Emergent communication through Metropolis-Hastings naming game with deep generative models

Tadahiro Taniguchi a†, Yuto Yoshida b†, Yuta Matsui b, Nguyen Le Hoang b, Akira Taniguchi a and Yoshinobu Hagiwa c

aCollege of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan; bGraduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan; cResearch Organization of Science and Technology, Ritsumeikan University, Kusatsu, Shiga, Japan

ABSTRACT
Constructive studies on symbol emergence systems seek to investigate computational models that can better explain human language evolution, the creation of symbol systems, and the construction of internal representations. Specifically, emergent communication aims to formulate a computational model that enables agents to build efficient sign systems and internal representations. This study provides a new model for emergent communication, which is based on a probabilistic generative model (PGM) instead of a discriminative model based on deep reinforcement learning. We define the Metropolis-Hastings (MH) naming game by generalizing previously proposed models. It is not a referential game with explicit feedback, as assumed by many emergent communication studies. Instead, it is a game based on joint attention without explicit feedback. Mathematically, the MH naming game is proved to be a type of MH algorithm for an integrative PGM that combines two agents that play the naming game. From this viewpoint, symbol emergence is regarded as decentralized Bayesian inference, and semiotic communication is regarded as inter-personal cross-modal inference. This notion leads to the collective predictive coding hypothesis regarding language evolution and, in general, the emergence of symbols. We also propose the inter-Gaussian mixture model (GMM) + variational autoencoder (VAE), a deep generative model for emergent communication based on the MH naming game. In this model, two agents create internal representations and categories and share signs (i.e. names of objects) from raw visual images observed from different viewpoints. The model has been validated on MNIST and Fruits 360 datasets. Experimental findings demonstrate that categories are formed from real images observed by agents, and signs are correctly shared across agents by successfully utilizing both of the observations of agents via the MH naming game. Furthermore, scholars verified that visual images were recalled from signs uttered by agents. Notably, emergent communication without supervision and reward feedback improved the performance of the unsupervised representation learning of agents.

1. Introduction

Constructive studies on symbol emergence systems, which are multi-agent systems that can make symbols or language emerge and use them for communication, are crucial for understanding human language and cognition, and for creating robots that can adapt to our semiotic communication [1–5]. Specifically, emergent communication aims to build a computational model that enables agents to build efficient sign systems, e.g. a list of words and a syntactic rule, and internal representations. Language (and symbol systems in general) features a dynamic nature. It changes dynamically through time in terms of forms and meanings. Despite the time-varying properties, symbol systems consistently enable individuals to communicate information about external objects. In other words, certain cognitive and social dynamics in multi-agent systems enable agents to form a symbol system in an emergent manner and offer a function of semiotic communication to the agents. The system is considered a complex system with an emergent property and referred to as symbol emergence system [4]. Importantly, such emerged symbols assist agents not only in communicating via signs but also in appropriately categorizing objects. A proper reciprocal reliance between sign sharing and perceptual category formation, including representation learning, is critical for the computational model of perceptual symbol systems [6].
Thesignificantchallengeinsemioticcommunication,thatis,communicationusingsigns,isanthatagentscanneitherinspecteachother’sbrainstatesnordirectlytransmit
meanings [7]. A symbol is a kind of sign emerging froma triadicrelationshipbetweenthe sign, the object, and the interpretant according to the terminologypopular in Peircean
semiotics [8]. The challenge in emergent communication isdevelopingnot onlyamodelfatenabletoformamodelsymbolsystemstowith bettersubversivcommunication
andcooperationbutalsoone thatcanexplainlanguagesimulationand symbolemergenceinhumans.
Thisnotionhasbeenalong-termchallengeinsymbol
e emergenceinrobotics [4].

Scholarsexaminedlanguagegames,suchasnaming
andreferentialgameswithexplicitfeedback,formodellingemergentcommunicationover time.Many stud-
iess inthisfieldwerebased onvariants of the Lewis
signalinggame [9]. Steels and related scholarsof artificial
life anddevelopmentalroboticsconductedawide
rang eofpioneering works as a synthetic approach to
language evolution [7,10–20]. They were not built upon
deep learning; hence, the constructive models of emerg-
et communication could not incorporate therepresenta-
tion learning of rawsensory data. Following the pub-
ication of key works by Foerster et al. [21] and Lazari-
ndou et al. [22], studies on emergent communication
have been revived. Many studies have been conducted
[23–31] due to the invention of deep reinforcement learn-
ing. The reason is that therepresentation-learning power
of deep learning is required to realize symbol emer-
gence based on raw sensory data. These models can
be regarded as emergent communication models based on
discriminativemodelsaccording to machine learning
perspectives.

However, as Tomaselloimportantlypointed out, this
typeofpointing-and-naminggamewithexplicitfeed-
back is not representative of the vastmajorityof word-
learning situations that children encounter in daily life
[32]. Therefore, language games with explicit rewards
orsupervisoryfeedbackarenotsuitable models from
the developmental point of view. In contrast, it is widely
known that a human infant holds the capability for joint
attention during the early developmental stage. This abil-
ity, where twoindividualssharetheir focus on an object
orrevent, serves as the foundation for language acquisi-
tion [33]. In other words, the assumption of joint atten-
tion is more plausible than the assumption of explicit
feedback in a language game from the developmental
perspective.

Atthesametime, generativemodelsarewidelyused
for modeling representation learning and concept forma-
tion based on multimodal sensory information [34–38].
In cognitive science and neuroscience, the generative
perspective of cognition, which is also known as the
free-energy principle and predictive coding, has become
dominant as a general principle of cognition [39–41]. The
world model-basedapproach to artificial intelligence also
follows this view [42,43].

With thiscontextin mind, this studypresentasnovel
emergent communication framework based on deep
probabilistic generative models (PGMs). We first define
the Metropolis-Hastings (MH) naming game. Hagiwara
et al. [44,45] initially introduced this type of game for a
specificprobabilisticmodel. In the currentstudy, the
MH naming game is generalized and formally defined
by generalizing the idea. The game appears to besimil-
to the original naming game; however, it is not. The
MH naming game does not require any explicit feedback
between agents but assumes the existence of joint atten-
tion inspired by developmental studies. The MH naming
game is completely based on PGMs and is mathemat-
ically demonstrated to be the same as the Metropolis-
Hastings algorithm for the model. The model represents
the generative process of the representation-learning and
sign-sharing processes of two agents as a whole. The
emergent communication is regarded as decentralized
Bayesian inference (see Theorem 2.1). Figure 1 pro-
vides an overview of the MH naming game. Semiotic
communication is defined as inter-personal cross-modal
inference when a speaker provides the name of a target
object, and a listener recalls the picture of the item from
the name.

The limitation of the models proposed by Hagiwara
et al. [44,45] is that they do not involve deep gener-
itive models and cannot enable agents to conduct symbol
emergence on raw images and to image ( i.e.recon-
struct) objects corresponding to signs. They also did
not provide a general theory for the MH naming game.
To address these aspects, the current study presents an inter-Gaussian mixture model (GMM) + variational auto encoder (VAE) or inter-GMM + VAE, a deep PGM, and an inference procedure for the model. The inference procedure is based on an MH naming game and a decomposition-and-communication strategy for modeling emergent communication based on deep probabilistic models [46].

The main contribution of this paper is twofold.

