity, the contralateral spatial preference of activity, and the projection relations between cortical areas, Sakagami and Tsutsui reported that the activity of M and C cells precedes that of CM cells, and process information through the ventral and dorsal pathways, respectively. They also reported that this information may be integrated in the activity of CM cells. A possible scenario for the entire process is depicted in Fig. (5).

In the next section, we further formalize, in a more generalized way, the stimulus-response relations of a subject who performs behavioral tasks that have already been learned, and also the situation that the activity of the brain is recorded. This allows us to define the concept of *discrimination* and show that the set of all discriminations can be considered the *observation space* of the task under consideration. We assert that with this formalization, it may also be possible to describe both the actual and potential results of such an experiment much more reliably than in the case of the conventional theory, and therefore that such a formalization will allow for relevant and testable predictions.

3 Observation space: A framework for the interpretation of the experimental results

3.1 S-R-A: stimulus, response and activity

Here we focus on a learning task that can be expressed by mapping relations from stimulus to response, that is, we consider the idealized situation in which the subject's behavior is determined only by external stimuli. Let S and R be a finite set of *stimuli* and a finite set of *responses* (or *behavior*), respectively. Then, the experimenter induces the subject to perform action that can be represented by a map $f : S \to R$. Furthermore, let A denote a set of neural *activity* that can be recorded from the subject's brain. Thus, in each of N trials of the experiment, the experimenter records a triplet of quantities consisting of the stimuli, response and activity: $(s_i, r_i, a_i) \in S \times R \times A$ $(i = 1, 2, ..., N)$. As a simplification, usually in this kind of experiment, it is assumed that any correlations between trials can be ignored, and thus each trial is considered to be independent. The distributions of quantities of $(s, r, a) \in S \times R \times A$ are then analyzed. Although there can be many relations among elements of this triplet, we here restrict ourselves to a map from a pair consisting of the stimulus and the response to the neural activity: $F : S \times R \to A$, taking into account of the experiment to be concerned in this paper. This means that we can only study activity with respect to the product space of stimulus and response, $S \times R$.

Because it is practically impossible to describe the state of the entire nervous system, measurements are taken from certain specific parts of the nervous system, focusing only on certain specific quantities. For example, the activity is measured as the mean firing rate of a single neuron, which we consider to be elementary unit. Here we refer to such an elementary unit a *measurement unit,* or simply a *unit.*

3.2 Discrimination of stimulus-response through activity

If the neurons are classified, on the basis of some kind of the statistical significance of the difference between the average activities for all stimulus and response pairs, forming $S \times R$, we can consider this classification of neurons to be represented by the *partition* of the space of $S \times R$. Under this assumption, let us consider what kind of information regarding the pair consisting of the stimulus and response could be represented by the behavior of a certain part or parts of the nervous system.

In our framework, the focus is not the neuronal activity itself, but, rather, the "discrimination" realized in the product space, $S \times R$, through the measured activity of a unit. To represent this properly, we define the concept of *discrimination* induced by a map.

Definition For a given map $\phi : U \to V$, a *discrimination* on the domain U induced by the map ϕ is an equivalence relation $\mathcal{D}_\phi \subseteq U \times U$ such that

$$\mathcal{D}_\phi = \{(a, b) \in U \times U \mid \phi(a) = \phi(b)\}. \tag{3}$$

(In standard terminology, such an equivalence relation is referred to as a "classification".)

According to this definition, $\mathcal{D}_\phi$ is a *quotient set* U/R_ϕ, where R_ϕ is an equivalence relation on U such that $aR_\phi b$ if and only if $\phi(a) = \phi(b)$. The equivalence relation R_ϕ discriminates U into disjoint subsets, i.e. the equivalence classes of U.

Though the range set V (or the image $\phi(U) \subseteq V$) may have some algebraic structure (for example, order), this definition of the discrimination does not take any such structure into account. This concept of the discrimination represents all the information that can be described if only the equality between elements of the set is taken into account. Thus, given a map between sets, that which can be realized on the domain set is represented by this discrimination. Moreover, note that the correspondence between a map and a discrimination is not one-to-one.

In the case that ϕ is a map from the product space of the stimulus and response to the observed activity of a measurement unit, only the *discrimination* induced by the map provides a method to extract meaningful information within the present framework. More precisely, for a map $F : S \times R \to A$ that represents the activity of a unit with regard to stimulus and response, the discrimination $\mathcal{D}_F$ on $S \times R$ induced by F represents the distinctions between the elements of $S \times R$ that can be obtained from A. Before continuing, we note here that the above definition can also be applied to the case of multiple measurement units. (Note that a discrimination realized in the case of multiple measurement units is a *refinement* of a discrimination realized with any single one of these units (see §3.3).)

If only "correct" responses are considered, as in the case of the previously studied ST task, according to the map $f : S \to R$, the activity that we study here can be expressed as a map from S only. The measurement of such activity is considered to be a discrimination induced by a map $\tilde{F} : S \to A$, where $\tilde{F}(s) = F(s, f(s))\ \ (s \in S)$. Although it is an interesting problem to investigate what happens if we also include "incorrect" responses, here we treat only correct responses.

3.3 Partition lattice as an observation space for finite S-R relations

The set of all discrimination on a finite domain U (i.e. all quotient sets on U) constitutes a partially ordered set ordered by *refinements*. This partially ordered set is called a *partition lattice* [3] (Fig. (6)).

–Fig. 6–

Definition Let U be a finite set. For any equivalence relations $\mathcal{D}$ and $\mathcal{D}'$ on U, $\mathcal{D}$ is a *refinement* of $\mathcal{D}'$ if and only if,

$$(a, b) \in \mathcal{D} \Rightarrow (a, b) \in \mathcal{D}' \ \text{ for all } \ a, b \in U. \tag{4}$$

Definition Let U be a finite set containing $k = |U|$ elements. The *partition lattice* Π_k of length k–1 of U is the partially ordered set of all equivalence relations (i.e. all discriminations) on U ordered according to refinements. That is, for any equivalence relations $\mathcal{D}_\phi$ and $\mathcal{D}_\psi$ on U, $\mathcal{D}_\phi \leqslant \mathcal{D}_\psi$ if and only if $\mathcal{D}_\phi$ is a refinement of $\mathcal{D}_\psi$.

A partially ordered set in which for any two elements there exists a *meet* (or *greatest lower bound*) and a *join* (or *least upper bound*) is called a *lattice* (see, e.g. Ref. [3]). The partition lattice Π_k is a lattice whose *meet* $\mathcal{D}_\phi \wedge \mathcal{D}_\psi$ is identified to the discrimination $\mathcal{D}_\phi \cap \mathcal{D}_\psi \subseteq U \times U$ (where $\cap$ represents set intersection) and the *join* $\mathcal{D}_\phi \vee \mathcal{D}_\psi$ is the *transitive closure* of the relation $\mathcal{D}_\phi \cup \mathcal{D}_\psi$ (where $\cup$ represents set union). Here, the transitive closure of a relation $R \subseteq U \times U$ is the minimal equivalence relation of U that contains R. The *bottom* (or the *least* element) of Π_k is the partition that differentiates all elements of U, and the *top* (or the *greatest* element) induces no differentiation of U.

As the number of elements of the domain set $k = |U|$ increases, the number of possible partitions $B(k) = |\Pi_k|$ increases rapidly. (The number $B(k)$ is known as the Bell number. Its recurrence formula is given by $B(k+1) = \sum_{i=0}^{k} \binom{k}{i} B(i) \quad (k = 0, 1, ...)$, provided that $B(0) = 1$.) In the case of the ST task with 2 attending conditions (§2.2, $|X| = 2$), there are $k = |X| \cdot 2^{|X|} = 8$ stimulus conditions, and hence the number of all the possible partitions is $B(8) = 4140$, whereas in the case $|X| = 3$ (§2.3), $k = 24$ and $B(24) \approx 4 \cdot 10^{17}$.

