Cognitive Distinctions as a Language for Cognitive Science: Comparing Methods of Description in a Model of Referential Communication

Thomas M. Gaul[1]   Eduardo J. Izquierdo[2]

[1]Cognitive Science Program, Indiana University Bloomington
tgaul@iu.edu
[2]Department of Electrical and Computer Engineering, Rose-Hulman Institute of Technology
izquierd@rose-hulman.edu

Abstract

An analysis of the kinds of language we use in scientific practice is critical to developing a more rigorous and sound methodology. This article argues that the methods of description commonly employed in cognitive science risk obscuring important features of an agent's cognition. We propose an alternative method to alleviate this problem, wherein the concept of *cognitive distinctions* plays a central role. A model of referential communication is developed and analysed as a platform to compare the ordinary method of description with description by cognitive distinctions. We demonstrate that the ordinary method fails to adequately capture an agent's perspective, whereas cognitive distinctions in a graph-theoretic formalism are, in contrast, capable of reflecting that perspective. We then consider how different methods of description relate to one another in a broader methodological framework. Finally, we explore the consequences of, and challenges for, cognitive distinctions.
Keywords: cognitive distinctions, method of description, conditions of observation, natural language, referential communication, interaction graph

1 Introduction

Language plays an outstanding role in how we think about scientific questions and go about answering them. This is evident not only in such questions and answers being articulated in propositions, but also in the metaphors we use; for instance, the use of informational concepts in biology (Oyama, 2000) and the computer metaphor in cognitive science (Newell and Simon, 1976; Winograd and Flores, 1986). That is, the language in which questions are posed greatly shapes our intuitions about the sorts of things we are studying and the form of our scientific practice. It should be clear, then, that a careful evaluation of such language is an essential part of evaluating our scientific paradigms, not only towards a greater refinement of our theories, but also to ensure that we do not miss blind-spots in adhering to a particular pattern of thought.

In cognitive science, the influence of language is most prominent in the various ontologies of cognitive systems: brains as computers, agents as dynamical systems (Van Gelder, 1998), extended minds (Clark and Chalmers, 1998), and so on. While these differences of theoretical framework are quite apparent in their consequences, there may be more subtle variations in language *within* a given framework whose consequences are equally significant. This article seeks to highlight how the language used to describe the behaviour of a system — in cognitive terms — influences the capacities we attribute to that system and the grounds on which we do so. Indeed, we show how the methods of description usually employed can in fact obscure important features of a system's cognition. This is not to say that the ordinary method *misrepresents* cognitive phenomena, but that its application is ill-suited to certain pragmatic contexts wherein we are concerned with the behaviour and perspective of non-human systems.

We use a simple model of communication as a platform to compare descriptions. It is well-suited to the task since, as language-users, we have strong intuitions about the nature of communication and what it looks like. As will be shown, how these intuitions are embedded in our language can greatly influence what we do and do not pay attention to or emphasize in a model. Moreover, communication is interesting in its own right, not only as a means of clarifying linguistic concepts in simpler cases and with greater evolutionary continuity, but also as a component of a general theory of adaptive behaviour (Beer, 1995, 1997; Williams et al., 2008). But work towards such a theory demands a clear account of our methods.

Prior to such an account, we also need a clearer view of the theoretical framework within which those methods are situated. The present article uses concepts from the work of Maturana and Varela on the biology of cognition (Varela, 1979; Maturana and Varela, 1980, 1987), and Varela's later work on enaction (Varela et al., 2017). From this perspective, we see cognition not as a process of computing outputs from sensory inputs, but rather as the continuous and dynamic interaction between an agent and its environment. In this framing, the actions of an agent are not functions of sensory inputs. They are, instead, compensations for external perturbations co-determined by the agent's internal dynamics; the same perturbation can elicit different behaviours depending on the internal state of the system perturbed. The set of all internal states of a system and the perturbations that induce transitions between them constitute what Maturana and Varela call that system's *cognitive domain* or *domain of interaction* (Maturana and Varela, 1980; Beer, 2014). The cognitive domain, in other words, is what structures the world of our experience and is determined by our actions within it. Within this framework, explaining cognition is a matter of showing how the dynamics of agent-environment systems give rise to coherent behaviour; it is not explained by how the agent represents a pregiven environment. Furthermore, subsets of the cognitive domain can be specified, forming specialised domains in their own right. For example, communicative interactions take place in what Maturana and Varela call a *consensual domain*, in which agents in maintained coupling serve as sources of mutual perturbation for one another, thus shaping each other's paths through their respective cognitive domain. We will not use the term 'consensual domain' in this article, nor will we yet commit to a particular definition of communication, but there will be frequent mention of the cognitive domain of both ourselves and the model agents we describe.

More formally, we can take a dynamical systems perspective on whole agent-environment systems and analyse how communicative behaviour is generated from the dynamic interaction between agent and environment (Beer, 1995). Toy models can be used to fully analyse a minimal case of a given behaviour from this perspective (Beer, 1996). Referential communication has been modelled frequently in this way, beginning with Williams et al. (2008). Communication is *referential* when it is about matters temporally and/or spatially displaced from the immediate "here and now." Though our concerns are with communication in general, reference is a useful feature of a model, as it acts as a constraint that stands in for the lives of otherwise independently acting agents. For instance, Williams et al. (2008) used an evolutionary algorithm to generate agents to solve a task in which a *sender* agent would have to move itself in the sensory range of a *receiver* agent such that, in response to the sender's particular pattern of movement, the receiver would move to some target location and stay there (and the receiver has no information about the target's location prior to interaction). When no restrictions are placed on how the agents may move, the sender either shepherds the receiver directly to the target location, or else sits at that location, waiting for the receiver to bump into it and stop. Both of these solutions, intuitively, do not capture something important about communication, that being that the receiver's behaviour should at some point be independent of the sender's after some period of mutual interaction.

Thus, to the end of comparing methods of description, this article proceeds as follows. First, we characterize what we see as the default method of description usually employed in modelling practice; this is the reference against which we compare the method we later introduce. Second, we identify its main problem as the conflation of conditions of observation and the cognitive capacities associated with them (we elaborate on these concepts in the next section). Third, we present an alternative method of description — by focusing on *cognitive distinctions* — that tries to alleviate this. Then, we present a model of referential communication and describe a particular sender-receiver pair using both methods before comparing them. Finally, we discuss further implications of the method and some challenges.

2 Methods of Description

We have thus far been rather vague about what we mean by communication, and this is deliberate, since how one chooses to do this depends on their method of description and how they apply that method to natural systems — language in humans, the waggle dance in bees (Frisch, 1967; Chittka, 2023), etc. — and artificial systems — the model presented below.