- By generalizing earlier studies, we establish the MH naming game to provide an emergent communication framework based on PGMs. In contrast to conventional language games, it assumes joint attention instead of explicit feedback between agents. We demonstrate that, in this framework, emergent communication is equal to the MH algorithm of a PGM that represents two agents. In other words, emergent communication is formulated as external and internal representation learning based on the decentralized Bayesian inference.

- We propose inter-GMM+VAE and its inference procedure as an MH naming game that enables two agents to undertake emergent communication, classify raw images and share signs that represent them in a cooperative manner. On two datasets, namely, MNIST and Fruits 360, we illustrate that emergent communication based on inter-GMM+VAE enables two agents to build categories and share signs at the same level as centralized inference.

2. Metropolis-Hastings naming game

The MH naming game is a form of language game played between two agents (Figure 1). In general, the game is played as follows. An agent views an object and tells the name based on its percept, that is, the internal state inferred from its observations. The agent says a word (i.e. a sign) corresponding to the object in a probabilistic manner (i.e. sampling a word from the posterior distribution over words). A counterpart, that is, a listener, determines whether or not it accepts the word based on its belief state. Afterward, they alternate their roles or take turns. This process does not involve explicit feedback from the listener to the speaker. In contrast, we assume joint attention, where the listener knows which object the speaker is looking at. In this section, we depict that the MH naming game can be derived as an approximate Bayesian inference procedure of a certain PGM that represents two agents as an integrative system.

The left panel in Figure 2 presents a PGM that integrates two PGMs that represent two agents with a latent variable $w_d$. This notion can be regarded as a PGM of a variant of multimodal VAEs [47]. When observing the $d$-th object from a different perspective at the same time, Agent $\star$ receives observations $o_d^\star$ and infers internal representation $z_d^\star$. Notably, $\star$ represents A or B throughout this study. The graphical model shown in Figure 2 left displays a latent variable $w_d$ shared by the two agents. In the context of multimodal VAEs, $w_d$ corresponds to a latent representation that integrates two modality information.
namely, visual and auditory information. From the viewpoint of a standard inference scheme, such as the Gibbs sampling and variational inference, information about \( z^A_d \) and \( z^B_d \) such as posterior distributions or samples in Monte-Carlo approximation are required to update \( w_d \). However, \( z^A_d \) and \( z^B_d \) are internal representations of Agents A and B, respectively. Therefore, each agent cannot look into the internal state of the other, which is the fundamental assumption of human semiotic communication. Metaphorically, if the brains of the two agents were connected, \( w_d \) would be an internal state of the connected brain and can be inferred by referring to the internal representations \( z^A_d \) and \( z^B_d \) of each agent. However, it is not the case. The question is whether or not agents can infer the shared variable \( w_d \) without connecting their brains, that is, without simultaneously referring to \( z^A_d \) and \( z^B_d \). Thus, playing the MH naming game is the solution.

Let us decompose the generative model into two parts following the symbol emergence in robotics toolkit (SERKET) framework (Figure 2 right) \([46,48]\). SERKET is a framework that enables the decomposition of a PGM into several modules and derives an inference procedure in a systematic manner. A total inference procedure can be divided into inter-module communication and intra-module inference, which is the same as the usual inference procedure of the elemental module \([46,48]\).

The graphical models corresponding to Agents A and B are structurally the same as PGMs for representation learning, such as VAEs. \( z^* \) is an internal representation of \( o^*_d \), which is inferred under the influence of the prior distribution, which has a variable \( w^*_d \). Here, lists \( o^* = \{ o^*_d \}_{d \in D^v}, w^* = \{ w^*_d \}_{d \in D^w}, w = (w_d)_{d \in D}, \) and \( z^* = \{ z^*_d \}_{d \in D^z} \) are defined, where \( D \) is the number of objects. In addition, \( o = \{ o^A, o^B \}, z = \{ z^A, z^B \}, \theta = \{ \theta^A, \theta^B \}, \) and \( \phi = \{ \phi^A, \phi^B \} \).

Let us regard the sampling process \( w^*_d \sim P(w^*_d|z^*_d, \phi^*) \) as the utterance of a sign \( w^*_d \). With this metaphorical assumption, the sampling of \( w^*_d \) can be regarded as a naming behavior for the object \( d \) by \( * \). Notably, \( w^*_d \) does not mean a latent variable for Agent \( * \), but a tentative sample for \( w_d \) drawn by Agent \( * \). The sign can be a word, a sentence, or even an image. With this assumption, the MH naming game is defined as follows.

Algorithm 1 for the PGM illustrated in Figure 2. The MH communication corresponds to a single communication event in the MH naming game as described in Algorithm 2\(^2\). The game consists of the following steps.

1. **Perception:** Speaker and listener agents (Sp and Li) observe the \( d \)-th object, obtain \( o^*_d \), and \( o^*_{Li} \) infers their internal representation \( z^*_d \) and \( z^*_{Li} \), respectively.

Algorithm 1 Metropolis-Hastings communication

```plaintext
1: procedure MH-communication(z\(Sp\), \(\phi\)\(Sp\), z\(Li\), \(\phi\)\(Li\), w\(d\))
2: \(w^Sp \sim P(w^Sp|z^Sp, \phi^Sp)\)
3: \(r = \min \left( 1, \frac{P(z^Sp|\phi^Sp, w^Sp)}{P(z^Sp|\phi^Sp, w^Sp)} \right) \)
4: \(u \sim \text{Unif}(0, 1)\)
5: if \(u \leq r\) then
6: return \(w^Sp\)
7: else
8: return \(w^Li\)
9: end if
10: end procedure
```

Algorithm 2 Metropolis-Hastings naming game

```plaintext
1: Initialize all parameters by sampling from prior distributions.
2: for \(t = 1\) to \(T\) do
3: // Agent A talks to Agent B.
4: for \(d = 1\) to \(D\) do
5: \(w^B \leftarrow \text{MH-communication}(z^A, \phi^A, z^B, \phi^B, w^d)\)
6: end for
7: // Learning by Agent B
8: \(\theta^B \sim P(\theta^B|z^B, \beta^B)\)
9: \(\phi^B \sim P(\phi^B|w^B, z^B, \alpha^B)\)
10: // Perception by Agent B
11: for \(d = 1\) to \(D\) do
12: \(z^B_d \sim P(z^B_d|\theta^B, \phi^B, w^B)\)
13: end for
14: // Agent B talks to Agent A.
15: for \(d = 1\) to \(D\) do
16: \(w^A_d \leftarrow \text{MH-communication}(z^B, \phi^B, z^A, \phi^A, w^d)\)
17: end for
18: // Learning by Agent A
19: \(\theta^A \sim P(\theta^A|o^A, z^A, \beta^A)\)
20: \(\phi^A \sim P(\phi^A|w^A, z^A, \alpha^A)\)
21: // Perception by Agent A
22: for \(d = 1\) to \(D\) do
23: \(z^A_d \sim P(z^A_d|\theta^A, w^A_d, \phi^A)\)
24: end for
25: end for
```

2. **MH communication:** Speaker tells the name \( w^Sp \) of the \( d \)-th object by sampling it from \( P(w_d|z^Sp, \phi^Sp) \). The listener determines if it accepts the naming with probability \( r = \min(1, \frac{P(z^Sp|\phi^Sp, w^Sp)}{P(z^Sp|\phi^Sp, w^Sp)}) \).
Appendix A.1). As a result, the MH naming game functions as a Metropolis-Hastings sampler of $P(w, z, \theta, \phi|\omega)$. In other words, the MH naming game is a decentralized approximate Bayesian inference algorithm.