Because we can consider a discrimination for a finite relation defined by a map from $S \times R$ to A, the neuronal activity observed in an experiment can be placed at a certain point in the partition lattice. Thus the *observation space* of such an experiment can be considered the partition lattice when this discrimination constitutes all the *information* that can be obtained in the experiment.

Considering the scheme described in Fig. (1), the concept of the observation space helps us construct a framework on the space of observed data (the upper-right box in the figure) with respect to which the experiment is characterized. In actual experiments, however, information regarding an internal process of the subject is obtained not only through discrimination of the activity but through other observables as well, including the latency, the power spectrum, correlations, mutual information and anatomical structure. We ignore such other factors in the present formalization, in order to extract important information derived from the behavior of observables that have been treated in conventional neurophysiological experiments.

3.4 Deductive process and intermediate expression

In the present paper, we use the term *deductive process* in an operational or behavioral sense; that is, we do not take into account purely internal inference processes (including unconscious processes), but we do take into account the process of the subject's response to externally imposed conditions. Let us consider the situation in which the selection process of the response R given certain stimuli S consists of multiple stages. In this case, the deductive process should consist of a composition of sub-processes representing some intermediate stages.

With the above operational definition, we regard a deductive process as an application of some externally observable *rule*, which is expressed by a map ϕ from a set of inputs (stimuli) S to a discrimination $\mathcal{D}_\phi$. Thus, when an input S is given, the deductive process is identified with a discrimination on S. Similarly, for any finite set T, a process associated with any map $\psi : T \to T'$ (where T' is an arbitrary set) is expressed as a discrimination $\mathcal{D}_\psi$ on T. Hence, an intermediate stage of the process is also identified with a discrimination on S by an external observer who treats only "correct" trials whose responses are expressed by a map of S. We call this discrimination of an intermediate stage an *intermediate expression.*

The following propositions relate a process with multiple stages to the partial order relation of a partition lattice.

Proposition 1 *For any finite sets U and W and any set V, let a map $\phi : U \to V$ be a composition of two maps, $\psi : U \to W$ and $\psi' : W \to V$; i.e. $\phi = \psi' \circ \psi$. Then, the intermediate expression (the discrimination at the intermediate stage) $\mathcal{D}_\psi$ on U is a refinement of the discrimination $\mathcal{D}_\phi$ on U. Thus on the partition lattice $\Pi_{|U|}$,*

$$\mathcal{D}_\psi \leqslant \mathcal{D}_\phi. \tag{5}$$

Proof Because $\phi = \psi' \circ \psi$, obviously $\psi(a) = \psi(b) \Rightarrow \phi(a) = \phi(b)$ for any $a, b \in U$, and thus $(a, b) \in \mathcal{D}_\psi \Rightarrow (a, b) \in \mathcal{D}_\phi$. □

Roughly speaking, Proposition 1 implies that any deductive process defined in an operational sense is described as “climbing up” toward a coarser element on the partition lattice. Apparently, an external observer describing the activity resulting from the stimuli S can only treat the discrimination on S, so that the amount of information on S that the observer can obtain should monotonically “decrease” as the process proceeds to later stages. The next proposition asserts that the integrated information of concurrently processed sub-processes corresponds to the *meet* of these discriminations on the lattice.

Proposition 2 *Given $\phi_1 : U \to V_1$ and $\phi_2 : U \to V_2$, for any finite set U and any sets V_1 and V_2, let ϕ be a* tuple *of ϕ_1 and ϕ_2, i.e. $\phi : U \to V_1 \times V_2$, $\phi(u) = (\phi_1, \phi_2)(u) = (\phi_1(u), \phi_2(u))$ $(u \in U)$. Then, the discrimination $\mathcal{D}_\phi$ is the* meet *of the discriminations $\mathcal{D}_{\phi_1}$ and $\mathcal{D}_{\phi_2}$. Thus, on the partition lattice $\Pi_{|U|}$,*

$$\mathcal{D}_\phi = \mathcal{D}_{\phi_1} \wedge \mathcal{D}_{\phi_2}. \tag{6}$$

Proof The equivalence $\phi(a) = \phi(b)$ holds iff $\phi_1(a) = \phi_1(b)$ and $\phi_2(a) = \phi_2(b)$ for any $a, b \in U$. Thus $(a, b) \in \mathcal{D}_\phi$ implies $(a, b) \in \mathcal{D}_{\phi_1}$ and $(a, b) \in \mathcal{D}_{\phi_2}$; i.e. $\mathcal{D}_\phi \leqslant \mathcal{D}_{\phi_1} \wedge \mathcal{D}_{\phi_2}$. Next, note that for any $\psi : U \to W$, if $\mathcal{D}_\psi \leqslant \mathcal{D}_{\phi_1} \wedge \mathcal{D}_{\phi_2}$ then $\psi(a) = \psi(b)$ implies $\phi_1(a) = \phi_1(b)$ and $\phi_2(a) = \phi_2(b)$. Hence, we have the equivalence relation $\phi(a) = \phi(b)$, and thus $(a, b) \in \mathcal{D}_\psi$ implies $(a, b) \in \mathcal{D}_\phi$, i.e. $\mathcal{D}_\psi \leqslant \mathcal{D}_\phi$. Therefore, Eq. (6) follows. □

An intermediate expression as a partition is defined in reference to only the stimulus set, which the experimenter can only partially obtain from actual internal processes. What can be described is, however, just the partition of the stimuli when the activity of a unit belonging to an intermediate stage is measured. Therefore, the deductive process should necessarily be described as a transition between discriminations.

4 Theories and verification: possibilities for the deductive process and derived discriminations

4.1 Two possibilities that express the deductive process

Let us denote the set of the attending conditions by X ($n = |X|$), the set of multi-dimensional stimuli by the product of the attributes of stimuli, $\boldsymbol{Y} = \prod_{x \in X} Y_x$, and the set of responses by Z. In this case, the inference process that the subject is to carry out in the experiment can be expressed as the map,

$$f : X \times \boldsymbol{Y} \to Z, \tag{1}$$

which represents the "correct" correspondence of the stimuli to the responses.

In order to understand the potential substructure of f, we consider higher-order maps with one argument corresponding to f (this process is called the "Currying"). We adopt here two higher-order maps that can be obtained directly from f on the domain represented by the product $X \times \boldsymbol{Y}$:

$$g : X \to Z^{\boldsymbol{Y}}, \quad g(x)(\boldsymbol{y}) = f(x, \boldsymbol{y}) \tag{7}$$

$$h : \boldsymbol{Y} \to Z^{X}, \quad h(\boldsymbol{y})(x) = f(x, \boldsymbol{y}). \tag{8}$$

The map g is described as that which chooses a map $(\boldsymbol{Y} \to Z)$ from $|X| = n$ maps, in a manner that depends on X, whereas the map h is described as that which chooses a map $(X \to Z)$ from $|\boldsymbol{Y}| = \prod_{x \in X} |Y_x|$ maps, in a manner that depends on $\boldsymbol{Y}$. As mentioned in §2, the inference process intended in the experiments conducted by Sakagami et al. is of the type Eq. (7). Therefore, f is restricted to express a conditional map such as Eq. (2). Then a domain determined by $g(x)$ can be restricted to the corresponding attribute of the stimulus. In other words, for any $x \in X$, there is a map $f_x : Y_x \to Z$ such that

$$f_x(p_x(\boldsymbol{y})) = g(x)(\boldsymbol{y}), \tag{9}$$

where p_x denotes the projection of the product $\boldsymbol{Y}$ to Y_x. By contrast, h is not restricted in this way. Under the map h, the mapping of the conditions to the response is carried out in parallel (i.e., simultaneously and independently) for all stimulus attributes. Although the internal processes of the subject may proceed in this manner, the inference process induced by the experimental condition used by Sakagami et al. seems to be represented by the process Eq. (7) rather than the process Eq. (8). For this reason, we consider the case Eq. (7) in the following, although we cannot deny the possibility that there exist parallel processings like that described by Eq. (8), as well as others, even for conditional tasks.