In the models on which this article builds, establishing what communication is generally follows a basic pattern. We consider certain examples of communication in natural systems and then abstract general characteristics of these behaviours in terms of spatio-temporal patterns. This generally results in a list of constraints on spatial trajectories that can be implemented in a model. For example, Williams et al. (2008) propose the following, applicable to a single sender-receiver pair and an object of reference:

1. The future behaviour of the receiver is sufficiently constrained by interaction with the sender,
2. The receiver's behaviour should vary with properties of the referent,
3. The sender-receiver interaction should have a degree of separation from the referent,

where 'behaviour' can be taken as synonymous with spatial trajectory. Most other models have followed Williams et al. (2008) in this regard, adding further complications to address specific questions such as information dynamics (Manicka, 2012), role negotiation (Campos and Froese, 2017), and behavioural flexibility (Yao et al., 2023). The exception to this is Fox and Bullock (2023) who use the teleosemantic notion of 'proper function' introduced by Millikan (1984, 1989):

> "Referential communication occurs when the signal-producing behaviour of one agent (the *signaller*) has the proper function to adapt a second agent (the *receiver*), via its sense organs, to some state of affairs, and when this second agent's signal-consuming behaviour has the proper function to be so adapted." (Fox and Bullock, 2023, italics in original)

Here, the proper function of a behaviour is the function that, when performed by the agent's ancestors, led to the genes for that behaviour being propagated. While we do not find this to be a very compelling definition of communication, the reasons for this do not concern us here. (For a critique of Millikan's and others' aetiological theories of function, see Christensen and Bickhard (2002).) However, when it comes to designing a task for model agents, the authors seem to use the same constraints as described above. Thus, there appears to be an implicit change of description in which cognitively loaded terms ('adapting to a state of affairs') are exchanged for more concrete spatio-temporal terms.

What is relevant here about these definitions is that they describe (or imply) the *conditions of observation* of those phenomena we consider to be instances of referential communication. They characterize — whether implicitly or explicitly — the spatial trajectories of the agents involved and what constraints they must satisfy in a given context in order that we call those trajectories instances of referential communication. Thus far, there is nothing inherently wrong with this method of description. We must begin somewhere in trying to model a given phenomenon, and the conditions of observation are all we have prior to an actual investigation. Hence, such conditions can serve as constraints on a task that model agents can solve.

Furthermore, it is certainly useful to describe the behaviour of particular agents in spatio-temporal terms, to the extent that this is a simple reformulation of what is expressed by a visual representation of the agent's spatial trajectory (in fact, we do this for the model presented below). However, matters become problematic when satisfaction of the conditions of observation (e.g., an agent solves the task) is taken to imply the cognitive capacity we associate with those conditions. The problem is that the conditions do not uniquely specify a cognitive capacity, as they say little about the actual operation of the agent or how the observed behaviour is situated within the agent's cognitive domain. We are suggesting here a differentiation between *behavioural* or *spatio-temporal* description and *cognitive* description.

This problem is exasperated when we use natural language verbs in describing the behaviour of a model agent (communicating, recognizing, searching, etc.). It may be contended that such descriptions are metaphorical or made tentatively in recognition of the imprecision of natural language. But what does the more precise description look like? There seems to be no clear way to move from a natural language description to a formal description beyond basic representations of physical space. This is not to say we need formal definitions of the concepts implicit in language, but rather a method of description that is amenable to formalisation and flexible with respect to the structure of an agent's cognitive domain, as opposed to one restricted to a human's cognitive domain. Here, we will be primarily targeting this verb-based method of description, as it is the only substantive cognitive description commonly used. When it is not applied directly to the agent, it serves as an implicit connection between spatio-temporal descriptions of the agent and the stated purpose for investigating it (e.g., to study *communication*).

What is needed, then, is a method of description that takes the perspective of the agent as central. Thus, we propose *cognitive distinctions* as a useful concept in this regard. We define a cognitive distinction as the *sufficient differentiation* of state trajectories following varied perturbations. This definition is clearest when we treat the system's states and behaviour as discrete (at least in approximation). For example, imagine an experiment in which a visual stimulus is presented that only varies in color. The subject must identify the color by name. Clearly, we can expect that minor variations in shade would be named together, while red and blue, for example, would be named differently. We can then get a sense of the structure of the cognitive domain by mapping these names onto a color space. Suppose we also track the internal state of the subject (brain activity, metabolism, etc.) before each presentation of a color stimulus: we can then connect these states by the stimulus that induced the transition. This can be done in terms of the complete color space (continuous) or in the labelled one (discrete). When carried out exhaustively on all internal states, this process generates the network of all possible color-naming interactions.

In this (unrealistic) example, the 'varied perturbations' are the different colors and the condition of 'sufficient differentiation' is the subject giving different names. We attribute a particular cognitive distinction to one of the subject's states when different color presentations on that state result in different names.

While we will reserve a fuller defense of cognitive distinctions for when we have a more concrete example, a few points are in order here. Firstly, a complete mapping of the cognitive domain would demand the presentation of all possible perturbations on all possible internal states. Obviously, this is an impossible task for any real system. But we need only a subclass of perturbations with some degree of structure to achieve something useful, as demonstrated in the color example. Moreover, as we intend to apply this method to simple model agents, there are far fewer degrees of freedom to worry about. Secondly, we attribute cognitive distinctions to *states* and not the whole agent; it is a local attribution, not a global one. Thirdly, the imaginary experiment described above may appear to contradict our dynamical perspective in mapping 'inputs' to 'outputs', but this is more an artefact of discretisation and pedagogical decisions. The temporal separation of apparent inputs and outputs breaks down in more complicated continuous interactions. Further, even in the discrete case, when we do not have complete control of the full space of perturbations, the actual perturbations any system experiences are in part determined by the system itself (we move our head to get a different view).

3 The Model

This model is largely based on those of Williams et al. (2008) and Yao et al. (2023). There are two agents, called *sender* and *receiver*. They exist in a one-dimensional periodic environment of circumference 2π, along which they can move in either direction. Each agent is equipped with a single continuous sensor sensitive to any object in the environment, with a range of $\pm\frac{\pi}{64}$ centered about the agent; the sensor is unsigned, so they cannot directly determine the direction of a stimulus.

All possible objects in the environment — of which there are two kinds — are points on the circle. The other objects besides agents are 'posts'. Posts are organized into sets such that, within a set, each post is placed in sequence $\frac{\pi}{32}$ apart. We will notate post-sets by $P = n$, for n posts in the set.