We have demonstrated that the MH naming game is a decentralized approximate Bayesian inference method of a PGM that integrates two agents into a system (Figure 2 left). The MH naming game (in Algorithm 2) realizes the inference of $P(w|\omega^A, \omega^B)$ without inspecting each other’s brain states. Notably, the MH naming game naturally involves role alternation. In contrast to referential and original naming games [7], the MH naming game does not use any feedback, such as rewards or supervisory signals. Instead, it assumes joint attention, that is, the two agents know that they simultaneously look at the same object from different perspectives.

Surprisingly, the MH naming game satisfies the following property.

**Theorem 2.1:** The MH naming game is a Metropolis-Hastings sampler of $P(w, z, \theta, \phi|\omega)$.

**Proof:** Regarding $z_d^* \in z^* \in z$, $\theta^* \in \theta$, and $\phi^* \in \phi$, they are drawn from the distribution conditioned on the values of the remaining variables. As a result, they are considered drawn from the Gibbs sampler, which is a type of MH sampler. Regarding $w_d \in w$, if the proposal distribution $w_d^{Sp} \sim P(w_d^{Sp} | z_d^{Sp}, \phi^{Sp})$ is used, then the acceptance ratio of the MH algorithm [49] becomes $r = \min(1, \frac{P(z_d | w_d, \omega^A, \omega^B)}{P(z_d | w_d^{Sp}, \omega^A, \omega^B)})$, where $(Sp, Li) \in \{(A, B), (B, A)\}$ (see Appendix A.1). As a result, the MH naming game functions as a Metropolis-Hastings sampler of $P(w, z, \theta, \phi|\omega)$. In other words, the MH naming game is a decentralized approximate Bayesian inference algorithm.

Inference (3) Learning: After MH communication was performed for every object, the listener updates its global parameters $\theta^L$ and $\phi^L$.

(4) Turn-taking: The speaker and listener alternate their roles and go back to (1).

Figure 3 illustrates a probabilistic graphical model of inter-GMM+VAE and its decomposition.

**Figure 3.** Probabilistic graphical model of Inter-GMM+VAE and its decomposition.

![Probabilistic Graphical Model](image_url)

We define a deep generative model for two agents-emergent communication called inter-GMM+VAE. Figure 3 illustrates a probabilistic graphical model of inter-GMM+VAE. The probabilistic generative process of inter-GMM+VAE is shown as follows.

**3. Inter-GMM+VAE**

**3.1. Generative model**

We define a deep generative model for two agent-emergent communication called inter-GMM+VAE. Figure 3 illustrates a probabilistic graphical model of inter-GMM+VAE. The probabilistic generative process of inter-GMM+VAE is shown as follows.

\[
\begin{align*}
    w_d & \sim \text{Cat}(\pi) \quad d = 1, \ldots, D \\
    \mu_k^*, \Lambda_k^* & \sim \mathcal{N}(\mu_k^*, (\alpha \Lambda_k^*)^{-1}) W \\
    & \quad \times (\Lambda_k^* | \nu, \beta) \quad k = 1, \ldots, K \\
    z_d^* & \sim \mathcal{N}(z_d^* | \mu_{w_d}^*, (\Lambda_{w_d}^*)^{-1}) \quad d = 1, \ldots, D \\
    o_{d}^* & \sim p^{\theta^*}(o_{d}^* | z_d^*) \quad d = 1, \ldots, D
\end{align*}
\]

where $* \in \{A, B\}$; the parameters $\mu_k^*, \Lambda_k^*$ are parameters of the $k$-th multivariate normal distributions of Agent $*$, and $\phi^* = (\mu_k^*, \Lambda_k^*)_{k=K}$. The parameters are assumed to be generated using the normal-Wishart distribution. The latent variable $z_d^*$ shared by the GMM and VAE components is assumed to be drawn from a multivariate normal distribution corresponding to the $k$-th sign, that is, $w_d = k$. The discrete variable $w_d$ which represents a sign of the $d$-th object, is considered to be generated from the categorical distribution $\text{Cat}(w_d | \pi)$. In this
research, we assume that the mixture ratio \( \pi \) is a uniform distribution. Assuming that the observations \( o^*_d \) of each agent is generated from a VAE decoder \( p_{o^*}(o^*_d|z^*_d) \) with latent variable \( z^*_d \), the total generation process is described as above (Equations (1)–(4)). Notably, inter-GMM+VAE can be regarded as a variant of multimodal VAEs [47].

Figure 3 depicts a graphical model of inter-GMM+VAE and its composition and decomposition relationships. Inter-GMM+VAE is obtained by composing two GMM+VAE in a manner similar to that of inter-multimodal Dirichlet mixture (MDM) is obtained by composing two MDMs in [45]. GMM+VAE is obtained by combining GMM and VAE. Composing graphical models, particularly VAE with structured or discretized latent variables, is examined for combining the complementary characteristics of traditional PGMs, such as GMM, HMM, and LDA, with deep generative models such as VAE [51–54]. In this study, we simply combine GMM and VAE. The notation \(+\), that is, the composition of two graphical models and their mutual (or simultaneous) inference, follows the convention in [46].

We also call a generative model that consists of (1)–(3) inter-GMM, which is a tail-to-tail composition of two GMMs. In addition, inter-GMM+VAE can be considered a composition of inter-GMM and two VAEs. The primary distinction between our proposed Inter-GMM+VAE model and previous models, such as Inter-GMM, Inter-DM, and Inter-MDM [44,45], lies in the introduction of the VAE, as well as the extension of the MH naming game through mutual inference (MI) between VAE and GMM. This makes Inter-GMM+VAE the first model that allows the MH naming game to incorporate representation learning directly from raw sensor data, like images.

### 3.2. Inference via the MH naming game

As explained in Section 2, the MH naming game acts as a sampling-based inference procedure of inter-GMM+VAE. However, \( \theta^* \) and \( z^*_d \) cannot be drawn from the analytical posterior distribution, in contrast to inter-DM and inter-MDM in [44,45], because inter-GMM+VAE involves VAE, that is, a deep generative model. Moreover, gradient-based optimization throughout the system cannot be employed because \( w \) is assumed to be inferred through the MH naming game, that is, Markov Chain Monte Carlo (MCMC). As a result, we use the decomposition-and-communication method employed in the (Neuro-)SERKET framework [46,55]. Mutual inference is performed between GMM and VAE. The parameters of a GMM module \( (\mu_k^d, \Lambda_k^d) \) are sent to a VAE module, and each VAE is trained with data-dependent prior distribution \( \mathcal{N}(z^*_d|\mu_{w^*d}, (\Lambda_{w^*d})^{-1}) \). After the optimization of VAE, \( z^* \) is sent to the GMM module, \( \phi^* \) is inferred using Gibbs sampling, and \( w \) is sampled as an utterance in the MH naming game. This MI process enables the approximate sampling of the internal variables \( (z^*, \phi^*, \theta^*) \) of each agent. Appendix A.2 presents a diagram that depicts the overall MH naming game, that is, the inference procedure, for illustrative purposes.