We can further decompose $g(x) : \boldsymbol{Y} \to Z$ into two maps representing the *transformation process* and the *selection process*. The transformation process consists of the conversion from a stimulus to a behavioral meaning Z, which is expressed by f_x $(x \in X)$ in Eq. (9), and the selection process consists of the projection of an attribute in accordance with the attending condition $x \in X$. If each of these processes is carried out as one stage of the total process, then there are two possibilities for this total process, one in which the selection process is first and the transformation process is second and one in which this order is reversed. These two cases are expressed as follows:

$$g(x) = \varepsilon \circ \pi_x^\varepsilon, \quad \boldsymbol{Y} \xrightarrow[\pi_x^\varepsilon]{} Y \xrightarrow[\varepsilon]{} Z, \tag{10}$$

$$g(x) = \pi_x^\lambda \circ \lambda, \quad \boldsymbol{Y} \xrightarrow[\lambda]{} \boldsymbol{Z} \xrightarrow[\pi_x^\lambda]{} Z. \tag{11}$$

Here Y is the sum (disjoint union) of Y_x $(x \in X)$ (i.e., $Y = \coprod_{x \in X} Y_x$), $\boldsymbol{Z}$ is the product of Z over all $x \in X$ (i.e., $\boldsymbol{Z} = \prod_{x \in X} Z = Z^n$), and

$$\pi_x^\varepsilon(\boldsymbol{y}) = y_x \quad (x \in X), \tag{12}$$

$$\varepsilon(y) = \langle f_1, ..., f_n \rangle (y), \tag{13}$$

$$\lambda(\boldsymbol{y}) = (f_1, ..., f_n)(\boldsymbol{y}), \tag{14}$$

$$\pi_x^\lambda(\boldsymbol{z}) = z_x \quad (x \in X), \tag{15}$$

where $\boldsymbol{y} = (y_1, ..., y_n)$, $\boldsymbol{z} = (z_1, ..., z_n)$, $(f_1, ..., f_n) : \boldsymbol{Y} \to \boldsymbol{Z}$ is the *tuple* of $f_1, ..., f_n$, i.e., process of *selective attention* (cf., e.g., Ref. [6]).

Note that the manner of decomposing the process f adopted here is not unique. In fact, there are infinitely many possible ways to decompose this process. For example, it is clear that the number of intermediate stages of the process can be arbitrarily increased by means of inserting any *injection* (injective or one-to-one map) at any stage. Thus, the pair of decompositions Eq. (10) and Eq. (11) is merely a simplified such set among infinite possibilities.

4.2 The possibility of verification through the observation of neural activity

If the experimenter could observe the neural activity A of the brain in a suitable manner, the discrimination $\mathcal{D}_F$ induced by the map from the stimuli to the observed activity, $F : X \times \boldsymbol{Y} \to A$, should be identified with an intermediate expression described in §3.4. In order to understand this more clearly, let us consider the relation between two kinds of processes obtained theoretically, the early selection process and the late selection process, with the process derived from the experiment of Sakagami and Tsutsui [11]. In other words, the relationship between the classification of neuron groups given in §2.2 and the scenario suggested by latency analysis (Fig. (5)) is the subject to be considered here.

–Fig. 7–

In early selection represented by Eq. (10), the intermediate expression of the stimuli at the intermediate stage, Y, is the discrimination on the stimuli due to the selection process π_x^ε; that is, the intermediate expression should be given by $\mathcal{D}_{\pi_x^\varepsilon}$, whereby just an attribute for each attending condition is discriminated. In late selection represented by Eq. (11), the intermediate expression at the intermediate stage, $\boldsymbol{Z}$, after the transformation process λ gives the discrimination $\mathcal{D}_\lambda$. This discrimination divides n behavioral meanings based only on $\boldsymbol{Y}$, which is independent of X. One example of each of the intermediate expressions $\mathcal{D}_{\pi_x^\varepsilon}$ and $\mathcal{D}_\lambda$ in the case that X consists of three attributes ($X = \{\mathtt{m}, \mathtt{sh}, \mathtt{c}\}$), is depicted in Figs. (7a) and (7b), respectively.

–Fig. 8–

Let us consider the implications of the cell types found in experiments. Though there is an arbitrariness in defining the resulting classes as intermediate stages directly according to the scenario of Fig. (5), we regard the union of MI, SI and CI, i.e. MI + SI + CI, as the first stage, the union of M and C, i.e. M + C, as the second stage, and the CM class as the third stage. In this case, it turns out that MI + SI + CI discriminates the stimulus space as shown in Fig. (8a). Similarly, M + C gives the discrimination shown in Fig. (8b).

Although the discrimination given by MI + SI + CI (Fig. (8a)) is identical to the discrimination given by the transformation from the multi-dimensional stimuli $\boldsymbol{Y}$ to the multi-dimensional action $\boldsymbol{Z}$, as shown in Fig. (7b), it cannot be determined whether MI + SI + CI expresses the activity after or before the process of transformation to behavioral meaning, because all the maps $f_{\mathtt{c}}$, $f_{\mathtt{sh}}$, and $f_{\mathtt{m}}$ were *injections* in the experiment under consideration [11]. In other words, the discrimination induced by $\boldsymbol{Y}$ itself and that induced by $\boldsymbol{Z}$ were identical in that experiment. To clarify this point, further experiments employing many-to-one maps (see below) can be carried out. In the case that MI + SI + CI expresses activity after transformation to behavioral meaning, it can be concluded that late selection is suitable for the entire process. In this case, the transformation process should be carried out through the union of the intrinsic cells' activity, MI + SI + CI, and the successive selection process is regarded as being divided into multiple stages, with the intermediate stage M + C. In the case that the union MI + SI + CI expresses activity before the transformation process, but the other union, M + C, expresses activity after transformation, MI + SI + CI merely expresses the discrimination of the TS, $\boldsymbol{Y}$, whereas the process between MI + SI + CI and M + C may consist not only of the transformation process but also of a part of the selection process, because the intermediate expression $\mathcal{D}_{\pi_x^\varepsilon}$ differs from the discrimination computed with M

+ C. The particular asymmetric property of C cells, whose responses are correlated with both the color and the shape conditions, suggests that the behavioral responses under these attending conditions are partially selected in *each* C cell if these cells express the already converted behavioral meaning under these attending conditions. In fact, in a different paper by Sakagami et al. [12], it is reported that 76% of C cells (out of 25 C cells) found in the VLPFC also recognize the color attribute under the color condition for a different set of target stimuli. Because the transformation process from the TS was not an injection in Ref. [12], this result suggests that the C cells (at least in most cases) express the behavioral meaning that has already been converted.

Now, suppose there exist other types of cells, C′ cells and S cells, with C′ cells reacting to the color attribute only under the color condition, and S cells reacting to the shape attribute only under the shape condition. We obtain the discrimination induced by the union M + S + C′. This is displayed in Fig. (8c). This discrimination is identical to the discrimination $\mathcal{D}_{\pi_x^c}$ induced by the projection displayed in Fig. (7a). These supposed classes, C′ and S, and the stage M + S + C′ construct one possible stage that may intervene between the stages MI + SI + CI and M + C.

To decide which process, early or late selection, provides a plausible interpretation of neural activity correlated with behavior, it is sufficient to check whether MI, SI and CI cells are correlated with the behavioral response under corresponding conditions. If MI, SI and CI are correlated with the behavioral response, in which case the discrimination induced by MI + SI + CI is the intermediate expression $\mathcal{D}_\lambda$, the late selection is suitable. If MI, SI, CI and M, as well as the supposed classes S and C′, are all uncorrelated with the response, in which case the discrimination induced by MI + SI + CI is identical to that induced by

$\boldsymbol{Y}$, and the discrimination induced by M + S + C′ is identical to $\mathcal{D}_{\pi_x^\varepsilon}$, the early selection is suitable. Actually, the possibility that the intermediate stage $\mathcal{D}_{\pi_x^\varepsilon}$ consists of the supposed stage M + S + C′ cannot be ruled out. However, it is likely that the process as a whole is late selection, because the results of experiment reported in Ref. [12] indicates that most of the C cells respond to the behavioral meaning. Even in this case, the selection process would likely be carried out in multiple stages intervened by M + C. Thus, the selection process would be divided into a process executed along the ventral pathway (for the color and the shape conditions) and a process executed along the dorsal pathway (for the motion condition). Again, note that the splitting of the selection process into these two selection sub-processes is merely one among an infinite number of possibilities. However, our approach provides a suitable framework to compare and analyze such possibilities, rather than to derive a conclusion regarding this specific case.