The Agents

Each agent is controlled by a five-neuron continuous-time recurrent neural network (CTRNN) governed by the following state equation (Williams et al., 2008):

$$\tau_i \dot{y}_i = -y_i + \sum_{j=1}^{N} w_{ji}\sigma(y_j + \theta_j) + g_i s \tag{1}$$

where y_i is the state of each of N neurons, τ_i is a time-constant, w_{ji} is a connection weight from the j^{th} neuron to the i^{th}, θ_i is a bias term, $\sigma(x) = 1/(1 + e^{-x})$ is the standard logistic activation function, and g_i is a connection weight from the sensor s to the i^{th} neuron. The output of a neuron is $o_i = \sigma(y_i + \theta_i)$. The network is fully interconnected, including self-connections, and the sensor has a single weighted connection to every neuron (Figure 1). Both sender and receiver share the same parameters. The sensor activation is defined by the following equation:

$$s(\tilde{d}) = \frac{1}{1 + e^{5(2\tilde{d}-1)}} \tag{2}$$

where $\tilde{d}$ is the absolute distance from the agent to the nearest object (another agent or otherwise), normalised to the sensor range. If the distance is greater than that range, s is set to 0.

Two neurons are designated to drive the + and − motors. We will call the neuron that drives the + motor the + motorneuron, and analogously for the − motor. In the receiver, the circuit is mirrored such that the neurons driving the motors are swapped compared to the sender (i.e., N_1 drives the + motor

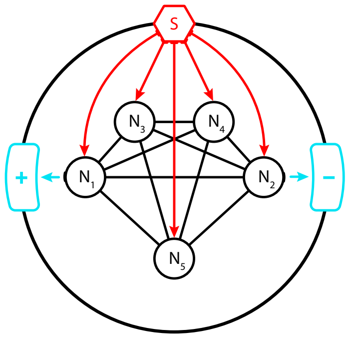


Figure 1: The agent configuration of a sender. All agents have a sensor (red) connected to all N neurons (black). The neurons are fully interconnected, including self-connections (not depicted). In the sender, the N_1 drives the $+$ motor, and the N_2 drives the $-$ motor. In the receiver (not depicted), this configuration is reversed.

in the sender, but the $-$ motor in the receiver). This constitutes the only difference in configuration between the agents; they therefore move in opposite directions for the same pattern of motorneuron activation. The velocity per time-step of an agent is given by: $v = \gamma(o_1 - o_2)$, where o_1 and o_2 represent the outputs of the motor neurons, and γ is a constant corresponding to the maximum velocity. For this study, the maximum velocity was set to $\gamma = 3$. Simulations were run with Euler step integration and a step size of 0.01.

The Task

The task is organised in three phases: 'Phase 1', 'Phase 2', and 'Phase 3' (Figure 2). The phases last for 250, 300, and 600 time-steps, respectively.

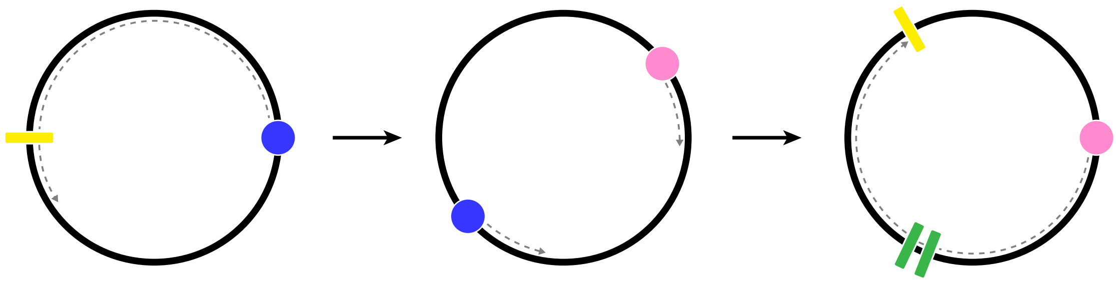


Figure 2: A task configuration. The task environment is one-dimensional and periodic (a circle). In Phase 1 (left) the sender (blue) interacts with the $P = 1$ target post-set (yellow). In Phase 2 (center) the receiver (pink) interacts with the sender. In Phase 3 (right) the receiver moves until it reaches the target post-set, avoiding the $P = 2$ distraction (green).

The agents are always initialised with the state of all their neurons set to 0; they are only initialised in the phase in which they first appear in a given permutation (a single instance of the task), meaning they maintain their state trajectory between phases. In Phase 1, the sender begins at position $0 \pm \frac{\pi}{32}$ and the target at π (agent positions are always initialised on a uniform Gaussian distribution, while the positions of post-sets remain constant). After interacting with (passing through) the target, the sender must then continue into the next phase, primed for different patterns of behaviour depending on whether the target was $P = 1$ or $P = 2$. Phase 2 begins with the target removed, the sender continuing its trajectory, and the receiver initialised $\pi \pm \frac{\pi}{32}$ units from wherever the sender is; the environment is now empty apart from the agents. The interaction between sender and receiver must be such that the receiver's state and/or position varies with respect to the target by the end of the phase. The beginning of Phase 3 removes the sender and sets the receiver's position to $0 \pm \frac{\pi}{32}$. There are two post-sets equidistant from this location, at $\pm\frac{2\pi}{3}$; one is designated the target and the other a distraction. The sender and receiver are evaluated based on how close the receiver is to the target post in Phase 3.

This is repeated over four permutations; we notate permutation n by $\texttt{Pu} = n$. $\texttt{Pu} \in \{1, 2\}$ have $P = 1$

as the target (the "target condition" is $P = 1$); they are thus identical for the first two phases. They differ in Phase 3 with respect to the arrangement of the post-sets: $\texttt{Pu} = 1$ has the target at $-\frac{2\pi}{3}$ and the distraction at $\frac{2\pi}{3}$, while $\texttt{Pu} = 2$ has the opposite configuration. $\texttt{Pu} \in \{3, 4\}$ are similarly related, except with $P = 2$ as the target. The permutations are always in the same order.

The agents are evaluated as a pair over five trials, where a trial is one cycle of the four permutations. Fitness is calculated according to the following equation:

$$f = \min \left\{ \frac{1}{|\{\texttt{Pu}\}|} \sum_{\texttt{Pu}} \left(1 - \frac{\overline{d} - d_c}{\frac{2\pi}{3} - d_c} \right), \, trial \in T \right\} \tag{3}$$

where $\{\texttt{Pu}\}$ is the set of permutations, $\overline{d}$ is the average distance of the receiver from the target during the last 250 time-steps of Phase 3, d_c is the minimum distance from the target necessary for perfect performance (the "close enough" range), and T is the total number of trials (here, 5). The fitness is thus calculated as the worst trial.

