Individuation Through Integration: A PID-Based Study of Androids and Insect Societies

Takashi Ikegami, Suzune Baba, Takahide Yoshida, and Hiroki Kojima
Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan
ikeg@sacral.c.u-tokyo.ac.jp

Abstract

This paper explores how individual agency and self-recognition can emerge from collective dynamics through the lens of Community First Theory (CFT), which holds that structured collectives precede and scaffold the formation of individuality. We present two case studies: (1) a humanoid android (Alter3) equipped with multimodal sensors and large language models that achieves self-modeling through coordinated module interaction, and (2) honeybee colonies in a planar artificial hive, where the opening of a foraging door triggers the spontaneous emergence of behavioral roles such as foragers, dancers, and observers. Using Partial Information Decomposition (PID), we quantify the informational structure of both systems, showing that key components contribute not just uniquely or redundantly, but synergistically to global behavior. We interpret this synergy as a functional marker of individuation. Furthermore, we propose the concept of ecological alignment—a top-down constraint imposed by the collective environment that channels the behavior of individual modules or agents—bridging embodied AI and social insects under a unified framework of collective intelligence.

1 Introduction

Most theories of individuality in collective systems begin with well-defined agents and study how group behaviors emerge from their bottom-up interactions. In contrast, our research begins with collectives: how structured communities give rise to individuated roles, functional units, and eventually agents with apparent autonomy. We propose a theoretical framework called the Community First Theory (CFT), which posits that individuality does not precede the collective—it emerges from it.

We introduce the concept of second-order individuality, which refers to agent-like structures that are spatially distributed and temporally extended. Unlike traditional "first-order" agents, which are typically defined by strong internal drives and autonomy, second-order individuals are formed from structured interactions within a collective. The motivation, agency, and selfhood of such individuals are not innate but are dynamically generated by the organization of the community itself. Ultimately, these second-order individuals feed back into the collective, regulating the behavior of the components that gave rise to them. This feedback is where alignment occurs: not as a top-down imposition, but as an emergent constraint shaped by the collective organization. In this paper, we explore how such alignment arises in both artificial and biological systems through the lens of synergy and role individuation.

To illustrate this idea, we present two contrasting yet complementary case studies: (1) an artificial android system, Alter3 [MMI21, YMI25], which develops a self-model through multimodal integration of sensory input, memory, and self-reflection via large language models (LLMs)[YMM+23], and (2) a honeybee colony composed of genetically identical individuals, in which role differentiation (e.g., foragers, dancers, pioneers) spontaneously emerges after the colony is allowed to interact with the external world [DDI23].

To quantitatively analyze the emergence of second-order individuality, we apply an information-theoretic framework based on Partial Information Decomposition (PID)[WB10, KBO+20, LRM+24]. PID allows us to decompose mutual information between subsystems into redundant, unique, and synergistic components. This decomposition enables us to identify which "elements" in a system contribute cooperatively to global patterns—whether these are physical roles in bee colonies or cognitive

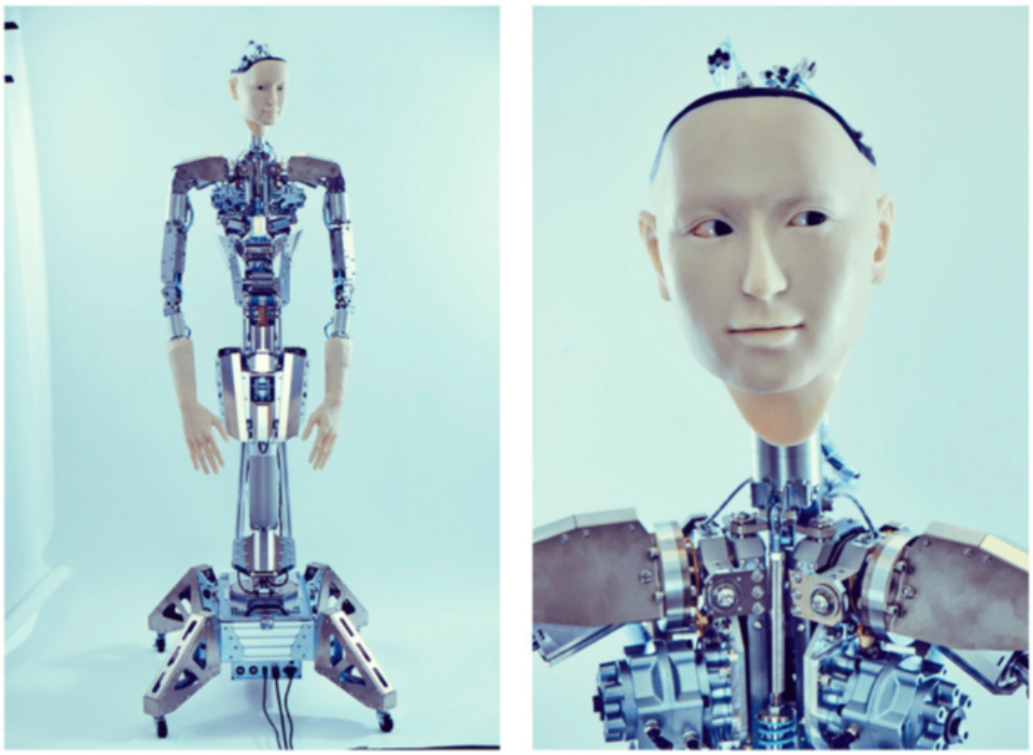

Figure 1: Appearance of the humanoid android ALTER3. The left panel shows a full-body view, highlighting its mechanical structure and articulated limbs. The right panel presents a close-up of the face, designed to evoke a subtle human likeness while revealing its robotic mechanisms.