### 3.3. Semiotic communication as an inter-personal cross-modal inference

Semiotic communication using sign \( w_d \) (see Figure 1) is divided into two parts. A speaker tells the name of an object \( d \) by sampling \( w^*_d \sim P(w_d|o^*_d) \), and a listener recalls an image of the object by sampling \( o^*_d \sim P(o^*_d|w^*_d = w^*_d) \). This process pertains to an ancestral sampling across the two agents.

If we consider inter-GMM+VAE as a variant of multimodal VAE, then two observations, namely, \( o^*_d \) and \( \phi^* \), are regarded as multimodal observations, as previously mentioned. The cross-modal inference is the process of inferring an observation from one modality to another in multimodal generative models (e.g., \( P(o^*_d|\phi^* \))). Furthermore, in inter-GMM+VAE, the cross-modal inference is performed across two agents. As a result, semiotic communication, in which the listener recalls an image from the sign \( w_d \) provided by the speaker, is considered an inter-personal cross-modal inference [45].

### 4. Experiment 1: MNIST dataset

#### 4.1. Conditions

**Dataset**: In this experiment, the MNIST dataset\(^4\) is used to validate the proposed model. The MNIST dataset consists of 28 × 28 pixels handwritten character images from 0 to 9. Agents A and B are assumed to observe the same object from different perspectives. In this experiment, we used raw MNIST data for the observations of Agent A, and MNIST data rotated 45° clockwise to the left for observations of Agent B. The total number of MNIST data used in this experiment was 10,000, with 1,000 MNIST data for each label. Figure 4 illustrates an example of the dataset used in this experiment.

**Compared method**: The proposed model, MH naming game (proposal), was assessed by comparing two baseline models and a topline model. In No communication (baseline 1), two agents independently form internal representations \( z \) and sign \( w \). No communication occurs between the two agents. In other words, the No communication model assumes two GMM+VAEs for Agents A and B and independently infers signs \( w^*_A \) and \( w^*_B \), respectively. All acceptance (baseline 2) is the same as the MH
naming game, whose acceptance ratio is always $r = 1$ in MH communication (MH-COM in Algorithm 2). Each agent always believes that the sign of the other is correct. In Gibbs sampling (topline), sign $w_d$ is sampled using the Gibbs sampler. This process directly uses $z^A_d$ and $z^B_d$, although no one can simultaneously examine the internal (i.e., brain) states of the two agents in human semiotic communication. As a result, the condition is not a model of emergent communication; instead, it is a topline as an inter-GMM+VAE centralized inference procedure.

Network architecture: Convolutional and deconvolutional neural networks were simply employed for an encoder and a decoder of VAE. Appendix A.3 presents the details.

Hyperparameters: The hyperparameters of inter-GMM+VAE were set to $\alpha = 1.0$, $m = 0$, $\beta = 0.05I$, and $\nu = 12$. The total number of signs was set to $K = 10$. The number of iterations of the MH naming game was $T = 100$. The dimension of the latent variables $z^A_d$ was set to 12, and the number of the training iterations of VAE for each update was set to 100. Adam, with a learning rate of 0.001, was used as an optimizer. The MI of VAE and GMM was conducted five times.

Evaluation criteria: ARI [56] was used to evaluate the unsupervised categorization performance of each agent through the MH naming game. An ARI close to 1 indicates high categorization performance, whereas an ARI close to 0 indicates low performance. In contrast to the precision calculated by comparing the estimated labels and ground truth labels, ARI can consider label-switching effects in clustering. The kappa coefficient $\kappa$ assessed the degree to which the two agents shared signs [57]. For more details, please refer to Appendix A.4.

Other conditions: Experiments 1 were conducted using an Intel Core i9-9900K CPU with $1 \times$ NVIDIA GeForce RTX2080 8GB GDDR6.

4.2. Result

Categorization and sharing signs: Table 1 presents the results of the ARI and the kappa coefficient values for each condition on the MNIST data. Figure 5 illustrates the confusion matrices of $w^A$ and $w^B$ for each condition. The vertical axis represents the ground truth indices, and the horizontal axis represents the estimated signs, which are ordered for viewing. The results demonstrate that the MH naming game leads two agents to categorize the objects at nearly the same level as the Gibbs sampling (topline), which is a centralized inference procedure. Additionally, symbols emerged and were used between the two agents. Interestingly, the MH naming game between two agents improved categorization without any additional supervision compared to the no communication conditions (i.e., perceptual categorization conducted by a single agent). This finding is regarded as an advantage of multimodal unsupervised learning. Inter-GMM+VAE is a multimodal extension of GMM+VAE as an integrative model, and the MH naming game is an approximated MCMC inference process. As a result, the MH naming game evidently utilizes various observations gathered from different agents to increase classification performance through the inference of $P(w_d|o^A_d, o^B_d)$. No communication, certainly, could not share signs and exhibited a worse categorization performance than the MH naming game. All acceptance could share signs to a certain extent. Although all acceptance attempts to make each agent mimic the use of signs of the other, the procedure did not result in their sharing of signs at the same level as the MH naming game. The reason is that each agent in all acceptance must accept signs produced by the other, whose categorization may be immature or even incorrect. As Figure 5 suggests, communication in the all acceptance condition did not address the confusion between categories 0, 5 and 6, whereas the MH naming game could. This is because all acceptance lacks a theoretical basis in Bayesian inference and cannot function as an approximate posterior sampler, i.e., an unsupervised clustering method. In terms of the MI of GMM and VAE, MI enhanced classification performance in each condition.

Imagination from signs: Figure 6 reveals images recalled from each emerged sign by each agent. The images corresponding to the sign $w$ were recalled by reconstructing observation $o$ from the mean vector of the $w$-th Gaussian distribution $\mu^w_o$. In the MH naming game, each agent successfully reconstructed each number. Different digits from the same sign $w$ were rebuilt by agents in no communication. In all acceptance, the agents could nearly imagine digits from signs. However, digits 4 and 9 led to slight confusion, which corresponds to labels 0 and 5, respectively in Figure 5 due to sorting.
Figure 5. Confusion matrices from Experiment 1. The vertical axis represents the index of the ground truth label, corresponding to the true MNIST label. The horizontal axis displays the index of the sign estimated by each agent.

Table 1. Experimental results for MNIST data: the means ± standard deviations of ARI and kappa coefficient $\kappa$ of 10 trials of are shown.

| Condition          | MI | ARI (Agent A) ± | ARI (Agent B) ± | $\kappa$ ± |
|--------------------|----|-----------------|-----------------|------------|
| MH naming game     | ✓  | 0.78 ± 0.04     | 0.78 ± 0.03     | 0.91 ± 0.02|
| MH naming game     | ✓  | 0.71 ± 0.04     | 0.72 ± 0.03     | 0.91 ± 0.03|
| No communication   | ✓  | 0.65 ± 0.04     | 0.68 ± 0.05     | 0.04 ± 0.04|
| No communication   |    | 0.60 ± 0.02     | 0.64 ± 0.03     | 0.01 ± 0.05|
| All acceptance     | ✓  | 0.68 ± 0.04     | 0.65 ± 0.03     | 0.81 ± 0.03|
| All acceptance     | ✓  | 0.61 ± 0.03     | 0.63 ± 0.05     | 0.83 ± 0.04|
| Gibbs sampling     | ✓  | 0.81 ± 0.03     | –               | –          |
| Gibbs sampling     | ✓  | 0.73 ± 0.04     | –               | –          |

Note: The highest scores are in bold, and the second-highest scores are underlined.