A given permutation is terminated and evaluated to 0 if any of the following conditions are met: (1) the sender is within d_c of the target at the end of Phase 1; (2) the agents are within d_c of each other at the end of Phase 2; or (3) the receiver interacts with the distraction *after* contact with the target. The first two constraints ensure continuous activation values (no sudden jumps between phases) while the third is meant to ensure that the receiver can only depend on previous interaction with the sender to successfully reach the target.

Evolution

Parameters for the CTRNN were evolved using a real-valued genetic algorithm (Beer, 1996). The following parameter ranges were used: time-constants $\tau \in [1, 30]$, connection weights between neurons $w \in [-16, 16]$, biases $\theta \in [-16, 16]$, and connection weights from the sensor to each neuron $g \in [-16, 16]$. A generational algorithm with rank-based selection was used on populations of 539[1] genotypes, each evolved for $10,000$ generations. Genotypes are 40-dimensional vectors of real numbers in the range $[-1, 1]$, where each number encodes the value of a single parameter by mapping it to the ranges specific above. The evolution begins with a population of random genotypes evaluated on the task. Successive generations are created by first selecting the top 5% of genotypes based on fitness (the 'elitist fraction'). These genotypes are copied directly into the next generation without modification. The remaining genotypes are copied and mutated by adding a random displacement vector whose direction is sampled from a uniform distribution of unit vectors and whose magnitude is sampled from a Gaussian distribution with mean 0 and variance 0.2. The whole population (including elites) is then reevaluated before creating the next generation. Evolutions were run until a genotype achieved a fitness of 0.95. 176 populations were evolved, producing 16 successful agent pairs.

4 Comparing Descriptions

This section analyses a particular agent pair in order to compare natural language description and description by cognitive distinctions. First, we provide an overview of the agents' spatial trajectories and begin analysing the neural mechanisms that support those trajectories. We then use Dynamical Systems Theory (DST) to provide a more comprehensive analysis, before applying both methods of description to the system; in particular, we present a provisional formalism for cognitive distinctions to articulate the description. Finally, we evaluate how the descriptions relate to each other and the agents' operation.

4.1 Behaviour and Neural Mechanisms of a Simple Agent

Figure 3 shows the behavioural trajectories of a sender-receiver pair. When the target condition is $P = 1$, the sender passes through the target seemingly unaffected during Phase 1. In Phase 2, sender and receiver move in trajectories of near constant velocity, crossing each other three times. In Phase 3 for $\texttt{Pu} = 1$, the receiver crosses the $P = 2$ distraction three times before passing through the target and slowing after contact to re-approach it from the opposite direction. For $\texttt{Pu} = 2$, the receiver simply

[1]The strange population size is due to concerns over compatibility with multithreading given the elitist fraction ($0.95 \times 539 \approx 512$).

passes through the target and changes its velocity to be slower and in the opposite direction. When the target condition is $P = 2$, the sender similarly passes through the target in Phase 1, except this time it begins slowing down before reaching the end of the phase. Then, in Phase 2, the sender has a very low velocity while the receiver passes by it at its initial speed. In Phase 3 for $\texttt{Pu} = 3$, the receiver passes through the $P = 1$ distraction before slowing down and changing direction just before reaching the target. For $\texttt{Pu} = 4$, the receiver very quickly turns and slows after passing through the target.

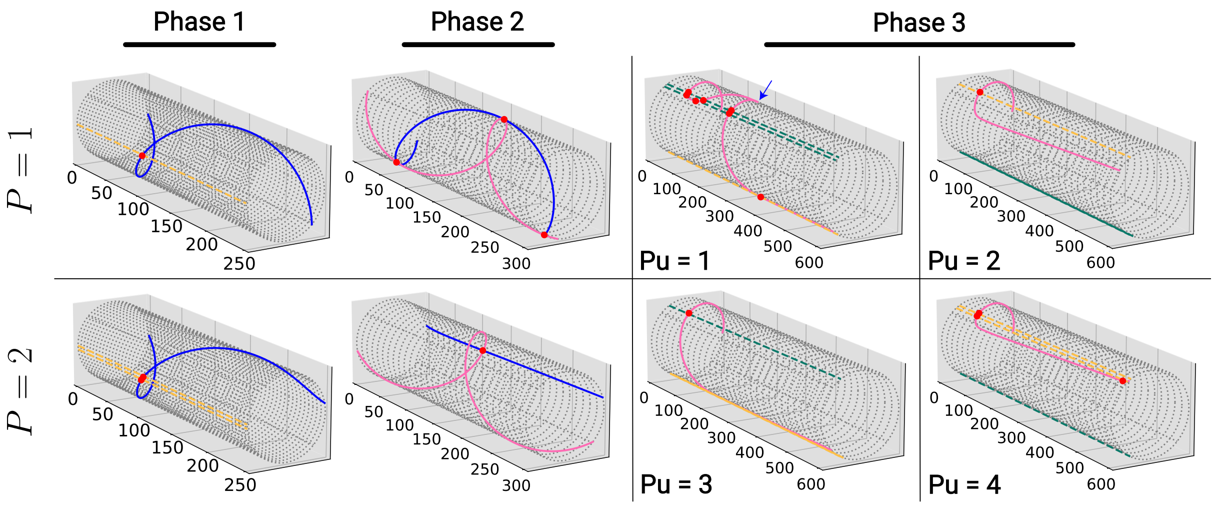


Figure 3: Position trajectories of 4 permutations. The numbered axis is time. The sender is blue, the receiver is pink, targets are yellow, distractions are green, and contact points are red. The plots in the top row are for permutations 1 and 2, target condition $P = 1$. The plots in the bottom row are for permutations 3 and 4, target condition $P = 2$. The blue arrow in $\texttt{Pu} = 1$ indicates the position at which the receiver returns to its *go*-attractor. Positions are initialized with no noise.

What should we make of this behaviour, and how should we describe it in a cognitive manner? At this point, we do not yet want to construct our description by cognitive distinctions prior to an explanation of the agents' operation. Part of this is rhetorical, in that our proposal is more convincing after we have permitted a full elaboration of the ordinary method of description. For instance, we might say that the sender "identifies" the target, "communicates" the target to the receiver, and that the receiver "finds" the target in its environment — clearly, these verbs (scare-quotes or not) do more to fuel our intuitions than inform us of anything useful. To use this caricature would be a dishonest representation of what actual descriptions in the literature look like. However, this is not to say that the language fundamentally changes (if it is present at all), but that mechanisms are sought that roughly correspond to the verbs used. This constitutes the justification of the language.

Another part is that we do not yet have sufficient information to construct a description by cognitive distinctions. Namely, we need to first have an account of the transitions between the agents' internal states that are induced by perturbations before we can reconstruct a network of such transitions. And again, the appropriateness of the description is best evaluated in light of a more thorough understanding of how the agents actually work.