modules in androids. These theoretical questions motivated our design, which integrates proprioception, vision, and memory into a single embodied architecture. While traditional robotic systems often treat modalities as independent inputs to be fused or weighted, our findings suggest that self-recognition requires higher-order information that emerges only when these modalities are combined. This perspective is formalized through Partial Information Decomposition (PID), which allows us to separate unique, redundant, and synergistic contributions to self-judgment. In particular, memory can sometimes produce negative unique information—indicating conflict or contradiction when it operates alone—yet plays a constructive role when integrated with other modalities[Inc17]. This points to the crucial role of multimodal synergy in stabilizing self-identification and supports the view that selfhood is not a property of individual signals, but of their interaction. Our work thus bridges embodied robotics, large language model reasoning, and information theory, offering a computational framework for investigating minimal selfhood in artificial agents.

2 Methods

2.1 Overview

This paper primarily introduces experiments conducted on the android ALTER 1[MMI21], focusing on the question of whether "self-recognition" can emerge within it. Following this, we present an analysis of data from social insects.

Alter3 is powered by the air compressor. It has vision from the webcam in thier eyes and auditory devices from the microphone. As for the current project We equipped ALTER3 with bend sensors, and integrate it with a multimodal LLM-based architecture [YBMI24]. .

Namely, ALTER3 possesses visual input (via webcam and pose estimation), proprioceptive sensing (via joint bending sensors), and episodic-like memory (via LLM-generated reflective summaries), enabling a form of self-referential processing analogous to human self-perception. Leveraging these

three modalities, ALTER3 attempts to infer agency over a perceived hand by autonomously executing finger movements, evaluating visual-proprioceptive congruence, and integrating outcomes from prior sensorimotor episodes.

2.2 System Architecture and Experimental settings

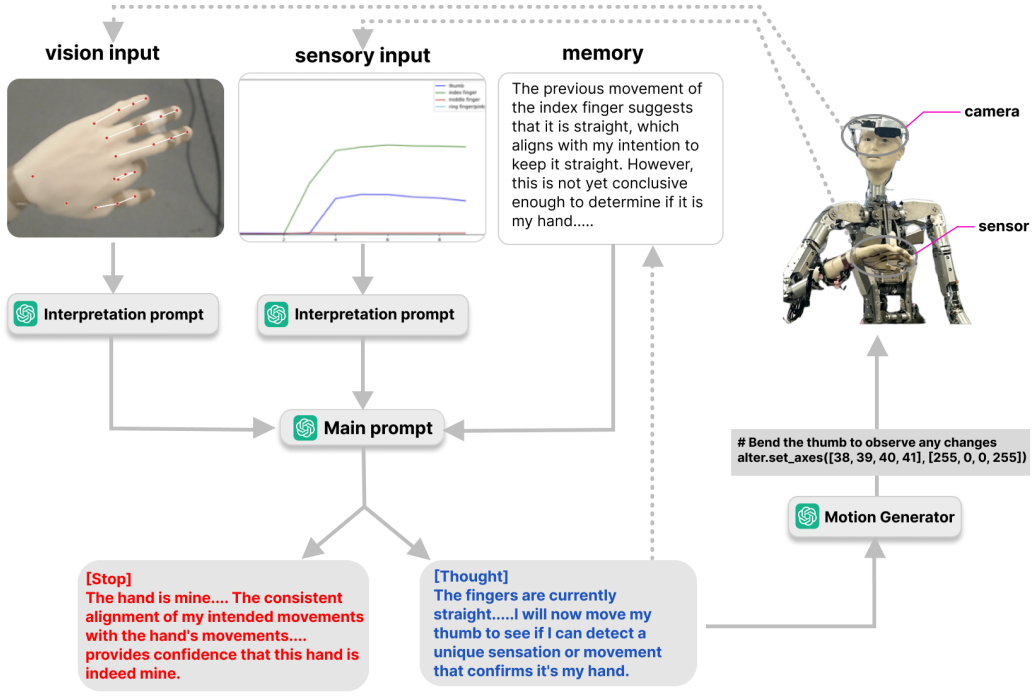

Figure 2: Experimental System. The vision input from the camera and the sensory input from the bending sensors are interpreted by the Interpretation Prompt and added to the Main Prompt. Based on this information, the main task is executed, and depending on the contents of its reasoning, the Motion Generator produces Python code to drive Alter3's movements. After the action, the information is reinterpreted, and the prior reasoning is accumulated as memory, repeating this process iteratively.

The experimental schema is as follows. We used OpenAI's GPT-4o model, which is capable of processing image inputs. The model received three types of information as input: (1) Proprioception, (2) Vision, and (3) Memory.