Figure 6. MNIST images recalled by each agent by signs.

Formation of internal representations: Figure 7 illustrates the latent variables $z^A$ and $z^B$ of VAE (i.e. internal representations of each agent). GMM and VAE with and without MI are shown to demonstrate the effect of MI. For visualization, principal component analysis (PCA) and t-SNE were employed [58]. The same color indicates the same digit. The figure suggests that mutual information (MI) in the VAE+GMM and Metropolis-Hastings (MH) naming game (i.e. MI across two VAE+GMM instances) influenced the distributions of internal representations, improving categorization performance, as shown in Table 1 and Figure 5.

Figure 8 displays the time course of the mean and standard deviation of the acceptance rate over 10 trials. For the MH naming game with MI, the results from the final MI iteration are shown, i.e. with the VAE trained using MI. The results indicate that with the VAE trained through MI, the two agents reached an agreement on sign usage more rapidly. This also suggests that MI leads to a better formation of internal representations.

5. Experiment 2: Fruits 360

5.1. Conditions

Dataset: To verify the proposed method on natural images, we used the Fruits 360 dataset. The Fruits 360 dataset consists of images of fruits and vegetables under a total of 131 categories with RGB channels and 100 × 100 pixels. We utilized raw Fruits 360 data for the observations of Agent A, and Fruits 360 data rotated 25° to the left for observations of Agent B, as in Experiment 1. This model assumes that the two agents were looking at the same objects from different viewpoints. This experiment employed a total of 2,350 Fruits 360 data points, with 235 Fruits 360 images used for each label. In this experiment, the study used 10 out of 131 categories (i.e. Corn Husk, Cherry Wax Red, Avocado, Corn, Raspberry, Pineapple, Eggplant, Lemon, Onion White, and Grape White 2). Figure 9 depicts the examples of the observations given to the agents in this experiment. The 10 categories were selected to ensure a discernible distinction among the fruits.

Additionally, to validate the dependency on the number of categories, the same experiment was conducted for
Figure 7. Visualization of the internal representations $z$ for MNIST data with (top) PCA and (bottom) t-SNE.

Figure 8. Time course of the acceptance rate during a trial under the conditions of the MH naming game, both with and without MI, on the MNIST dataset.

The MH naming game, No Communication, and Gibbs sampling using 20 and 40 categories from the Fruits 360 images. The number of signs $K$ for the two conditions was set to 20 and 40, respectively. In all conditions, MI was employed. The categories used are shown in Appendix A.5.

**Compared method:** In addition to the conditions used in Experiment 1, the study used inter-DM [44] and inter-GMM to assess the contribution of representation learning by VAE. Inter-DM and inter-GMM categorize using the Dirichlet mixtures and Gaussian mixture models, respectively. Observations were represented as bag-of-features (BoF) representations for inter-DM and inter-GMM. To obtain the BoF representations, we utilized the KAZE (KAnade-Lucas-Tomasi Zero-mean normalized cross correlation Enhanced) features [59], which capture local image feature representations. These KAZE features were then transformed into BoF representations using k-means clustering, with a codebook size of 200. The BoF representations were standardized for inter-GMM.

**Network architecture:** Convolutional and deconvolutional neural networks were employed simply as the encoder and decoder of the VAE, respectively. RGB image ($64 \times 64 \times 3$) was fed into the network as an input. Appendix A.3 illustrates the details.

**Other conditions:** The same hyperparameters and evaluation criteria in Experiment 1 were utilized. Data augmentation was employed to train the VAE. Each pixel
of the original image was normalized to [0, 1] and replicated three times. These replicas were then perturbed with Gaussian noise (zero mean, variance of 0.01) added to their pixel values.

### 5.2. Result

#### Categorization and sharing signs

Table 2 presents the results of the ARI and kappa coefficient values for each condition on the Fruits 360 dataset. Among the compared approaches, the MH naming game with MI, which is the proposed method, notably marked the highest score in ARIs. The performance of the MH naming game with MI was competitive with that of Gibbs sampling, which serves as the top-line method. We conducted a t-test and found no significant difference between the scores of the MH naming game with MI and Gibbs sampling in terms of ARI at the 5% significance level ($p = 0.12$ and 0.11 for A and B, respectively). According to the theory, the suggested approach and Gibbs sampling can sample from the same posterior distribution. This finding supports the theoretical implication (Theorem 2.1).

Considering $\kappa$, the all acceptance without and with MI took the first and second highest scores, respectively. This is because all acceptance compelled the agents to agree on the use of signs. The MH naming game with MI achieved a score of $0.97 \pm 0.01$ with probabilistic acceptance behavior, surpassing that of the MH Naming Game without MI. In all aspects, inter-GMM+VAE using MH naming game with MI outperformed both inter-DM and inter-GMM. This result demonstrates that VAE representation learning may identify acceptable representations for emergent communication.

Table 3 presents the results with an increased number of categories. The results indicate that under more challenging conditions, agents using the MH naming game were able to categorize objects more effectively than those using No Communication, achieving performance comparable to Gibbs sampling. The ARI decreased as the number of categories increased. This is because the number of similar categories grows as the count of objects, i.e., fruits, increases. Consequently, this renders the categorization task more difficult. See Appendix A.5. The results demonstrate that the MH naming game is effective for Inter-GMM-VAE even when the number of categories is larger than what the theory suggested.

Figure 10 displays the time course of the mean and standard deviation of the acceptance rate over 10 trials, in the same manner as Figure 8 for the MNIST dataset. The results exhibit the same trend as observed with the MNIST dataset. It was confirmed that MI leads to the formation of internal representations, enabling agents to share signs more readily.

#### Imagination from signs

Figure 11 presents images recalled from each emerged sign by each agent. In the MH naming game, each agent successfully recalled each fruit image. Alternatively, all acceptance recalls the same fruit for 7 and 9 and collapsed imagery for 2, 3, and 5.
This result is due to the inability of the agent to appropriately create internal representations and fruit categories. In the all acceptance condition, even though the agents achieved a high level of agreement in the usage of signs as indicated by $\kappa$, the clustering performance was inferior to that of the MH naming game. During the interaction, when a listener receives a name, they are compelled to merge observations from a fruit into a category to which they originally assigned the name. This can lead to unwarranted confusion in the categorization of objects. We believe that such a negative effect of all acceptance led agents to form ambiguous categories, causing the prototype of the category, i.e. the image corresponding to the mean vector, to become distorted.

The results reveal that the MH naming game on inter-GMM+VAE enabled two agents to cooperatively create internal representations, execute categorization, and share signs via the decentralized Bayesian inference.

6. Conclusion and discussion

This work detailed a new model for emergent communication based on a deep PGM. It defined the MH naming game was defined by generalizing prior works [44,45] and demonstrated that the MH naming game is the same as a form of MH algorithm for a PGM, which is an integrative model that combines two agents performing representation learning and participating in the naming game. From this viewpoint, symbol emergence and semiotic communication are regarded as decentralized approximate Bayesian inference and inter-personal cross-modal inference, respectively. To achieve emergent communication and symbol emergence based on raw images, the study proposed a deep generative model called inter-GMM+VAE. An MH naming game between two GMM+VAEs and MI between GMM and VAE comprised the inference process. Experiments using the MNIST and Fruits 360 datasets illustrated that the model enables two agents to simultaneously form internal representations and categories and share signs. Moreover, the study demonstrated that a listener could reconstruct appropriate images from the signs of a speaker.