We divide our analysis into subtasks corresponding to each phase and compare permutations within the phase. As we have already described and shown in Figure 3, there are two behavioural trajectories observed in Phase 2. The neural traces of the motorneurons during these trajectories are shown in the first column of Figure 4. We can see in these plots a sort of stepping mechanism in which repeated contact causes N_1 (cyan) to move in discrete-like steps between relatively stable or slow-moving states. As these steps accumulate, the neuron eventually reaches a point where it swings, at different rates, into a more active state where its output is near 1.0 (second and third columns). The dotted line in Figure 4 indicates this threshold point. Thus, the receiver enters Phase 3 with different levels of sensitivity to perturbation-induced transitions, corresponding to discrete-like steps; we label these steps *low*, *mid*, and *high*, where *high* is nearest the threshold.

The rate of these transitions is also important in light of the different slowing speeds we see in Phase 3. We can sort the patterns of this phase into three categories: *reset* in $\texttt{Pu} = 1$ such that neuron N_1 returns to *low* (this point is indicated by the blue arrow in Figure 3); *fast-stop* in $\texttt{Pu} \in \{2, 4\}$; and *slow-stop* in $\texttt{Pu} \in \{1, 3\}$. Notice also how not only the step-mechanism, but also the particular post-set

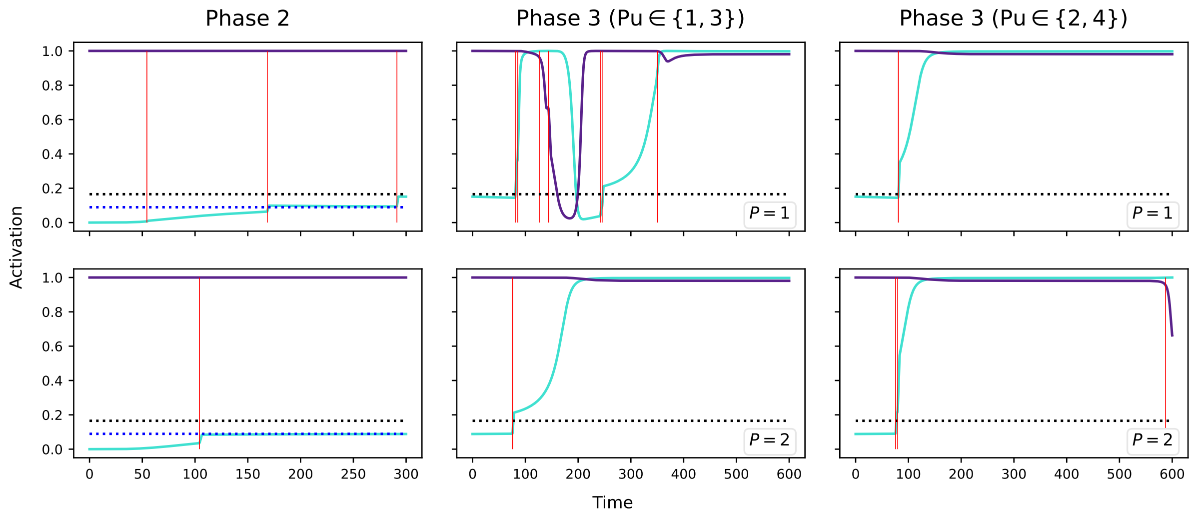


Figure 4: Neural traces of receiver in phases 2 and 3. The traces in the top row have target condition $P = 1$, and the ones on the bottom have $P = 2$. Neuron N_1 is cyan and neuron N_2 is purple. Vertical red lines represent contact points (when the sensor value is > 0.5). The dotted grey line is the threshold (saddle) point; the dotted blue line is the attractor (*mid*).

determine whether the N_1 passes the threshold. That is, if we interpret the difference between $P = 1$ and $P = 2$ as a difference of magnitude (in the sense that the sensor is active for longer passing through $P = 2$ than $P = 1$), magnitude correlates with larger steps in N_1.

We now have two matters left to explain: (i) why does a $P = 2$ perturbation on *high* induce the *reset* behaviour, and (ii) how do all these patterns and mechanisms relate to one another? To answer these questions, we need a more global picture of the agent's operation achieved by a dynamical analysis. (For an introduction to DST, see Garfinkel et al. (2017) or Strogatz (2018); for its application to brain-body-environment systems, see Beer (1995).) The following analysis

4.2 Dynamical Analysis

We selected this agent because it can be reduced to a two-dimensional dynamical system (Figure 5); the procedure is as follows. First, we found that two of the interneurons could be lesioned without a noticeable effect on performance. Then we fixed the remaining interneuron to maintain an output value of 1. Thus, the whole system can reasonably be analysed in the space of the two motorneurons. The only qualitative change in the reduction is a saddle point becoming an unstable fixed-point for certain values of s (red line in Figure 6), but this can be explained by noticing that there are two positive eigenvalues at these points in the full state-space (and thus two unstable directions). When the system is reduced, these are the only eigenvalues remaining, the other stable directions disappearing.

We now note some important properties of the dynamics. The phase-space under no sensory stimulation contains two attractors, separated by a stable saddle manifold (Figure 5). These attractors correspond to two basic behaviours: the attractor on the left corresponds to fast movement, and the other to slow movement in the opposite direction. We will refer to these as the *go*- and *stop*-attractors, respectively. To understand what happens as the system undergoes sensory perturbation, we can look at the bifurcation diagram (Figure 6). When a small sensory perturbation is applied to the system, the saddle and *go*-attractor are very quickly annihilated by each other; this allows the system to move into the basin of the *stop*-attractor (the *stop*-basin). The state of the system can return to the *go*-basin, as well, if a perturbation allows the limit-cycle to persist long enough, i.e., if the perturbation is *slow* enough around the appropriate values of s. In fact, if we notice the geometry of the *go*-basin in Figure 5, we see that the state of the system need only reach the lower part of the limit-cycle in order to be in that basin when $s = 0$. The remaining feature critical to the agent's operation is the timescale between the *go*-attractor and the saddle point (Figure 7a). Here, because the timescale is so slow in this region relative to the task, the system may act as if at a fixed-point despite being in a transient.