- **Proprioception**: Four bend sensors were attached to the thumb, index, middle, and ring/pinky fingers of Alter3's right hand. Sensor values were collected via Arduino and normalized for interpretation. The output values of each bending sensor were plotted as graphs, which were input to GPT-4o as images. LLM was asked to infer the hand's state from the sensor data. The prompt used was:"Your task is to determine the state of the hand from the sensor values shown in the graph."
- **Vision**: The camera captured first-person RGB images. Finger joint angles were estimated from these images using MediaPipe, a framework developed by Google. These joint angles were then mapped to finger positions and visualized as images, which were input into GPT-4o. LLM was asked to infer the state of the hand. The prompt used was: "Your task is to determine whether each finger is bent or straight."
- **Memory**: All intermediate reasoning steps (i.e., prompts and their outputs) were accumulated in a memory buffer. Up to the five most recent steps were retained, and this accumulated context was included in subsequent prompts to the LLM.

These three information sources were jointly input into the Main Prompt, which determine whether the hand was thier own. LLM can produce either [Thought] command (to continue reasoning) or [Stop] command (to end the trial).

When outputting [Thought], the model was required to describe the next action (i.e., which fingers to move) and the reasoning behind that choice. The generated action was passed to the motion generation module, which executed the corresponding movement. The resulting sensorimotor feedback was then observed through sensors and the camera, and the LLM updated its internal belief accordingly. When the model reached sufficient confidence, it ended the trial and issued a final judgment.

An excerpt of the Main Prompt is shown below:

> Your task is to determine whether it is your hand there. Output instructions: Two types of output are possible: [Thought] or [Stop]. You are free to move your fingers of your hand. The ring and little fingers move in the same way. Continue verifying until certainty is achieved. If you want to check by moving, first output [Thought] and write the reason and the next action. Once you have gained confidence in your answer, write your reasons and conclusions after [Stop].

Temperature was set to 0.0 for interpretation prompts and 1.0 for main task prompts to simulate deliberation and uncertainty.

The experiment was conducted under the following 14 conditions:

- Real Hand: All information that correspond to the actual hand movements were used. The Trials were conducted under seven conditions by varying the combinations of available information: Vision, Proprioception, and Memory(VPM), Vision and Proprioception(VP), Proprioception and Memory (PM), Vision and Memory(VM), Vision only(V), Proprioception only(P) and Memory only(M)
- Fake Hand: All information were inconsistent with the actual actions performed. Even in this case, trials were conducted under the same seven conditions: VPM, VP, PM, VM, V, P and M

All trials were repeated $N = 10$ times per condition.

3 Results

The results of the experiment using ALTER3 are presented below. We conducted the previously described spontaneous trial-and-error task under all eight possible combinations of the three modalities: Vision (V), Proprioception (P), and Memory (M). If ALTER3 successfully inferred that its real hand was its own and correctly identified the fake hand as not its own, we labeled the outcome as state 1 (correct inference); all other outcomes were labeled as state 0.

In the second part, we analyze these inference results using Partial Information Decomposition (PID).

3.1 Confidence Dynamics over Iterations

The temporal evolution of confidence scores as Alter3 iteratively evaluated whether the observed hand was its own (Figure 3). Each line corresponds to a single trial, with confidence levels plotted across the number of reasoning steps (i.e., thoughts). The confidence score (ranging from 0 to 100) assigned by a GPT model to its own textual outputs at each step.

The red lines represent the tests under the *real hand* condition, and the blue lines correspond to the *fake hand* condition.

A clear difference in dynamics emerges between conditions. In the real-hand trials, confidence scores tended to increase steadily, with many trajectories converging toward high certainty (above 90) within 2–4 iterations. By contrast, in the fake-hand trials, confidence scores often stagnated or declined over time, reflecting uncertainty or conflict in multimodal interpretations. Some trials terminated prematurely with low confidence, indicating failure to reach a coherent self-model.

The divergence in trajectories suggests that Alter3's self-recognition process is modulated by the coherence of sensorimotor feedback. When interacting with its real hand, proprioceptive and visual

signals reinforce each other, leading to rising confidence. In contrast, fake-hand conditions produce conflicting or noisy inputs, hindering convergence.

These patterns align with the accuracy statistics reported earlier, and further support the role of iterative, embodied reasoning in enabling reliable self-recognition.