Theoretical extensions: The proposed generative model-based approach to emergent communication is relatively generic and leaves potential for future expansions. In inter-GMM+VAE, the sign $w$ is assumed to be a categorical variable (i.e. a discrete sign). However, the MH naming game itself does not restrict $w$ as a categorical variable. A conceivable path is extending $w$ to word sequences while considering compositionality. The number of sign types, which correspond to Gaussian components, is fixed in inter-GMM+VAE. To render it flexible, using Bayesian nonparametrics (e.g. Dirichlet process GMM) is a possible solution [60–62]. In addition, the generative model for an agent can be replaced with other sophisticated models. The current study employed GMM+VAE for simplicity. It is known that a multinomial VAE performs object categorization. For example, using a multimodal VAE instead of the unimodal VAE is one possible extension [54]. Another task is to investigate improved models and network architecture. Another problem is extending the MH naming game from a two-agent party to an N-agent game.

Collective predictive coding hypothesis: The MH naming game suggests that humans can collectively categorize objects by integrating different perceptions. This integration occurs when we name objects based on individual perceptions and probabilistically decide whether to accept or reject another’s naming based on our beliefs during representation learning. In general, PGMs are designed to predict observations by inferring representations (latent variables). In contrast, predictive coding is
viewed as a theory for understanding brain function [40]. This perspective leads to the collective predictive coding hypothesis [63]. This hypothesis suggests that by fusing the fragmented sensory observations of each participant, we can collectively infer latent representations, leading to a deeper understanding of the world. Furthermore, the hypothesis argues that symbol systems, especially language, emerge from this collective predictive coding process. From a representation learning perspective, in the Inter-GMM+VAE model, the VAE forms internal representations under the top-down probabilistic constraints imposed by the GMM. At the same time, external representations, such as signs or language, are constructed in a bottom-up manner by the MH naming game. This process could be called social representation learning. It is intriguing to consider whether humans use the MH naming game in a similar way when they create and share symbols. Okumura et al. have shown that human subjects in experimental semiotic studies do, to some extent, conform to the acceptance probability calculated by MH communication [64]. Another promising line of research is the study of symbol emergence in humans, based on the theoretical framework discussed in this paper.

**Notes**

1. Code available at https://github.com/is0383kk/Symbol Emergence-VAE-GMM.
2. In Algorithm 2, agents alternate turns after naming all objects. Alternatively, agents could swap roles for each individual object. From the perspective of the MCMC algorithm, this variation relates to the order in which latent variables are sampled. While multiple possibilities exist to dictate this order, we have chosen a fixed sequence for simplicity in this paper.
3. Galke et al. found that role alternation and memory restriction are critical for human language evolution but are frequently overlooked by contemporary studies that use referential game research [50].
4. The MNIST database: http://yann.lecun.com/exdb/mnist/.
5. The performance seemingly remains relatively lower than the existing emergent communication models based on deep reinforcement learning. Notably, however, they use explicit feedback and negative samples to improve discriminability (i.e. discriminative models). We consider that their assumption is not natural from the developmental perspective, although specific performance measures could be improved. Therefore, the direct comparison between their and our assumptions is not straightforward.
6. Even though Gibbs sampling stands as a topdown method, its ARI hovers around 0.8. It is important to note that the ARI is calculated by comparing unsupervised clustering results to the ground truth, which, in this case, are human annotations. While Gibbs sampling is theoretically considered the proper method for appropriate posterior inference, it doesn’t guarantee that the clustering results will mirror human categorization tendencies. In our study, we aimed to see if the inference outcome of the MH naming game matches that of Gibbs sampling. We operated under the assumption that both the MH naming game and Gibbs sampling serve as sampling methods, pulling variables from an identical posterior distribution. Consequently, the remaining classification discrepancies between Gibbs sampling and the MH naming game are considered to originate from a shared root cause.

7. Fruits360: https://www.kaggle.com/moltean/fruits.
8. We used slightly different settings, i.e. 25 degrees instead of 45 degrees, to test for the existence of degree-specific effects, since the method should at least be independent of the particular degree of rotation.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This study was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [grant numbers JP21H04904, JP23H04835, and JP18K18134] and by the MEXT Grant-in-Aid for Scientific Research on Innovative Areas 4903 (Co-creative Language Evolution) [grant number 17H06383].

**Notes on contributors**

Tadahiro Taniguchi received his Ph.D. degree from Kyoto University, in 2006. From 2005 to 2008, he was a research fellow of the Japan Society for the Promotion of Science. From 2008 to 2010, he was an assistant professor at the Department of Human and Computer Intelligence, Ritsumeikan University. From 2010 to 2017, he was an associate professor in the same department. From 2015 to 2016, he was a visiting associate professor at the Department of Electrical and Electronic Engineering, Imperial College London. Since 2017, he has been a professor at the Department of Information Science and Engineering, Ritsumeikan University, and a visiting general chief scientist at Panasonic (Holdings) Corporation. He has been engaged in research on machine learning, emergent systems, and symbol emergence in robotics.

Yuto Yoshida received his B.E. and M.E. degrees from Ritsumeikan University, Japan, in 2020 and 2022 respectively. His research interests include symbol emergence and artificial intelligence.

Yuta Matsui received his B.E. degree from Ritsumeikan University, Japan, in 2023. His research interests include symbol emergence and machine learning.

Nguyen Le Hoang earned his B.E. degree in 2013 from the University of Information Technology – VNUHCM, Vietnam. He went on to obtain his M.E. degree in 2020 from the University of Technology – VNUHCM, Vietnam. His primary research interests lie in symbol emergence and artificial intelligence.

Akira Taniguchi received his Ph.D. degrees from Ritsumeikan University, Japan, in 2018. From 2017 to 2019, he was a Research Fellow of the Japan Society for the Promotion of Science. From 2019 to 2022, he was a Specially Appointed Assistant Professor at the College of Information Science and
Engineering, Ritsumeikan University, Japan. Since 2022, he is currently a Lecturer at the same department. His research interests include intelligent robotics, artificial intelligence, and symbol emergence in robotics. He is also a member of RSJ, JSAI, ISCIE, and SICE.

Yoshinobu Hagiwara received his Ph.D. degree from Soka University, Japan, in 2010. From 2010 to 2013, he was an Assistant Professor at the Department of Information Systems Science, Soka University. From 2013 to 2015, he was a Specially Appointed Researcher at the Principles of Informatics Research Division, National Institute of Informatics. From 2015 to 2018, he was an Assistant Professor at the Department of Human & Computer Intelligence, Ritsumeikan University. From 2018 to 2021, he was a Lecturer at the same department. He is currently an Associate Professor at the Research Organization of Science and Technology, Ritsumeikan University. His research interests include intelligent robotics, human-robot interaction, machine learning, and symbol emergence in robotics. He is a member of IEEE, RSJ, IEEJ, JSAI and SICE.