4.3 Applicability of the theory

In the theory presented in this paper, for a certain behavioral task, we introduced the concepts of stimuli S, responses R, and activity A expressed as an S-R map relation, and we related them to intermediate expressions.

We associated the sets $X \times \boldsymbol{Y}$ and Z with the stimuli S and responses R, respectively, and we also provided the correspondence between the discriminations statistically differentiated by the firing rates of neurons averaged over trials and the discriminations differentiated by intermediate expressions. The classification of the sets $X \times \boldsymbol{Y}$ and Z directly reflects the assumptions that we made. Furthermore, it is evident that information regarding the mean firing rate alone is not sufficient to obtain a useful description of the activity of neurons (e.g., Refs. [1],[2],[5],[13],[14],[19]). Our present framework itself is, however, independent of neural representations, such as various types of temporal coding.

Whenever an observer attempts to obtain a correspondence between the formalization of some phenomenon and the phenomenon itself, the ambiguity that stems from the observer's *arbitrariness of discrimination* is inevitable. Hence the results obtained from the present theory—as those of any theory—must be considered to be dependent on the hidden conditions, namely the tacit assumptions. Conversely, one of the most important advantages of the present formalization is to reveal such tacit assumptions.

5 Discussion

We constructed in the present paper a formal theory of tasks in neurocognitive experiments. This formal theory can provide an appropriate description of conditions that are applied to experiments and extract the structure of internal cognitive process in relation with neuron's activity. Our approach has provided a new classification of discrimination at an animal's "cognitive" level. Conventional theories provide merely an identification in terms of Boolean logic, which is typically seen in association tests with pigeons, and other identifications using simple stochastic renewals of maps, which is typically seen in experiments on reinforcement learning in T-shaped mazes with rats. The present theory, on the other hand, provides another more dynamic identification in terms of higher-order maps.

In general, brain sciences should work to bridge the gap between the neuronal level and the functional level. Physiological approaches, which usually give rise to the reduction of function to neuronal characters, possess predictive ability concerning behavior at the neuronal level but lack, in general, explanatory power with regard to cognitive functions. On the other hand, cognitive approaches, whose extreme example is simple behaviorism, possess explanatory capability regarding cognitive function but lack predictive power concerning neuronal level phenomena [18]. Thus, physiological and cognitive levels provide descriptions at two opposite extremes.

The present theory provides a description at intermediate levels between these two extremes. To describe neural systems at various levels, a variety of dynamical models have been adopted. The present theory can be viewed as a dynamical model in the sense of clarifying animal cognitive behavior and related neuronal activity by providing various types of mappings and their relations restricted to a given neurocognitive experimental task. The present results thus represent an attempt to reveal the underlying mechanism. We actually demonstrated a method to formulate a cognitive experiment. Formulating the measurement and analysis of the task, we elucidated the "observation space." As one example, we associated two formal processes with early and late selection. This approach allows us to classify experimental situations systematically from an operational point of view. Furthermore, this work may also provide a framework for further experiments which may be capable of clarifying a mathematical structure of the process of discrimination of stimuli and that of transformation process of cognition to behavior.

Thus, the present theory may provide a concrete method for the interpretation of cognitive function in terms of brain activity, namely a scheme of hermeneutic device [4]. Such a formalization of the experimental context would make tacit assumptions explicit. If this scheme were sufficiently powerful, it would then be possible not only to use it in analyzing existing neurocognitive experiments but also in devising new experiments.

Although the present framework is predictive, it is "static" in the sense that it does not treat more complex dynamics of the brain and mind. As stated in a previous paper [17], the dynamics of a description can be formulated using a certain type of functional map (see also Refs. [8],[15],[16]). The use of such a map may provide a direction for the development of the present framework. In addition, if the formal language is sufficiently potent, it may provide a tool to describe the function of the system itself at another cognitive level. Such a capability is related to Rosen's "relational biology" [9]. These extensions of our framework are left to a future study.

6 Acknowledgment

We would like to express our sincere thanks to Masamichi Sakagami, Hideaki Saito, Minoru Tsukada and Okihide Hikosaka for fruitful discussions on the experiments conducted by Sakagami et al. We also thank Shigeru Kuroda, Otto Rössler and Hans Diebner for important comments on the present theory.

A Appendix

A.1 Discrimination derived using ANOVA

Sakagami and Tsutsui [11] classified measured cells according to the results of 2-factor analysis of variance (ANOVA) with repeated measures on factors $Y_{\mathtt{m}}$ and $Y_{\mathtt{c}}$ (attributes of TS) under several attending conditions. (Tests for factor X were not carried out.) In this analysis, for a pair of levels $(i, j) \in Y_{\mathtt{m}} \times Y_{\mathtt{c}}$, a random variable of the activity of cell A_{ij} is expressed as a linear model,

$$A_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}, \tag{16}$$

where μ is a constant, α_i and β_j are the main effects of the factors $Y_{\mathtt{m}}$ and $Y_{\mathtt{c}}$, respectively ($\alpha_{\mathtt{l}} + \alpha_{\mathtt{r}} = \beta_{\mathtt{p}} + \beta_{\mathtt{y}} = 0$), and γ_{ij} is the $Y_{\mathtt{m}}$–$Y_{\mathtt{c}}$ interaction ($\gamma_{\mathtt{lp}} + \gamma_{\mathtt{ly}} + \gamma_{\mathtt{rp}} + \gamma_{\mathtt{ry}} = 0$). The error term ϵ_{ij} is an independent random variable following Gaussian distribution with mean 0 and variance σ^2, $\mathrm{N}(0, \sigma^2)$. Then, for each attending condition $x \in X$, there are three null hypotheses,

$$H_{x\mathtt{m}}: \ \alpha_{\mathtt{l}} = \alpha_{\mathtt{r}} \ (= 0),$$

$$H_{x\mathtt{c}}: \ \beta_{\mathtt{p}} = \beta_{\mathtt{y}} \ (= 0),$$

$$H_{x\mathtt{i}}: \ \gamma_{\mathtt{lp}} = \gamma_{\mathtt{ly}} = \gamma_{\mathtt{rp}} = \gamma_{\mathtt{ry}} \ (= 0).$$

Therefore, for two attending conditions (`m` and `c`), there are six hypotheses in all. Because the effect of the interaction, however, was not considered as criterion in their classification, we consider only four hypotheses regarding the main effects, $H_{\mathtt{mm}}, H_{\mathtt{mc}}, H_{\mathtt{cm}}$ and $H_{\mathtt{cc}}$. A cell is then classified into one of $2^4 = 16$ classes, as shown in Fig. (3), according whether or not each of these four hypotheses is rejected (i.e., these classes form a *Boolean lattice* $\mathbf{2}^4$). We call these 16 classes *statistically distinct classes*, or simply *classes*. Let $\tilde{H}$ be the proposition

“A null hypothesis H is rejected by the test with a given significance level,” and let $\mathring{H}$ be the negation of $\tilde{H}$, but it should be noted here that not rejecting H does not imply that H is true.

Using this notation, the classes CI and MI are represented as the cases $\mathring{H}_{\mathtt{mm}}\tilde{H}_{\mathtt{mc}}\mathring{H}_{\mathtt{cm}}\tilde{H}_{\mathtt{cc}}$ and $\tilde{H}_{\mathtt{mm}}\mathring{H}_{\mathtt{mc}}\tilde{H}_{\mathtt{cm}}\mathring{H}_{\mathtt{cc}}$, respectively. (For simplicity, the logical product is denoted by juxtaposition.) Similarly, the class C corresponds to $\mathring{H}_{\mathtt{mm}}\mathring{H}_{\mathtt{mc}}\mathring{H}_{\mathtt{cm}}\tilde{H}_{\mathtt{cc}}$, the class M to $\tilde{H}_{\mathtt{mm}}\mathring{H}_{\mathtt{mc}}\mathring{H}_{\mathtt{cm}}\mathring{H}_{\mathtt{cc}}$, and the class CM to $\tilde{H}_{\mathtt{mm}}\mathring{H}_{\mathtt{mc}}\mathring{H}_{\mathtt{cm}}\tilde{H}_{\mathtt{cc}}$.