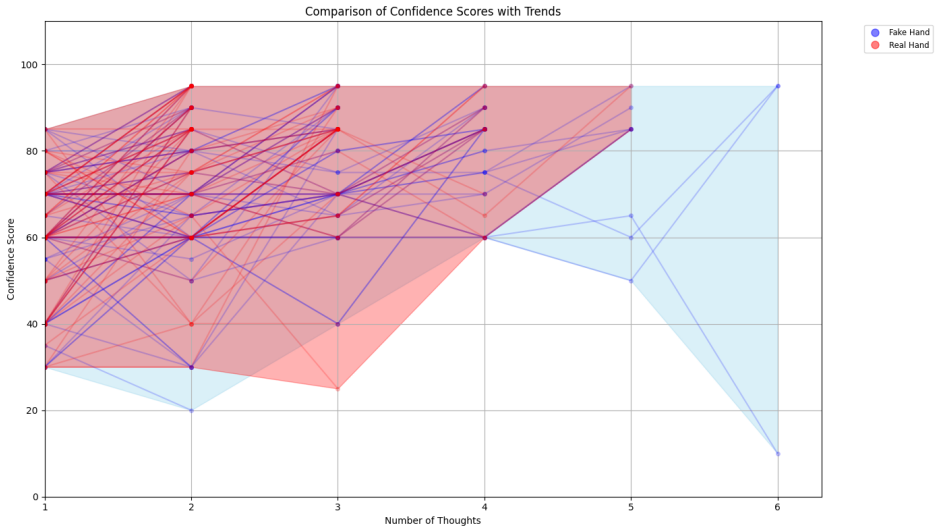

Figure 3: Temporal evolution of confidence scores across reasoning steps. The x-axis represents the number of thoughts (i.e., successive stages of reasoning), and the y-axis indicates the confidence score (ranging from 0 to 100) assigned by a GPT model to its own textual outputs at each step. The red region illustrates the distribution of confidence trajectories for trials involving the model's correct inference that the observed hand was its own. The blue region corresponds to trials involving incorrect inference (i.e., a fake hand). Confidence levels tend to increase steadily across steps when the hand is correctly recognized, in contrast to the more variable and generally lower confidence observed in incorrect trials.

To quantify the observed divergence in confidence dynamics, we compared confidence scores at the first and final reasoning steps (Step 1 vs Step 6). As shown in figure 3, both conditions started from similar confidence levels, but diverged significantly by the final step.

3.2 Judgment Accuracy

When Alter3 had access to all three modalities—vision, proprioception, and episodic memory (VPM)—it consistently identified its own hand correctly.

To systematically evaluate the contribution of each sensory modality, we conducted 10 trials under seven conditions: (1) VPM, (2) vision + proprioception (VP), (3) proprioception + memory (PM), (4) memory + vision (MV), (5) vision only (V), (6) proprioception only (P), and (7) memory only (M).

In real-hand trials (Figure **??**, left), body ownership was attributed in 9/10 trials under VPM and VP, 8/10 under PM and P, 7/10 under V, 5/10 under MV, and only 3/10 under M. This suggests that proprioception plays a critical role, while memory alone or memory combined with vision leads to weaker self-attribution.

In fake-hand trials (Figure **??**, right), misattributions occurred in 5/10 trials under M, 2/10 under V, 1/10 under VP, MV, and P, and never under VPM or PM.

These contrasts reveal a clear synergistic effect: accuracy with all three modalities surpassed that of any single modality, and the vision–proprioception pair outperformed either channel in isolation. Hence, reliable self-recognition in Alter3 emerges not from individual cues but from their cross-modal integration.

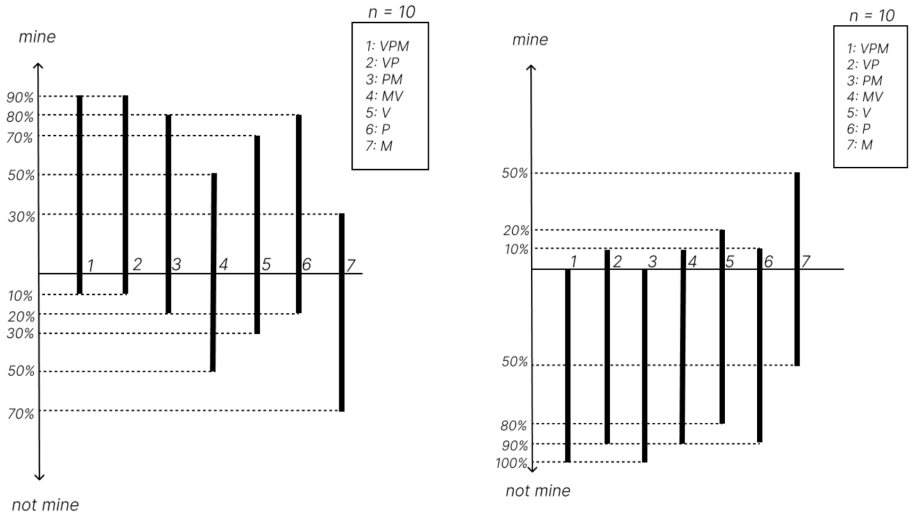

Figure 4: Proportions of body ownership judgments under seven sensory conditions. Left: real-hand trials. Right: fake-hand trials. Each bar reflects the percentage of trials in which the android judged the observed hand as "mine" (upward) or "not mine" (downward). Sensory conditions: 1 = VPM (vision, proprioception, memory), 2 = VP, 3 = PM, 4 = MV, 5 = V, 6 = P, 7 = M. Trials ended when the android voluntarily committed to a final judgment.

3.3 Modal Contribution

To quantify the role of each sensory modality in self-recognition, we computed the mutual information (MI) between the android's final judgment (real vs. fake) and the available sensory inputs. We considered each modality individually—vision (V), proprioception (P), and memory (M)—as well as their pairwise combinations (VP, PM, MV) and the full tri-modal combination (VPM).

Among the individual modalities, proprioception contributed the most information, with an MI substantially higher than that of vision or memory alone. Vision on its own yielded relatively low MI, suggesting that visual input is insufficient for accurate body ownership judgments unless combined with proprioceptive feedback. Memory alone provided the least information and in some cases introduced ambiguity, likely due to its reliance on past episodic associations that may not correspond to the current sensorimotor context.