ORCID

Tadahiro Taniguchi © http://orcid.org/0000-0002-5682-2076
Akira Taniguchi © http://orcid.org/0000-0003-0678-1103

References

[1] Steels L. The synthetic modeling of language origins. Evol Commun J. 1997;10.1.
[2] Deacon TW. The symbolic species: the co-evolution of language and the brain. New York: W. W. Norton & Company; 1998.
[3] Taniguchi T, Uğur E, Hoffmann M, et al. Symbol emergence in cognitive developmental systems: a survey. IEEE Trans Cogn Dev Syst. 2018;11(4):494–516. doi: 10.1109/TCDS
[4] Taniguchi T, Nagai T, Nakamura T, et al. Symbol emergence in robotics: a survey. Adv Robot. 2016;30(11–12): 706–728. doi: 10.1080/01691864.2016.1164622
[5] Taniguchi T, El Hafi L, Hagiwara Y, et al. Semiotically adaptive cognition: toward the realization of remotely-operated service robots for the new normal symbiotic society. Adv Robot. 2021;35:1–11.
[6] Barsalou LW. Perceptual symbol system. Behav Brain Sci. 1999;22:277–660.
[7] Steels L. The talking heads experiment: origins of words and meanings. Berlin: Language Science Press; 2015.
[8] Chandler D. Semiotics the basics. Oxfordshire: Routledge; 2002.
[9] Lewis D. Convention: a philosophical study. Toronto: John Wiley & Sons; 2008.
[10] Steels L, Kaplan F. AIBO’s first words: the social learning of language and meaning. Evol Commun. 2000;4(1):3–32. doi: 10.1075/eoc.4.1.03ste. Available from: https://www.jbe-platform.com/content/journals/10.1075/eoc.4.1.03ste
[11] Steels L. A self-organizing spatial vocabulary. Artif Life. 1995;2(3):319–332. doi: 10.1162/artl.1995.2.319
[12] Steels L, Belpaeme T. Coordinating perceptually grounded categories through language: a case study for colour. Behav Brain Sci. 2005;28(4):469–489. doi: 10.1017/S0140525X05000087
[13] Steels L, Loetzsch M. Perspective alignment in spatial language. In: Coventry KR, Tenbrink T, Bateman JA, editors. Spatial language and dialogue. Northamptonshire: Oxford University Press; 2008.
[14] Spranger M. The evolution of grounded spatial language [Ph.D. thesis]. Brussels: Vrije Universiteit Brussel; 2011.
[15] Spranger M, Loetzsch M, Steels L. A perceptual system for language game experiments. In: Steels L, Hild M, editors. Language grounding in robots. New York: Springer; 2012. p. 89–110.
[16] Spranger M. The evolution of grounded spatial language. Vol. 5, in Computational models of language evolution. Berlin: Language Science Press; 2015.
[17] Vogt P. The physical symbol grounding problem. Cogn Syst Res. 2002;3(3):429–457. doi:10.1016/S1389-0417(02)00051-7. Available from: https://www.sciencedirect.com/science/article/pii/S1389041702000517
[18] Vogt P. The emergence of compositional structures in perceptually grounded language games. Artif Intell. 2005;167(1):206–242. doi:10.1016/j.artint.2005.04.010. Available from: https://www.sciencedirect.com/science/article/pii/S0004370205001001
[19] De Beule J, De Vylder B, Belpaeme T. A cross-situational learning algorithm for damping homonymy in the guessing game. In: Artificial Life. 2006;X:466–472.
[20] Bley L. Language strategies for the domain of colour, Computational models of language evolution, Berlin: Language Science Press, 2015.
[21] Foerster JN, Assael YM, de Freitas N, et al. Learning to communicate with deep multi-agent reinforcement learning. In: International Conference on Neural Information Processing Systems (NIPS); 2016. p. 2145–2153.
[22] Lazaridou A, Pysakhovich A, Baroni M. Multi-agent cooperation and the emergence of (natural) language. In: The International Conference on Learning Representations (ICLR); 2017.
[23] Graesser L, Cho K, Kiela D. Emergent linguistic phenomena in multi-agent communication games. In: Conference on Empirical Methods in Natural Language Processing (EMNLP); 2019.
[24] Chaabouni R, Strub F, Altché F, et al. Emergent communication in semantic spaces. In: Advances in Neural Information Processing Systems (ICLR); 2021.
[25] Lazaridou A, Baroni M. Emergent multi-agent communication in the deep learning era; 2020. ArXiv preprint arXiv:200602419.
[26] Tucker M, Li H, Agrawal S, et al. Emergent discrete communication in semantic spaces. In: Advances in Neural Information Processing Systems (NeurIPS); 2021.
[27] Mu J, Goodman N. Emergent communication of generalizations. In: Advances in Neural Information Processing Systems (NeurIPS); 2021.
[28] Jiang J, Lu Z. Learning attentional communication for multi-agent reinforcement learning: intention sharing. In: International Conference on Learning Representations (ICLR); 2020.
[29] Kim W, Park J, Sung Y. Communication in multi-agent reinforcement learning: intention sharing. In: International Conference on Learning Representations (ICLR); 2020.
[30] Li F, Bowling M. Ease-of-teaching and language structure from emergent communication. Adv Neural Inf Process Syst (NeurIPS). 2019;32.
[31] Eccles T, Bachrach Y, Lever G, et al. Biases for emergent communication in multi-agent reinforcement learning. Adv Neural Inf Process Syst (NeurIPS). 2019;32.

[32] Tomasello M. Constructing a language: a usage-based theory of language acquisition. Massachusetts: Harvard University Press; 2005.

[33] Cangelosi A, Schlesinger M. Developmental robotics: from babies to robots. Massachusetts: MIT Press; 2015.

[34] Kingma DP, Welling M. Auto-encoding variational bayes; an introduction. In: International Conference on Learning Representations (ICLR); 2013. ArXiv preprint arXiv:1312.6114.

[35] Suzuki M, Nakayama K, Matsuo Y. Joint multimodal learning with deep generative models. In: International Conference on Learning Representations (ICLR); 2016.

[36] Taniguchi A, Hagiwara Y, Taniguchi T, et al. Improved and scalable online learning of spatial concepts and language models with mapping. Auton Robots. 2020;44:1–20.

[37] Nakamura T, Nagai T, Iwahashi N. Grounding of word meanings in multimodal concepts using lda. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2009. p. 3943–3948.

[38] Nakamura T, Nagai T, Funakoshi K, et al. Mutual learning of an object concept and language model based on MLDA and NPYLM. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2014. p. 600–607.

[39] Friston K. The free-energy principle: a unified brain theory? Nat Rev Neurosci. 2010;11(2):127–138. doi:10.1038/nrn2787

[40] Hohwy J. The predictive mind. Northamptonshire: Oxford University Press; 2013.

[41] Ciria A, Schillaci G, Pezzulo G, et al. Predictive processing in multimodal cognitive robotics: a review. Neural Comput. 2021;33(5):1402–1432. doi:10.1162/neco_a_01383

[42] Friston K, Moran RJ, Nagai Y, et al. World model learning and inference. Neural Netw. 2021;144:573–590. doi:10.1016/j.neunet.2021.09.011. Available from: https://www.sciencedirect.com/science/article/pii/S0893608021003610

[43] Okada M, Kosaka N, Taniguchi T. PlaNet of the Bayesians: reconsidering and improving deep planning network by incorporating bayesian inference. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020. p. 5611–5618.

[44] Hagiwara Y, Kobayashi H, Taniguchi A, et al. Symbol emergence as an interpersonal multimodal categorization. Frontiers in Robotics and AI. 2019;6:134. Available from: https://www.frontiersin.org/article/10.3389/frobt.2019.00134

[45] Hagiwara Y, Furukawa K, Taniguchi A, et al. Multiagent multimodal categorization for symbol emergence: emergent communication via interpersonal cross-modal inference. Adv Robot. 2022;36(5–6):239–260. doi:10.1080/01691864.2022.2029721

[46] Taniguchi T, Nakamura T, Suzuki M, et al. Neuro-serket: development of integrative cognitive system through the composition of deep probabilistic generative models. New Gener Comput. 2020;38:1–26.