How precisely do these classifications for measured cells reflect the information concerning stimulus conditions that the cells carry? If we can ignore *error of the first kind* (that is, statistical error arising in the case that a hypothesis is rejected when it is in fact true), rejecting a hypothesis on the basis of a main effect implies discrimination of the stimulus conditions in the corresponding factor. On the other hand, because we can never confirm a hypothesis in a definite sense, and therefore even if there is no rejection of a hypothesis, it cannot be concluded with certainty that there is no discrimination in those stimulus conditions for that cell. Theoretically, even if a hypothesis is true, the nonexistence of a difference between the stimuli is never guaranteed, due to a more general problem: Measured quantities may not sufficiently reflect the information concerning stimuli that a cell may carry. If there is interaction between the factors, there are generally more complex relationships between the dependence of the cell behavior on the various stimuli (however, see A.2). Furthermore, nothing definite can be stated about the statistical difference between the activities for different attending conditions, because no test for X has been carried out. Sakagami and Tsutsui, however, demonstrated that the activity patterns of all observed CM cells correspond to the behavioral response Z, and hence, it appears possible that CM cells discriminate stimulus conditions in a manner similar to the behavioral response.

A.2 Correspondence with discriminations

Now, let us reconsider the statistical method and the results of the ST task described in §2.3 from the viewpoint introduced in §3. The 16 statistically distinct classes obtained using the 2-factor ANOVA ($Y_{\mathtt{m}}$ and $Y_{\mathtt{c}}$), displayed in Fig. (3), do not have a simple correspondence with discriminations.

As mentioned in A.1, because it cannot be claimed that there is no difference between any two stimulus conditions, the possible correspondence between statistically distinct classes and *discriminations* must contain all the refinements of the discriminations. In other words, letting Π be the *partition lattice* consisting of all discriminations on some domain, and letting Γ be the subset of Π corresponding to a class, Γ must be a *down-set* of Π, that is, for any $\mathcal{D}_\phi \in \Gamma$ and $\mathcal{D}_\psi \in \Pi$, $\mathcal{D}_\psi \leq \mathcal{D}_\phi \Rightarrow \mathcal{D}_\psi \in \Gamma$.

A concrete correspondence is derived using the statistical model Eq. (16). Let us denote the statement that a hypothesis H itself is true by the same symbol H, and its negation (i.e., the statement that H is false) by $\bar{H}$. If we assume that no error of the first kind occurs, $\mathring{H}$ implies H or $\tilde{H}$, while $\bar{H}$ implies $\tilde{H}$ only. Though the $Y_{\mathtt{m}}$–$Y_{\mathtt{c}}$ interaction was not taken into account when classifying the results of the neuronal activity in the experiment, the correspondence between $\bar{H}$ and $\tilde{H}$ is not affected by whether or not there is an interaction, because each subset in the case without an interaction contains the subset in the case with an interaction.

Finally, if the domain $\boldsymbol{Y}$ is merged with another factor that was not statistically tested in the ST experiment, such as the attending condition X, the correspondence between the classes and the discriminations becomes more complex. In this case, the correspondence between the classes and all possible relations between discriminations for different attending conditions should be considered.

References

[1] Abeles, M., *Corticonics: Neural Circuits in the Cerebral Cortex.* (Cambridge University Press, Cambridge, 1991).

[2] Aertsen, A., Erb, M. and Palm G., Dynamics of functional coupling in the cerebral cortex: an attempt at a model-based interpretation. *Physica D* **75** (1994) 103–128.

[3] Birkhoff, G., *Lattice theory,* 3rd ed. (Colloquium Publications vol. 25, Amer. Math. Soc., Providence, RI, 1967).

[4] Erdi, P and Tsuda, I. Hermeneutic approach to the brain: Process versus device, *Theoria et Historia Scientiarum,* **VI** (2) (2002) 307–321.

[5] Fujii, H., Ito, H., Aihara, K., Ichinose, N. and Tsukada, M. Dynamical cell assembly hypothesis—Theoretical possibility of spatio-temporal coding in the cortex. *Neural Networks* **9** (1996) 1303–1350.

[6] Johnston, W. A. and Dark V. J., Selective Attention, *Annual Review of Psychology* **37** (1986) 43–75.

[7] Kampis, G., *Self-modifying Systems in Biology and Cognitive Science—A New Framework for Dynamics, Information and Complexity.* (Pergamon Press, Oxford, 1991).

[8] Kataoka, N. and Kaneko K., Functional Dynamics I: Articulation Process, *Physica D* **138** (2000) 225–250.

[9] Rosen, R., *Life Itself: a comprehensive inquiry into the nature, origin, and fabrication of life.* (Columbia University Press, New York, 1991).

[10] Sakagami, M. and Niki, H., Encoding of behavioral significance of visual stimuli by primate prefrontal neurons: relation to relevant task conditions, *Exp. Brain Res.* **97** (1994) 423–436.

[11] Sakagami, M. and Tsutsui, K., The hierarchical organization of decision making in the primate prefrontal cortex, *Neurosci. Res.* **34** (1999) 79–89.

[12] Sakagami, M., Tsutsui, K., Lauwereyns, J., Koizumi, M., Kobayashi, S. and Hikosaka, O., A Code for Behavioral inhibition on the Basis of Color, But Not Motion, in Ventrolateral Prefrontal Cortex of Macaque Monkey, *J. Neurosci.* **21** (2001) 4801–4808.

[13] Singer, W., Neuronal synchrony: A versatile code for the definition of relations? *Neuron* **24** (1999) 49–65.

[14] Steinmetz, P. N. et al., Attention modulates synchronized neuronal firing in primate somatosensory cortex, *Nature* **404** (2000) 187–190.

[15] Takahashi, Y., Kataoka, N., Kaneko, K. and Namiki, T., Function Dynamics, *Japan Journal of Industrial and Applied Mathematics* **18** (2001) 405–423.

[16] Tsuda, I. and Tadaki, K. A logic-based dynamical theory for a genesis of biological threshold, *BioSystems* **42** (1997) 45–64.

[17] Tsuda, I. and Hatakeyama, M., Making Sense of Internal Logic: theory and a Case Study. In *Sciences of the Interface,* eds. by Diebner, H. H., Druckrey T. and Weibel P. (Genista, Tübingen, 2001).

[18] Tsuda, I. Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems, *Behavioral and Brain Sciences* **24** (2001) 793–847.

[19] Tsukada, M. A proposed model of the hippocampal-cortical memory system and temporal pattern sensitivity of LTP in hippocampal neurons. *Concepts in Neuroscience* **3** (1992) 213–224.

Figure Captions

Fig. 1. The circular process involved in the construction of scientific theories: (a) Observation of an object or phenomenon in the real world produces observed data; (b) from the observed data, a formal theory or hypothesis is constructed; (c) deducing from the theory, predictions regarding other phenomena are made; (d) comparing these predictions with newly gathered data, the theory can be verified or refuted. The inductive process (b) itself cannot be formalized, and in any theory, a hypothesis remains until it is rejected. This scientific process often confronts difficulty in the treatment of complex and open systems.