Pairing vision with proprioception (VP) significantly increased MI compared to either modality alone, highlighting a synergistic interaction between exteroceptive and interoceptive signals. The PM and MV combinations also improved information content, but to a lesser extent. The highest MI was achieved when all three modalities—vision, proprioception, and memory—were integrated (VPM), indicating that cross-modal integration is critical for robust self-recognition.

These results are summarized in Figure 5, which visualizes the MI values across all modality conditions. The figure clearly illustrates that no single modality is sufficient, and that the combination of sensory inputs plays a key role in enhancing judgment accuracy.

3.4 Synergy and Redundancy

To further dissect the nature of multimodal integration, we applied Partial Information Decomposition (PID) to partition the joint MI into unique, redundant, and synergistic components. Figure 6 presents a comparison between two PID methods—Williams & Beer (WB) and the I_{ccs} approach. In both decompositions, synergy—information available only when all three modalities are combined—was prominent. Notably, I_{ccs} yielded higher total synergy than WB, reinforcing the hypothesis that self-recognition in Alter3 involves emergent, higher-order cross-modal relationships. Conversely, redundancy and some unique information terms varied across methods, highlighting methodological sensitivity in quantifying distributed informational structures.

These results support the view that body ownership and self-recognition are not reducible to any single modality, but arise from the dynamic integration of distributed sensory and memory sources.

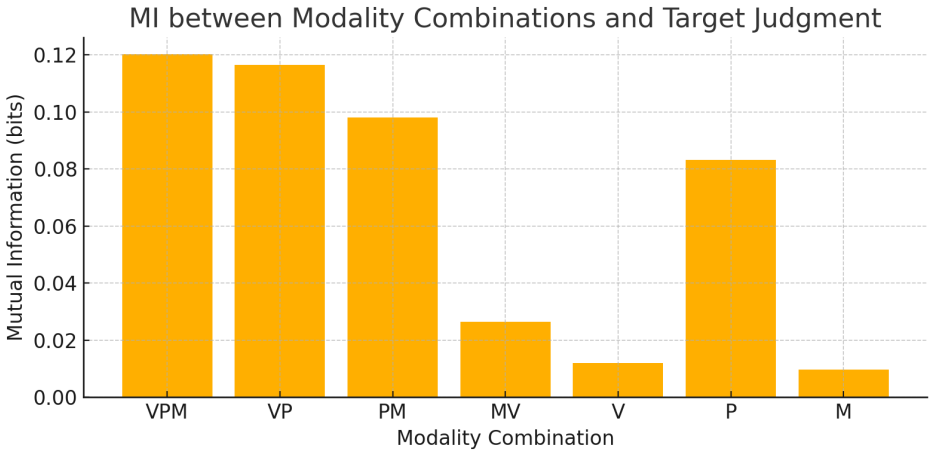

Figure 5: Mutual information (MI) between the target judgment (real vs. fake) and various combinations of sensory modalities. Full multimodal integration (VPM) yields the highest MI, while vision or memory alone contributes relatively little.

4 Comparative Insights from Biological Collectives

Life is not something that can be assembled merely by putting together component parts like LEGO blocks. This reflects the holistic nature of living systems. In the case of ALTER, simply combining devices does not result in a functioning system. What matters is how these elements are integrated. We observed that a network of information mediated by a large language model (LLM) achieved this integration without a predefined blueprint—through a synergy effect. But what about biological systems that form real communities? Using the example of an artificial bee hive, we examine how a cluster of individual bees begins to function as a cohesive colony—interpreted here as an instance of synergy.

We examined analogous dynamics in honeybee colonies, where collective behavior leads to emergent role differentiation. In a previous study [GRM+18, DDI23], approximately 1,000 adult worker bees from a single-cohort population, along with a naturally mated queen, were housed in an artificial hive constructed as a single-layer, two-dimensional planar array within a transparent box. Unlike typical multilayer hives, this setup allowed unobstructed visual access to the full colony at all times.

Each individual bee was uniquely identified using a 2D QR code (bCode) affixed to the thorax, enabling precise tracking of both position and orientation every second over the course of a 7-day period. The experiment was conducted under controlled conditions: the hive was kept in a dark and quiet room, and its glass surface was cleaned twice daily to ensure high detection accuracy across day and night cycles. There were no larvae or pupae in the hive, and thus no brood care was observed.

A key feature of the design was the delayed opening of the hive entrance. For the first two days, the door remained closed, preventing any external foraging. On the third day, the entrance was opened, and the worker bees began exploring the outside environment and returning to the hive with information and resources. To describe the collective behavior of honeybee colonies, we focused on the emergence of various roles. In particular, we examined: (D) bees performing the waggle dance; (F) bees observing the dance from the front row; (K) the total kinetic energy of the hive, representing the overall level of synchronized activity; (P) a subset of bees that initiate increases in kinetic energy; and (O) the number of bees outside the hive.

Before the entrance was opened, bees already displayed synchronized, spontaneous bursts of movement—what we term *endogenous bursts*, arising without external stimuli. Following the opening, functional role differentiation rapidly progressed: foragers, dancers, observers, and other behavioral specializations emerged, transforming the hive into a coherent superorganism.

To quantify this transition, we measured the mutual information between overall hive activity and the appearance of dancers and bursts, then applied the same Partial Information Decomposition (PID) methodology used for Alter3. In this framework, the hive itself corresponds to the "self," while individual bees assume specialized roles akin to the sensory and memory subsystems in the android.