[47] Suzuki M, Matsuo Y. A survey of multimodal deep generative models. Adv Robot. 2022;36(5–6):261–278. doi:10.1080/01691864.2022.2035253

[48] Nakamura T, Nagai T, Taniguchi TS. An architecture for connecting stochastic models to realize a large-scale cognitive model. Front Neurorobot. 2018;12:25. doi:10.3389/fnbot.2018.00025

[49] Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57:97–109. doi:10.1093/biomet/57.1.97

[50] Galke L, Ram Y, Raviv L. Emergent communication for understanding human language evolution: what’s missing? In: Workshop on Emergent Communication at ICLR; 2022.

[51] Johnson MJ, Duvenaud DK, Wiltshko A, et al. Composing graphical models with neural networks for structured representations and fast inference. Adv Neural Inf Process Syst (NIPS). 2016;29.

[52] Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning. In: Advances in Neural Information Processing Systems (NIPS); 2017.

[53] Razavi A, Vinyals O. Generating diverse high-fidelity images with VQ-VAE-2. In: Advances in Neural Information Processing Systems (NeurIPS); 2019. p. 14866–14876.

[54] Kuniyasu R, Nakamura T, Taniguchi T, et al. Robot concept acquisition based on interaction between probabilistic and deep generative models. Frontiers in Computer Science. 2021;3:3.

[55] Nakamura T, Nagai T, Taniguchi TS. An architecture for connecting stochastic models to realize a large-scale cognitive model. Front Neurorobot. 2018;12:25. doi:10.3389/fnbot.2018.00025. Available from: https://www.frontiersin.org/article/10.3389/fnbot.2018.00025

[56] Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(1):193–218. doi:10.1007/BF01908075

[57] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46. doi:10.1177/00131644600200104

[58] Van der Maaten L, Hinton G. Visualizing data using t-sne. J Mach Learn Res. 2008;9(11):2579–2605.

[59] Alcantarilla PF, Bartoli A, Davison AJ. KAZE features. Adv Robot. 2020;34(17):1003610.

[60] Teh YW, Jordan MI, Beal MJ, et al. Hierarchical Dirichlet process. In: Frontiers in Statistical Learning. Berlin: Springer; 2006. p. 600–607.

[61] Blei DM, Jordan MI. Variational inference for Dirichlet process mixtures. Bayesian Analysis. 2006;1(1):121–143. doi:10.1214/06-BA104

[62] Nakamura T, Nagai T, Iwahashi N. Multimodal categorization by hierarchical Dirichlet process. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2011. p. 1520–1525.

[63] Taniguchi T. Collective predictive coding hypothesis: symbol emergence as decentralized bayesian inference; 2023. PsyArXiv preprint psyarxiv.com/d2yz6.

[64] Okumura R, Taniguchi T, Hagiwara Y, et al. Metropolis-Hastings algorithm in joint-attention naming game: experimental semiotics study; 2023. ArXiv preprint arXiv:230519936.

[65] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174. doi:10.2307/2529310.
Appendix

A.1 Acceptance rate in the MH naming game

Given \( P(z) \) and \( Q(z^* | z) \) are target and proposal distributions, respectively, where \( z^* \) is a proposed sample, then the acceptance rate of MH algorithm is as follows:

\[
R(z, z^*) = \min(1, \frac{P(z^*) Q(z | z^*)}{P(z) Q(z^* | z)})
\]  
(A1)

\[
\bar{R}(z, z^*) = \frac{P(z^*)}{P(z)} \frac{Q(z^* | z)}{Q(z | z^*)}
\]  
(A2)

In the MH naming game, the target distribution for \( w_d \) is \( P(w_d | z^*_p, z^*_l, \phi^*_p, \phi^*_l) \) and the proposal distribution is \( P(w_d | z^*_p, \phi^*_p) \).

\[
\bar{R}(w^*_l, w^*_p)
\]  
(A3)

As a result, the acceptance rate in the MH naming game satisfies the condition of the MH algorithm.

---

**Figure A1.** Overall procedure of MH naming game for Inter-GMM+VAE.
A.2 Overall procedure of the MH naming game for inter-GMM+VAE

Figure A1 presents a diagram depicting the overall MH naming game (i.e. inference procedure). In Figure A1, $\psi^*$ denotes the parameter of the inference network.

A.3 Network architecture

Figure A2 displays the network structure of VAE used in Experiment 1. $H$ and $W$ represent the height and width of the image, respectively; $K$ denotes the number of kernels, $C$ pertains to the number of filters in the output, and $S$ stands for the number of strides. The layers Conv, Conv Transposed, Linear, ReLU and Sigmoid denote the convolution, transposed convolution, and fully-connected layers, the ReLU function, and the Sigmoid function, respectively. Figure A3 illustrates the network structure of VAE used in Experiment 2 in the same manner.

A.4 ARI and kappa coefficient

ARI is extensively used to evaluate clustering performance by comparing clustering results with ground truth labels. In contrast to the precision calculated by comparing the estimated labels and ground truth labels for evaluating the classification systems trained by supervised learning, ARI can consider label-switching effects in clustering. ARI is defined by the following equation (A9). Notably, RI denotes RandIndex.

$$\text{ARI} = \frac{RI - \text{Expected RI}}{\text{Max RI} - \text{Expected RI}}$$ (A9)

For more details, please refer to [56].

A.5 Fruit 360 categories used in experiment 2

Table A1 presents sample images from the Fruit 360 dataset used in Experiment 2. For the 10, 20, and 40 category conditions, we utilized fruits with IDs ranging from 1–10, 1–20, and 1–40, respectively. The fruit names adhere to the nomenclature used in the original Fruit 360 dataset.
**Table A1.** Summary of Fruits 360 dataset used in Experiment 2.

| ID | Name               | ID | Name               | ID | Name               | ID | Name               |
|----|--------------------|----|--------------------|----|--------------------|----|--------------------|
| 1  | Avocado            | 11 | Banana             | 21 | Apple Pink Lady    | 31 | Mandarine          |
| 2  | Cherry Wax Red     | 12 | Blueberry          | 22 | Cactus fruit       | 32 | Mangostan          |
| 3  | Corn               | 13 | Cauliflower        | 23 | Cantaloupe 1       | 33 | Nectarine Flat     |
| 4  | Corn Husk          | 14 | Fig                | 24 | Carambula          | 34 | Passion Fruit      |
| 5  | Eggplant           | 15 | Ginger Root        | 25 | Chestnut           | 35 | Pear Abate         |
| 6  | Grape White 2      | 16 | Kohlrabi           | 26 | Clementine         | 36 | Pear Stone         |
| 7  | Lemon              | 17 | Lychee             | 27 | Cucumber Ripe      | 37 | Pepper Red         |
| 8  | Onion White        | 18 | Pepper Green       | 28 | Granadilla         | 38 | Pitahaya Red       |
| 9  | Pineapple          | 19 | Walnut             | 29 | Guava              | 39 | Rambutan           |
| 10 | Raspberry          | 20 | Watermelon         | 30 | Hazelnut           | 40 | Tomato not Ripened |