We can now recast our previous analysis in dynamical terms. First, we plot the steps of the receiver in state-space (cyan points in Figure 7b). We see that these points lay along either the initial transient

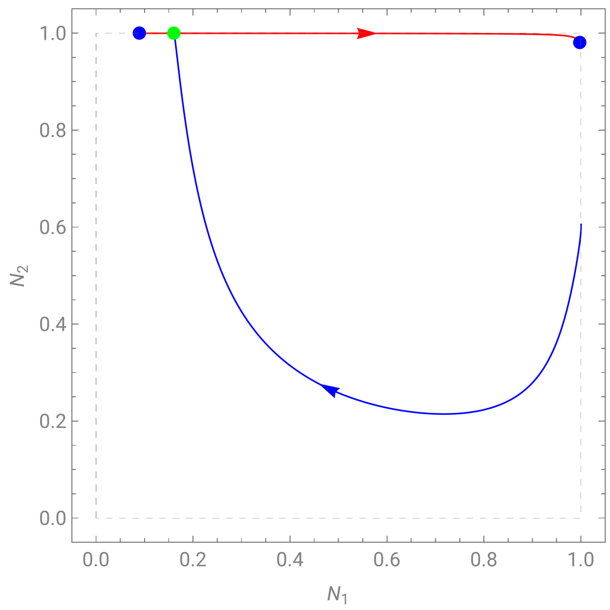


Figure 5: Phase portrait of a two-dimensional CTRNN in the output space of the motorneurons, with $s = 0$. There are two attractors (blue points) and a saddle (green point). The basins of attraction are delineated by the stable saddle manifold (blue line) and the attractors are connected by the saddle's unstable manifold (red line). The dotted grey lines indicate the bounds of the output space, $[0, 1]^2$. The agents start in the basin of the left attractor.

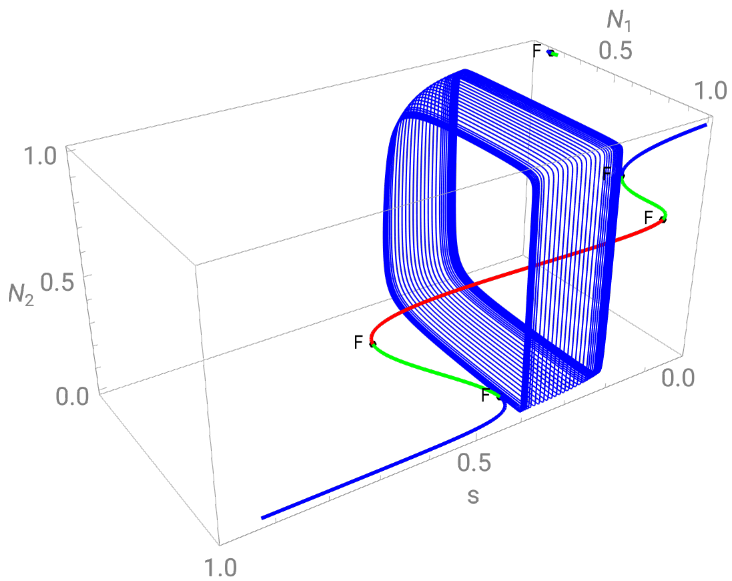
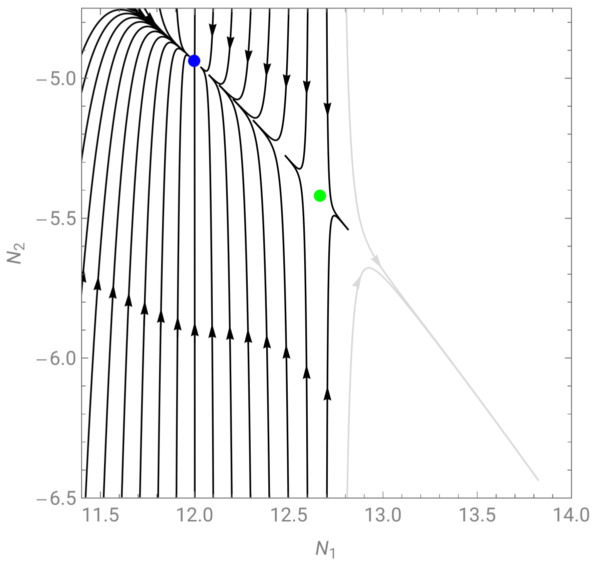
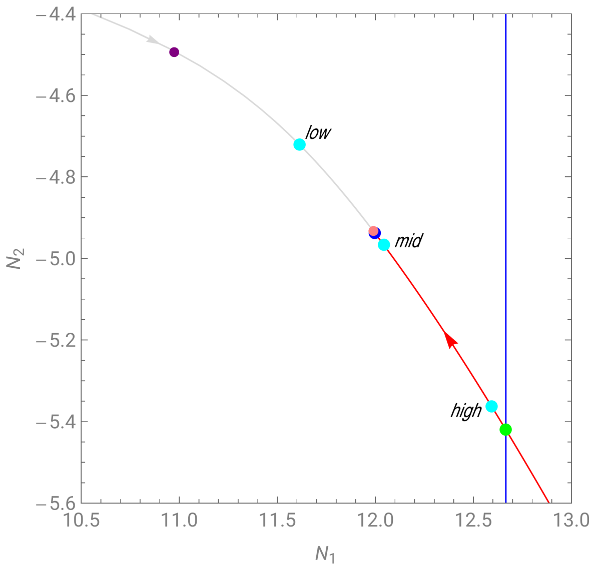


Figure 6: Bifurcation diagram of a two-dimensional CTRNN in output space. s is treated as a parameter of the system, where each value has a corresponding phase portrait (in the (N, N_2) plane). Blue lines that extend across s represent stable equilibria, red line unstable equilibria, and green lines saddle points. Closed blue curves are limit-cycles at particular values of s, but represent a continuum of limit-cycles in that range. Black points labelled 'F' are fold, or saddle-node, bifurcation points. The limit-cycle is created and destroyed by infinite-period bifurcations.



(a)



(b)

Figure 7: Phase portraits of the agent in state-space with $s = 0$. (a) Slow trajectories (black) approach the saddle's (green) unstable manifold (not depicted) before moving towards one of the attractors (blue). Fast trajectories (grey) are only separated by 0.1 units in state-space from the nearest slow trajectories. All were integrated for 145 steps. (b) States just before contact. All points lie on either the unstable manifold (red) or the initial transient from the starting state (grey). The saddle's stable manifold is in blue. Cyan points correspond to the receiver's state just before contact with the sender (*high* being nearest the saddle). The purple point corresponds to the sender in Phase 1 just before interaction with a post-set. The pink point corresponds to the receiver in the $P = 2$ target condition just before interaction with a post-set.

(grey) or the unstable saddle manifold (red). Hence, the apparent stability of the steps is due to those points being in a region with a very slow timescale. Moreover, the sequence of steps approaches the basin boundary, and it is in fact this boundary that constitutes the threshold point in Figure 4. We also note that the speed of the transition to the *stop*-attractor is dependent on how far a trajectory is pushed into the *stop*-basin, since trajectories further into the basin move faster (Figure 7a).