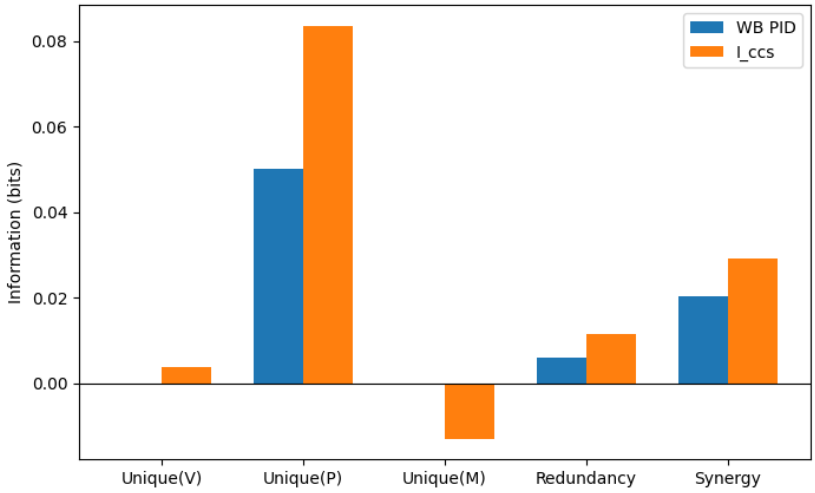

Figure 6: Comparison of PID components calculated using the Williams–Beer (WB PID) and I_{ccs} methods. Synergy and redundancy were computed with respect to the judgment state—whether the android correctly inferred that the observed hand was its own (true inference) or not (false inference). The three information sources were V (vision), P (proprioception), and M (memory). Negative values indicate suppression or conflict within individual modalities.

Our analysis revealed that bursts—and particularly the bees initiating them (pioneer bees)—were strongly correlated with the hive's global activity. Using the time series data for these variables (recorded in seconds, though analyzed in minute-level resolution), we divided the data into windows of 1000 minutes each. For each window, we computed the frequency distributions of K, D, F, and P, and then calculated the mutual information between these distributions and the hive state, defined as the level of foraging activity, measured by O. From this, we further computed the redundancy and synergy in the information structure of the hive. In terms of mutual information, the level of activity K showed a strong correlation with the hive activity O. Synergy with respect to hive activity was indeed observed among P, K, and D. Figure ** illustrates how the amount of this synergy varies over time across individual bursts.

PID showed that these roles contributed synergistically rather than redundantly, indicating that the transformation was not a mere accumulation of behaviors, but a functional differentiation within the collective. Informational synergy increased significantly after the hive was opened, paralleling the pattern observed in Alter3, where multimodal integration strengthened coherent decision-making.

This convergence between artificial and biological systems supports the *Community First Theory (CFT)*: structured collectives provide the scaffolding from which individual agency and functional differentiation emerge. Synergy, in both cases, marks not just an increase in information processing, but a signature of individuation itself.

5 Discussion

In ALTER's setup, LLMs communicate through natural language as both input and output, they enable general-purpose interaction among functional modules. This architectural affordance has led to increasing interest in orchestrating multiple LLMs through structured prompts for multi-agent reasoning. Recent studies such as Project-Sid and Lyfe Agent [AAB+24, GCS+24] have explored this direction. In Project-Sid, each LLM module is stateless and writes its output to a shared database; inter-module communication occurs through data retrieval and storage operations. Lyfe Agent places greater emphasis on memory structures, incorporating both short-term and long-term memory.

This architecture allows for flexible and adaptive system behavior, and offloads complex reasoning processes to cloud-based LLM APIs, thereby reducing dependence on local computation. We have

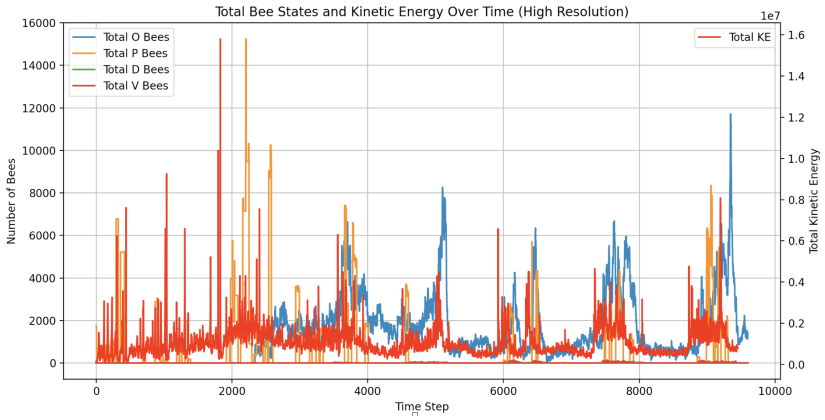

Figure 7: Temporal dynamics of bursting behavior. After approximately 30 hours, the hive entrance opens, allowing bees to exit. As some bees do not return, the total number of bees gradually decreases. We plotted the following five variables over time: (D) the number of bees performing the waggle dance; (V) the number of bees observing the dance from the front row; (K) the total kinetic energy of the hive, representing the overall level of synchronized activity; (P) a subset of bees initiating bursts in kinetic energy; and (O) the number of bees outside the hive.

applied this approach to ALTER3 by scaling up to a network of over 20 LLMs, incorporating modules responsible for defining ALTER3's personality, as well as meta-level LLMs that rewrite the prompts of other LLMs. By leveraging real-time APIs, ALTER3 can now generate context-sensitive conversations on the fly (submitted to ALIFE2025). We regard this system as a practical instantiation of Marvin Minsky's "Society of Mind" theory [Min86], which posits that the mind is not a singular entity but an emergent phenomenon arising from the interaction of many semi-autonomous agents.