Fig. 2. The behavioral task investigated by Sakagami and Tsutsui [11]. (a) A schematic depiction of the task. Two visual stimuli, a cue stimulus (CS) and a target stimulus (TS), were presented on a computer monitor to a monkey. The CS specified the attending condition, i.e. the attribute of the TS to which the monkey should pay attention. The CS was also used as a fixation spot. The TS was a multi-dimensional visual stimulus, which appeared at one of four locations at random. The monkey responded by releasing or continuing to press a lever. Neural activity in the prefrontal cortex was recorded with a microelectrode. (b) Time sequences of trials for “go” and “no-go” responses. When the monkey pressed the lever, a trial started, and the CS was presented. After 1–2 s, the TS was presented for 200 ms. After 1–2 s from the stop of TS presentation, the light expressing CS was dimmed and remained dimmed during 1.2 s. In the case of “go” response, the monkey has to release the lever within this dim period. In the case of “no-go” response, the monkey has to continue pressing the lever during this period.

Fig. 3. Sixteen possible classes statistically distinguished via the analysis of variance, taking two factors, $Y_{\mathtt{c}}$ and $Y_{\mathtt{m}}$ into consideration. These classes are arranged on a Boolean lattice $\mathbf{2}^4$. The activity of each class is represented by eight boxes, which correspond to the stimulus conditions $X \times \boldsymbol{Y}$. The arrangement of the stimulus conditions in these boxes is displayed in the lower-right inlet. Also, the "correct" behavioral responses Z are depicted in this frame for reference. In each row (corresponding to the attending condition), boxes drawn by the same pattern imply "identical" in the sense that it cannot be said that all activity of cells responding to a corresponding stimulus is not statistically identical. The classes labeled MI, CI, M, C and CM are those examined by Sakagami and Tsutsui. MI and CI cells discriminate $Y_{\mathtt{m}}$ and $Y_{\mathtt{c}}$ factors, respectively, under both attending conditions. M discriminates $Y_{\mathtt{m}}$ under the motion condition, `m`, only, and C discriminates $Y_{\mathtt{c}}$ under the color condition, `c`, only. CM cells discriminate $Y_{\mathtt{m}}$ under the motion condition, `m`, and $Y_{\mathtt{c}}$ under the color condition, `c`. The numbers in the figure indicate the number of cells belonging to the corresponding classes among the total of 523 cells in the dorsolateral prefrontal cortex. The significance level of the statistical test was 0.01. (See Appendix A.1 for statistical analyses.)

Fig. 4. Statistical discriminations of M, C and CM cells in the case that the shape condition is included (the middle row). C cells exhibit discrimination that corresponds to the behavioral meaning Z for the color and shape conditions, whereas the discrimination of M cells corresponds to the behavioral meaning for the motion condition only. CM cells exhibit discrimination that is equivalent to the behavioral meaning Z under all conditions (see the right inlet).

Fig. 5. A possible scenario of information processing, based on the work of Sakagami and Tsutsui [11], for multi-dimensional visual stimuli. Information regarding the color and shape of the multi-dimensional stimulus is processed along the ventral pathway (via the temporal cortex), and that regarding the motion is processed along the dorsal pathway (via the parietal cortex). Actually, Sakagami and Tsutsui found evidence suggesting the existence of M, C and CM cells (and small numbers of MI and CI cells, too) in the dorsolateral prefrontal cortex. MI, SI and CI cells are assumed to act in the early stage of the process. There is evidence that C cells integrate information from CI and SI cells, whereas M cells are organized directly from MI cells. CM cells integrate the information of M and C cells and this makes a determination of behavior.

Fig. 6. An example of a Hasse diagram for a partition lattice: a partition lattice of length 3 (Π_4). Π_4 orders all partitions of four components, that is, all possible ($B(4) = 15$) equivalence relations of a set consisting of four components. In this ordering the least element segregates each component, and the greatest element segregates no components. The partial order represents refinements of the components; i.e. a lower element connected by line(s) refines those above it.

Fig. 7. Intermediate expressions of two hypotheses concerning the internal process. (a) The case of early selection. The selection process is responsible for the discrimination of the stimulus space. Hence, each attribute is discriminated under the respective attending conditions. (b) The case of late selection. The transformation process is responsible for the discrimination of the stimulus space. In other words, the transformation process decides the

behavioral meaning for each element of $\boldsymbol{Y}$.

Fig. 8. Discriminations at the intermediate stages displayed in Fig. 5. (a) The union of cell types, MI, SI and CI discriminates each column of $\boldsymbol{Y}$, but it is impossible to decide whether this discrimination is associated with the stimulus or the behavioral meaning. (b) The discrimination derived from the union of cell types M and C. Because the cell type C responds to the behavioral meaning of the color and shape conditions, the behavioral meanings for these two attending conditions are not discriminated. (c) The discrimination derived from the union of cell types M and S, and the supposed cell type C$'$ which is assumed to respond to the color attribute only under the color condition. (See Fig. 4 for the arrangement of partitions representing the stimuli.)

Figures

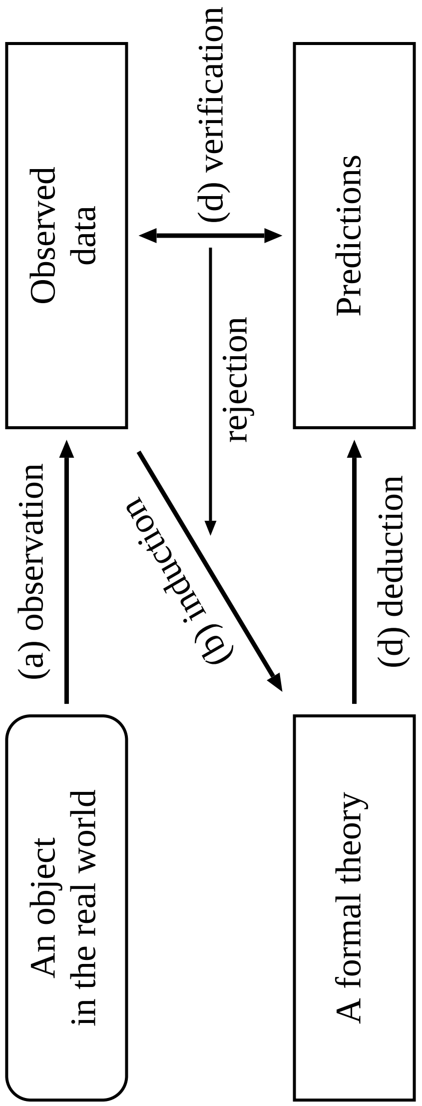

Figure 1. [upside ←]

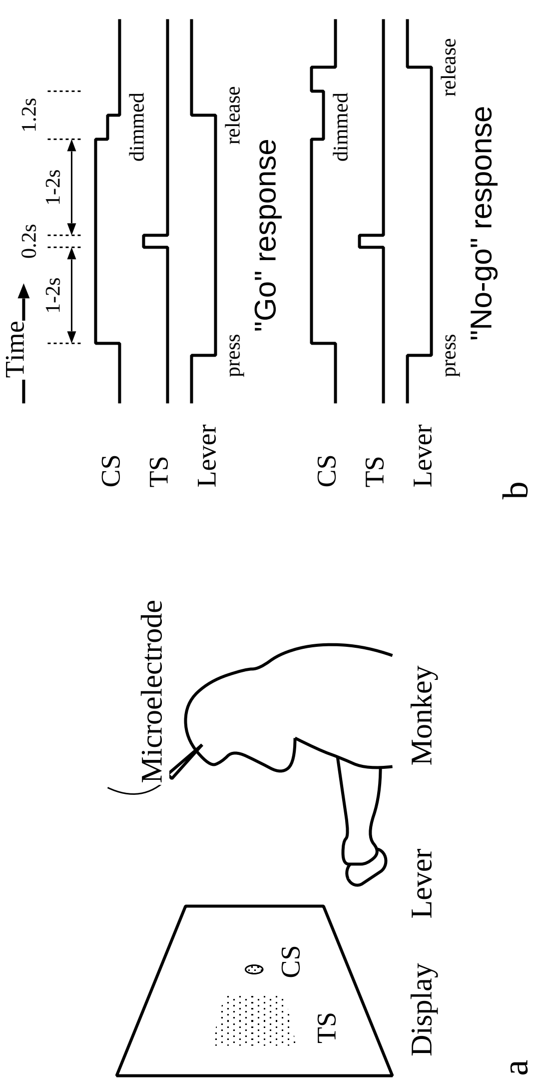

Figure 2. [upside ←]