Finally, Figure 6 allows us to see how the *reset* behaviour works: when the state is sufficiently close to the basin boundary (*high*) and the perturbation is sufficiently large ($P = 2$), the system spends enough time in the monostable and limit-cycle regimes ($s \gtrsim 0.198$) to be pulled into the lower part of the state-space before the phase-portrait returns to its $s = 0$ form. This explains why the receiver crosses $P = 2$ three times in $\texttt{Pu} = 1$. After the receiver passes through once, it is well into the *stop*-basin, where the agent reverses direction. This causes it to run into the posts again, thus accelerating its motion with the $-$ motorneuron near 1.0 and the $+$ motorneuron near 0.0 for $s \gtrsim 0.394$ (Figure 6); this is sufficient to bring the agent into the lower part of its state-space by the time it passes through the post-set, where it then moves through the *go*-basin until it approaches the attractor and changes direction again (blue arrow in Figure 3). Then, after it passes through $P = 2$ once again, the state is brought just past the stable saddle manifold where its initial speed is slow, thus generating the *slow-stop* behaviour.

4.3 Cognitive Descriptions

We now come to finally apply the methods of description that motivated this investigation. An ordinary (verb-based) description might look like this: "the sender communicated the target ($P = 1$ or $P = 2$) to the receiver, and the receiver identified the first post-set it ran into as target or distraction and then adjusted its behaviour sensitive to information stored during communication." Compare this with a sentence from Campos and Froese (2017), "We know that the agents have to *decide* what role they should take before starting *communication*." (pp. 6; emphasis added). Similarly, Yao et al. (2023) writes, "The receiver needs to *recognize* environmental labels and develop ways of *reacting* to them that are also sensitive to the information stored in *communication*." (pp. 461; emphasis added). These descriptions make reference to the cognitive capacities of the agents as global properties using natural

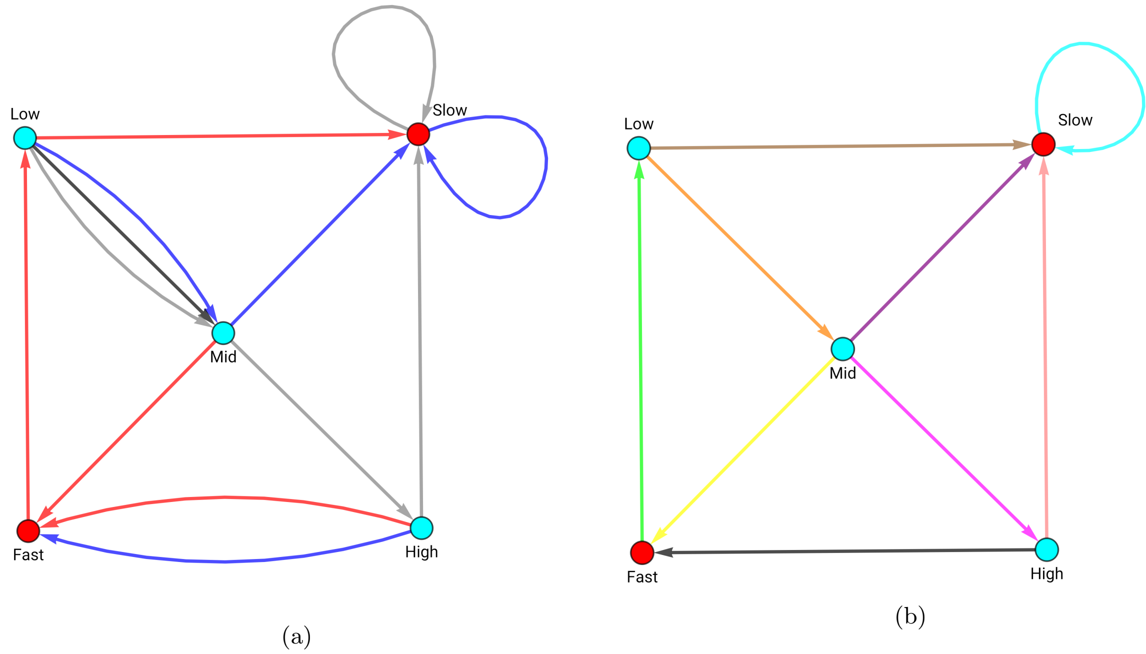


Figure 8: Partial interaction graphs of a single agent. Cyan vertices represent the stable 'steps' in Figure 7b. Red vertices represent the *slow-stop* and *fast-stop* state-trajectories of the agent. (a) Partial interaction graph without explicit perturbation classes. Edges represent state-transitions after perturbation, where blue is interaction with $P = 1$, red with $P = 2$, black with 'slow agent', and grey with 'fast agent'. Redundant edges indicate a failure to distinguish those perturbations. (b) Partial interaction graph with perturbation classes. Since no states have the same response-profile, every class is given a unique color.

language verbs. Thus, we hope, our verb-based description is a plausible representation of what occurs in the literature.

To substantiate our description, we point out the verb-mechanism correspondences. The sender's communication was simply a matter of changing velocity by transitioning between the *go*- and *stop*-basins. The receiver's information storage corresponds to the step-mechanism, leaving it in different states by the end of Phase 2. Identification of post-sets corresponds to the control of different stopping (or reset) behaviours, based on the magnitude of the perturbation relative to its state.

To construct a description by cognitive distinctions, we first split the behavioural trajectories of each agent at every perturbation. This results in five behavioural patterns: *low*, *mid*, *high*, *slow-stop*, and *fast-stop*. We differentiate the steps as distinct patterns, since we know that the same set of perturbations can elicit a different response-profile in each. Similarly, we will classify perturbations as 'fast agent', 'slow agent', $P = 1$, and $P = 2$. With these labels, we can construct a graph where behavioural patterns are vertices and perturbations are edges (Figure 8a). We call this a (partial) *interaction graph*. This partition of the behaviour and perturbations constitutes the basis for the concept of cognitive distinctions to apply; it implies a specification of *sufficient differentiation* as that which varies with respect to task outcomes. Thus, we are defining distinctions relative to the behaviour observed in the task. Note also that, since the sender and receiver have identical parameters, we can combine information from both agents to generate the graph.

More generally, we construct a graph from a set of time series by first creating a finest partition of the perturbation space such that the final graph exhibits a reversible dynamics (one edge per color per node). Then, separate states and state-patterns by the perturbations observed (assuming we can treat perturbations in such a discrete manner; we return to this point in the Discussion). This permits the construction of paths representing the behaviour of each agent in each time series. The final graph is formed by combining these paths.