We computed the frequency distributions of K, D, F, and P, and then calculated the mutual information between these distributions and the hive state, defined as the level of foraging activity, measured by O. From this, we further computed the redundancy and synergy in the information structure of the hive. In terms of mutual information, the level of activity K showed a strong correlation with the hive activity O. Synergy with respect to hive activity was indeed observed among P, K, and D. Figure. 8 illustrates how the amount of this synergy varies over time across individual bursts.

These results advance the central proposition of the Community First Theory (CFT): that functional individuality arises not prior to collectivity, but through it. In both the Alter3 android and the honeybee colony, meaningful roles—whether sensory modules or foragers—emerge only within structured ensembles engaged with external environments. Agency, in this view, is scaffolded by and contingent upon community. In the android experiment, visual input, interoception, and introspective memory collaboratively contribute to self-recognition, resulting in a synergistic effect. In the honeybee experiment, 1,000 bees are individually identified and tracked inside an artificial hive over one week. Two days after the start, a door to the outside is opened. Gradually, role differentiation among individual bees emerges, and the overall activity of the hive increases. This too can be interpreted as a manifestation of synergy among differentiated roles. In both the LLM modules within the android and the role modules within the honeybee hive, alignment occurs individually through the formation of a collective. Alignment refers to top-down constraints imposed across heterogeneous modules. It adjusts behavior and communication to conform with the goals or values of the collective community.

This has direct implications for the theme of ILIAD2025, which focuses on the alignment of artificial and natural intelligences within complex systems. Our findings suggest that *alignment is not merely a property of isolated agents*, but an emergent consequence of how modules or individuals interact within collectives. By applying Partial Information Decomposition (PID), we gain a computational handle on this process, identifying whether different inputs contribute uniquely, redundantly, or synergistically to system behavior.

Crucially, in the Alter3 system, periods of high synergy corresponded to increased decision confidence and accuracy. This suggests that "informational synergy may serve as an operational signature of alignment": a measure of internal coherence among system components that reflects not only correct outputs, but agreement among modalities or agents.

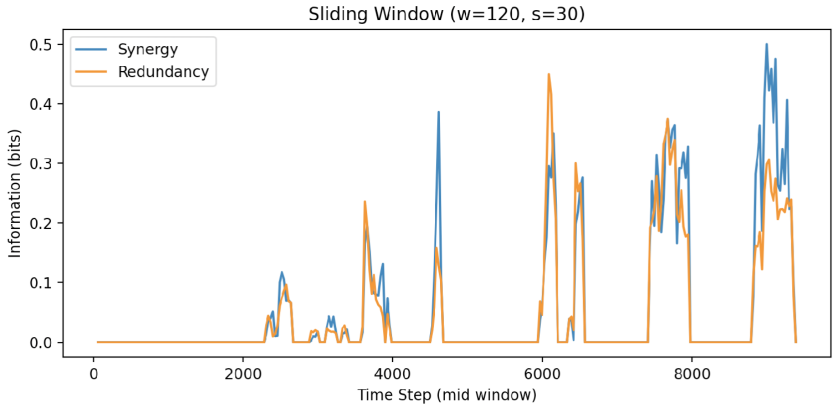

Figure 8: Time evolution of synergy and redundancy associated with the bursting time series, calculated using the Williams–Beer method. The analysis considers three sources—K (total kinetic energy), P (initiators of kinetic bursts), and D (dancing bees)—with respect to the hive state represented by O (number of bees outside). A sliding window of 120 seconds was used.

Rather than encoding alignment purely through pre-defined objectives, our findings support a *structuralist and emergent approach*: building collectives that naturally give rise to aligned, individuated intelligence. In the case of the android, it is the presence of a physical body—its boundary condition or constraint—that enables the emergence of synergy effects. Similarly, in the case of honeybees, it is the hive that serves as a boundary condition, allowing such synergy to arise. At the same time, both systems are exposed to flows of information from the outside. For the android, this occurs through visual, auditory, and interoceptive inputs. For the honeybees, it is through individuals leaving the hive and returning with new information from the external environment. These inflows of novel information contribute to the emergence of further synergy effects. It is important to emphasize that these synergy effects constitute what we refer to here as *alignment*.

Acknowledgement The analysis of the humanoid robot Alter3 is primarily based on the master's thesis of Suzune Baba (The University of Tokyo, 2025), titled "Is it Possible for Humanoid Robot 'Alter3' to Possess Self-Awareness? — Multimodality and the Formation of Selfhood."

This work has been partially supported by the MEXT Grant-in-Aid for Scientific Research "Exploring the Qualia Structure with Large Language Models and a Humanoid Robot" (Grant No. 24H01546) for the android-related research, and by "Community First Theory: A Theory and Experiments on the Evolution of Individuality, Diversity, and Spontaneity" (Grant No. 24H00707) for the honeybee-related research.