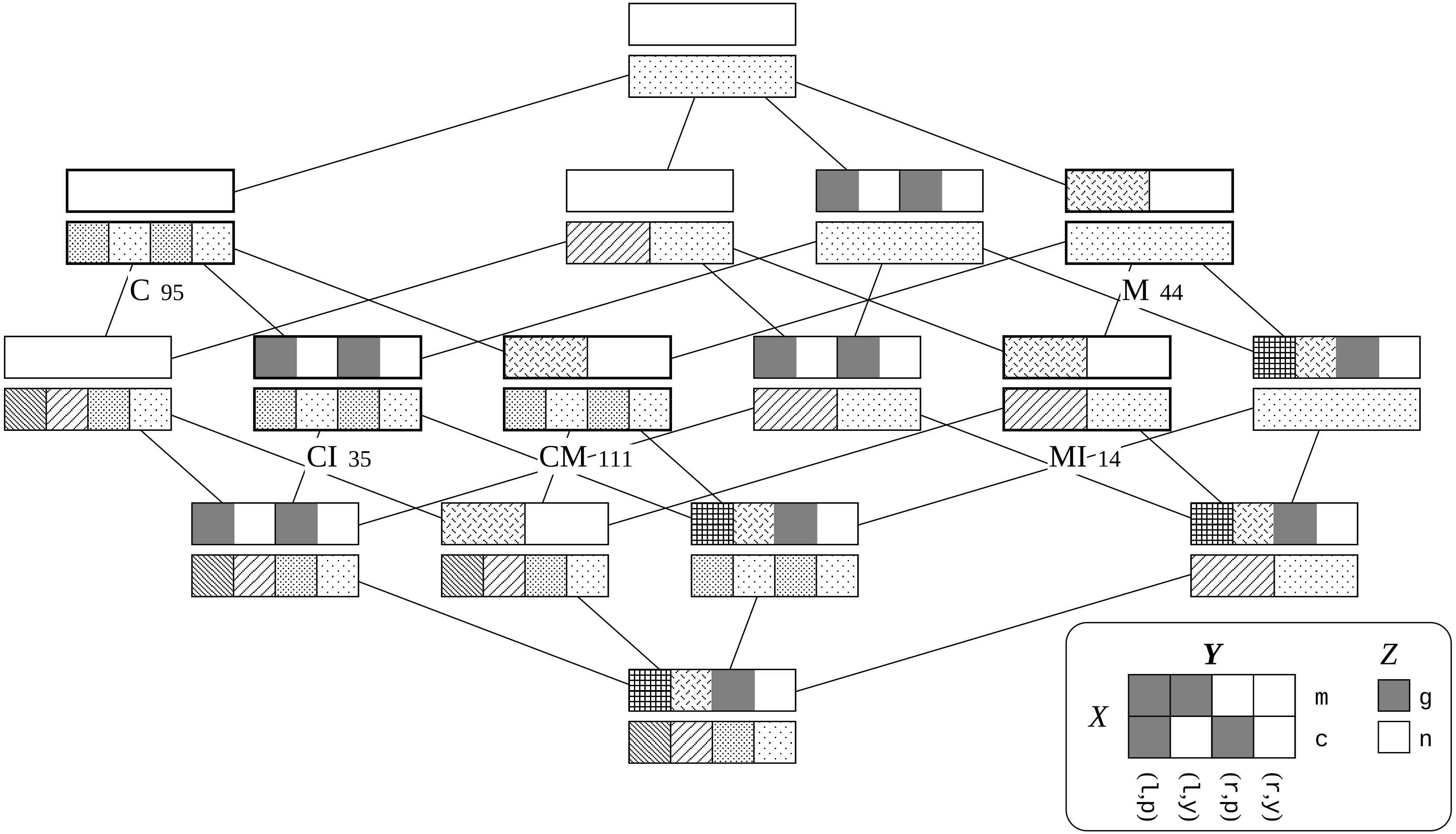

Figure 3. [upside ←]

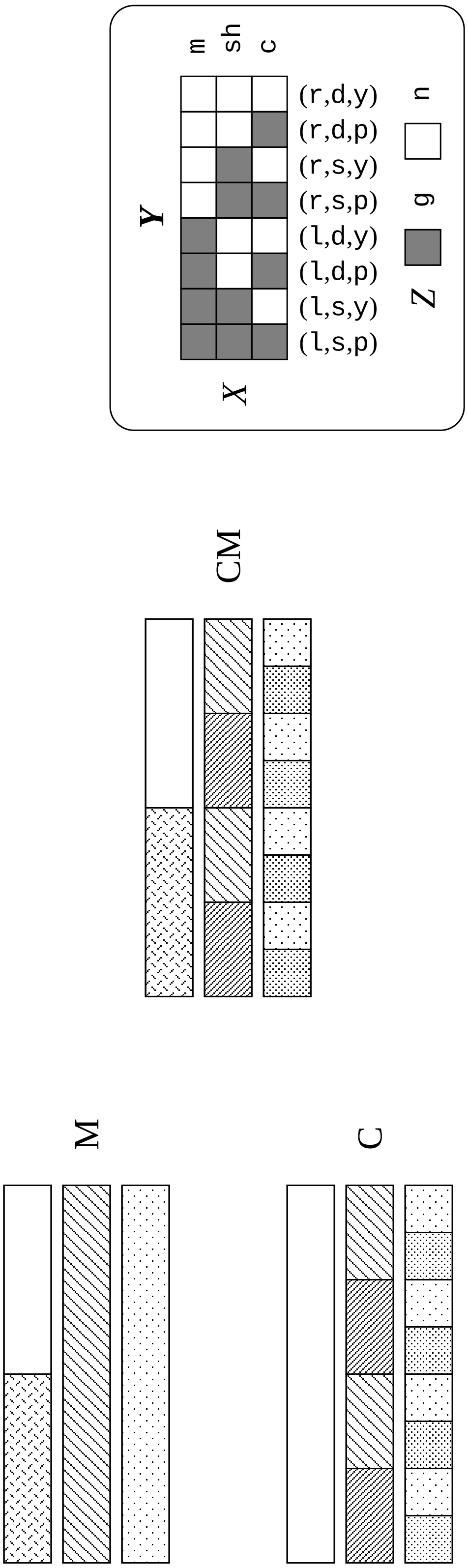

Figure 4. [upside ←]