References

[AAB+24] Altera. AL, Andrew Ahn, Nic Becker, Stephanie Carroll, Nico Christie, Manuel Cortes, Arda Demirci, Melissa Du, Frankie Li, Shuying Luo, Peter Y Wang, Mathew Willows, Feitong Yang, and Guangyu Robert Yang. Project sid: Many-agent simulations toward ai civilization, 2024.

[DDI23] Itsuki Doi, Weibing Deng, and Takashi Ikegami. Spontaneous and information-induced bursting activities in honeybee hives. *Scientific Reports*, 13(1):11015, July 2023.

[GCS+24] José Luis González, Long Cheng, Sheng Shen, Pietro Liò, and Yunlong Bian. Lyfe agents: Brain-inspired low-cost generative agents for social interactions. *arXiv preprint arXiv:2401.10061*, 2024.

[GRM+18] Tim Gernat, Vikyath D. Rao, Martin Middendorf, Harry Dankowicz, Nigel Goldenfeld, and Gene E. Robinson. Automated monitoring of behavior reveals bursty interaction

patterns and rapid spreading dynamics in honeybee social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(7):1433–1438, 2018.

[Inc17] Robin A. A. Ince. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7), 2017.

[KBO+20] David Krakauer, Nils Bertschinger, Eckehard Olbrich, Jessica C. Flack, and Nihat Ay. The information theory of individuality. *Theory in Biosciences*, 139(2):209–223, June 2020.

[LRM+24] Andrea I. Luppi, Fernando E. Rosas, Pedro A. M. Mediano, David K. Menon, and Emmanuel A. Stamatakis. Information decomposition and the informational architecture of the brain. *Trends in Cognitive Sciences*, 28(4):352–368, April 2024.

[Min86] Marvin Minsky. *The Society of Mind*. Simon & Schuster, New York, NY, 1986.

[MMI21] Atsushi Masumori, Norihiro Maruyama, and Takashi Ikegami. Personogenesis through imitating human behavior in a humanoid robot "alter3". *Frontiers in Robotics and AI*, 7:532375, 2021.

[WB10] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, April 2010. arXiv:1004.2515.

[YBMI24] Takahide Yoshida, Suzune Baba, Atsushi Masumori, and Takashi Ikegami. Minimal self in humanoid robot" alter3" driven by large language model. *arXiv preprint arXiv:2406.11420*, 2024.

[YMI25] Takahide Yoshida, Atsushi Masumori, and Takashi Ikegami. From text to motion: grounding gpt-4 in a humanoid robot "alter3". *Frontiers in Robotics and AI*, 12:1581110, May 2025.

[YMM+23] Takahide Yoshida, Atsushi Masumori, Norihiro Maruyama, John Smith, and Takashi Ikegami. Development of concept representation of behavior through mimicking and imitation in a humanoid robot alter3. In *Artificial Life Conference Proceedings 35*, volume 2023, page 42. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , 2023.

Appendix

Three-Variable PID Structure

When applying Partial Information Decomposition to three input sources (Vision V, Proprioception P, Memory M, in our case of Alter3) and a target variable S, the complete redundancy lattice contains 18 distinct information atoms. These atoms represent all possible ways information can be distributed among the three sources.

Based on the redundancy lattice structure, the total mutual information is partitioned into the 18 atoms:

Synergistic Information

- $\{V, P, M\}$: Synergistic information when all three sources are considered together
- $\{V, P\}$: Vision and Proprioception together
- $\{V, M\}$: Vision and Memory together
- $\{P, M\}$: Proprioception and Memory together

Unique Information

- $\{V\}$: Vision alone
- $\{P\}$: Proprioception alone
- $\{M\}$: Memory alone

Redundancy

- $\{V\}\{P\}\{M\}$: Redundant information among all sources
- $\{V\}\{P\}$: Redundant between Vision and Proprioception
- $\{V\}\{M\}$: Redundant between Vision and Memory
- $\{P\}\{M\}$: Redundant between Proprioception and Memory

Other Information Atoms

- $\{V, P\}\{V, M\}$
- $\{V, P\}\{P, M\}$
- $\{V, M\}\{P, M\}$
- $\{V, P\}\{V, M\}\{P, M\}$
- $\{V\}\{P, M\}$
- $\{P\}\{V, M\}$
- $\{M\}\{V, P\}$

Components Used in This Study

In our three-variable PID analysis, we focused on the following key components:

- **Synergy**: $\{V, P, M\}$
- **Redundancy**: $\{V\}\{P\}\{M\}$
- **Unique Information**: $\{V\}$, $\{P\}$, $\{M\}$

All PID calculations were performed using the `dit` Python library, which implements both Williams-Beer and I_{ccs} decomposition methods.

Williams-Beer (WB) vs. I_{ccs} Approaches

The two PID methods used in this study differ primarily in how they calculate redundancy:

Williams-Beer PID [WB10]:

- Redundancy is defined as the minimum mutual information across sources
- Conservative approach that tends to underestimate redundancy
- Always produces non-negative unique information values

I_{ccs}[Inc17]:

- Uses pointwise common change in surprisal to calculate redundancy
- Allows for negative unique information values, indicating conflicting or misleading contributions
- More sensitive to nonlinear interactions and context-dependent coordination