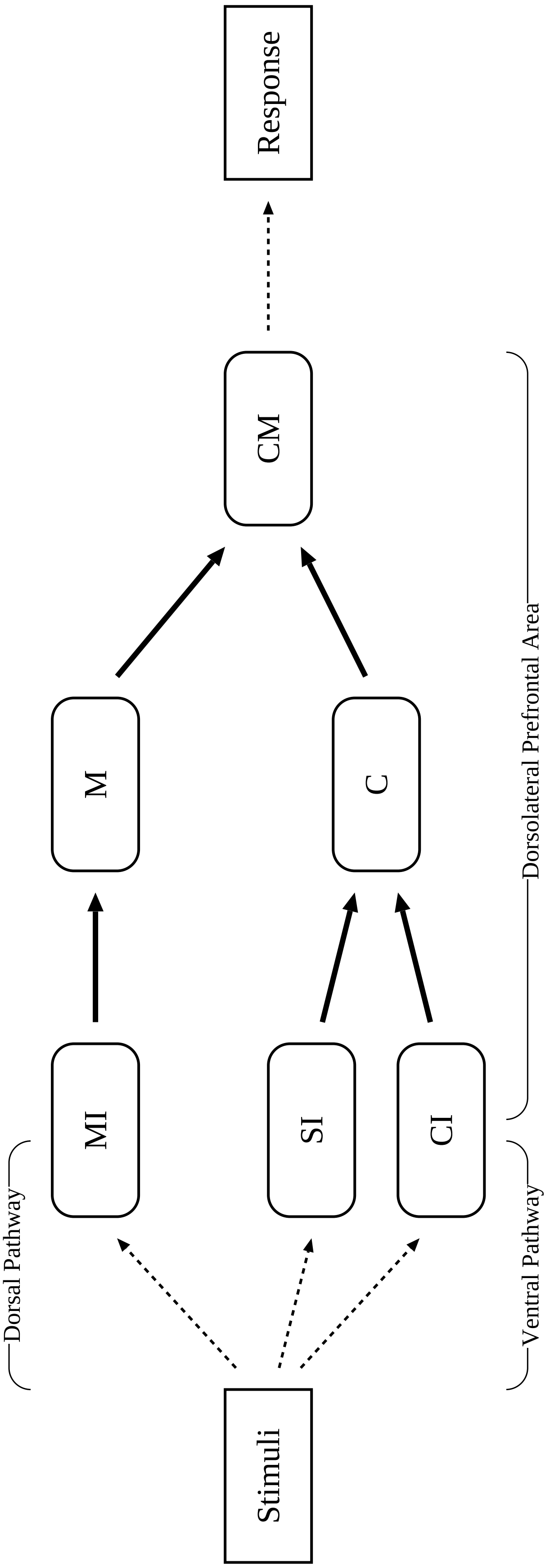

Figure 5. [upside ←]

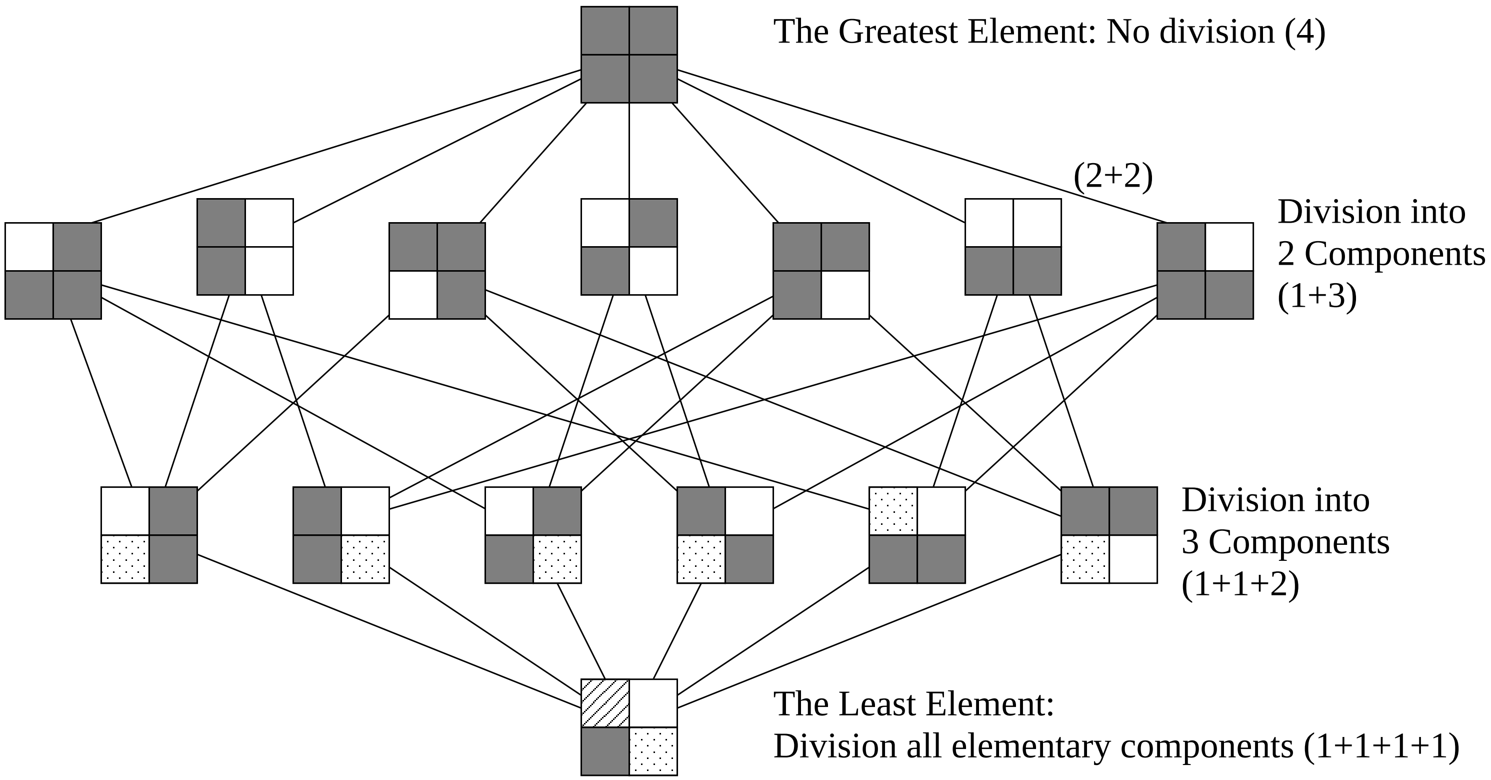

Figure 6. [upside ←]

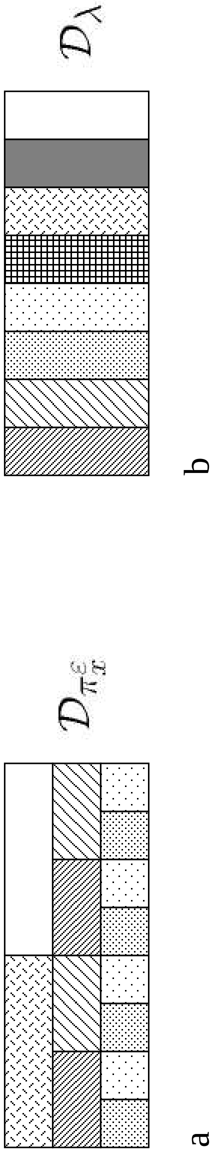

Figure 7. [upside ←]

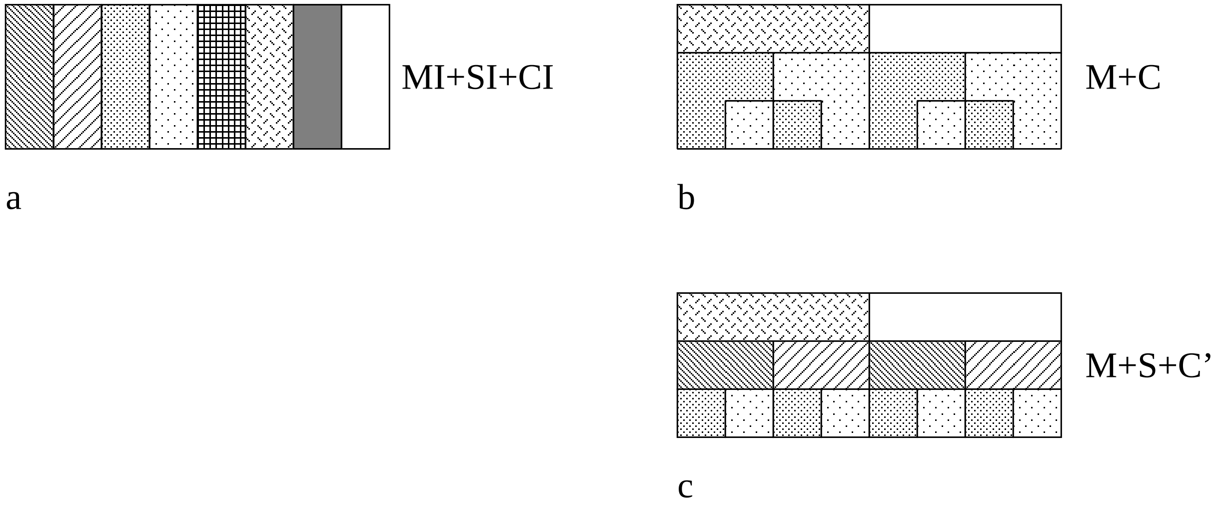

Figure 8. [upside ←]