There are a few things to notice in the interaction graph. Not every vertex has 4 outgoing edges, since this graph was only constructed from the behaviour observed in Figure 3 (hence we call it a *partial* interaction graph). Also notice that any particular behavioural trajectory is simply a path through the graph.

More importantly, however, is the rich picture of the agents' perspective we get from this description. For instance, notice that when *low* is perturbed by 'fast agent' (grey), 'slow agent' (black), or $P = 1$ (blue), it always transitions to *mid*; the agent *cannot distinguish* between the perturbations in this state. Moreover, what perturbations the agent can differentiate changes based on its state. This suggests that the perspective of the agent — the world that it experiences — is far more dynamic than a natural language description would imply. In particular, consider the *high* vertex; here, the agent fails to distinguish the post-sets, despite this state being critical to the receiver's success in Phase 3 — where "identifying the post-set as target or distraction" is supposed to happen.

Furthermore, we can make the graph more directly represent the experience of the agent by further constructing *perturbation classes* as further partitions of the set of perturbations already defined (Figure 8b). Here, each vertex has its own partition, where each class is an indistinguishable set of perturbations. The figure gives each perturbation class a unique color. Two vertices can use the same set of colors if their partitions are identical (this does not occur in our graph). Unfortunately, while this representation better reflects the experience of the agent, it is more opaque to the observer-community.[2] Part of this is due to the graph being only partial, but it is also a consequence of our being concerned with the task, and not the full scope of interactions the agent can engage in (we also return to this point in the Discussion). Thus, it is useful to consider both forms, one reflecting the relation between the agent's cognitive domain and our description of its environment, the other reflecting the perspective of the agent independently of that description, i.e., its *Umwelt* (Von Uexküll, 1992).

A few points bear mentioning here. The first of these is that describing agents in this way requires an account of both its behaviour *and* its environment, as well as the dynamics between them. And more than this, the environment must be explicitly considered from multiple perspectives, reinforcing the observer-community's role. The world of the agent is not pregiven, but must be constructed if we are to fully understand that agent in cognitive terms. Secondly, the graph, on its own, cannot predict the behaviour of an agent in physical space. This requires the association of particular states, patterns, and perturbations to the graph's various components. Put another way, *the same graph can simultaneously describe many different sets of behaviours* (or, at least, different to *us*). Again, it is the observer-community that relates the perspective of the agent to a description of its behaviour. Thirdly, the method we propose readily permits exploring the emergence of distinctions fundamental to certain kind of behaviour. That is, the structure of the cognitive domain may provide insight into how new behavioural repertoires emerge, including communication and perhaps even language. For instance, we might imagine that an agent-object distinction would result in a subgraph of communicative interactions, such that interaction with a conspecific would constrain an agent's behaviour to this subgraph. However, this is quite speculative and only meant as a demonstration of the method's potential application.

Let us broaden our scope once again to evaluate and compare the methods of description we have been concerned with. We first consider how they relate to explanations of the behaviour described. When we use a natural language description, the operation of the agent is very unconstrained. This in itself is not a problem, as we can interpret such a description as specifying the conditions of observation such that the ambiguity with respect to operation is, in fact, a virtue. That is, we want to assume as little as possible about the operation when we specify conditions of observation. But problems arise when natural language is interpreted beyond this role as a cognitive description of the agent. As we have seen, the expectations implicit in natural language can obscure features of an agent that may challenge our intuitions about what it is we attribute to that agent; we might attribute 'recognition' or 'identification' without considering whether the agent has the capacity to differentiate what is supposedly recognized or identified. This is clear in the example above, where the receiver cannot distinguish between $P = 1$ and $P = 2$. Another example is both agents failing to distinguish $P = 1$ from each other (see the *low* vertex in Figure 8a). It then becomes unclear whether it is even appropriate to talk about 'communication' or 'recognition' when such distinctions are not present (we elaborate on this point in the Discussion).

The correspondence between cognitive distinctions and operation is, in contrast to the ordinary method, much closer. It is not that distinctions are themselves descriptions of the operation of an agent, but that they better specify what conditions such operation much satisfy. For instance, we might expect a separation of state trajectories following different perturbations, but without specifying what sort of dynamics facilitate such separation. Moreover, since cognitive distinctions are explicitly state-dependent, we do not assume global uniformity of an agent's operation. This is what allows distinctions to better

[2]I use 'observer-community' instead of 'observer' to emphasise the inherently social and linguistic context in which we make descriptions and explain phenomena. Varela (1979) puts it succinctly: "...the knower is not the biological individual." (pp. 276).

capture the global structure.

To return to a point made in the Introduction, we emphasize that the method of description we propose is not "better" in some absolute or positivist sense, but that it is more appropriate to certain tasks than ordinary language, and *vice versa*. We can situate these methods in a broader framework that makes clear the pragmatic role they play. We have, throughout this article, used dynamical descriptions, neural descriptions, spatio-temporal/behavioural descriptions, and cognitive descriptions. They all mutually constrain one another, and each facilitate different needs. We used dynamical descriptions to explain the particular neural patterns we observed, and how those patterns relate in a global structure; we used spatio-temporal descriptions to specify the conditions of observation for referential communication, as well as to set a target for explanation of a particular agent; and we used cognitive descriptions to explore the perspective of the agent, to consider the attribution of cognitive capacities, and as a preliminary to specifying the appropriate conditions of observations. We can further consider the language in which we make these comparisons as cognitive descriptions of our own activity as linguistic organisms with scientific concerns. When we confuse these methods and aims, we risk epistemological error by forgetting what we — the observer-community — are doing.

5 Discussion

This article sought to demonstrate problems in using natural language to generate cognitive descriptions of model agents. We outlined what we consider to be the ordinary methods of description and suggested that they fail as cognitive descriptions by conflating conditions of observation with cognition as it is realized in an actual agent. We proposed the concept of 'cognitive distinctions' as the basic object for a method of description that remedies this issue. We developed and analysed a model of referential communication and described it using the ordinary method and cognitive distinctions. Finally, we extrapolated the immediate consequences of using either method and compared them to establish that the ordinary method fails to capture the perspective of an agent adequately, where cognitive distinctions can.

To clarify the point of all this: we want to establish a more rigorous and epistemologically sound methodology for doing cognitive science, and for analysing simple models of cognitive behaviour in particular. This requires an evaluation of the language we use to describe cognitive systems and, more generally, a linguistic and epistemic self-awareness that engenders a mode of inquiry in which acknowledgement of the observer-community is a necessary precondition to sensibly talk about the cognition of any system. Furthermore, the development of formalisms that require such awareness for their interpretation is one part of establishing the methodology we seek. That is, not only do formalisms provide a rigorous framework in which to articulate theoretical concepts, but they also serve as guides for how we ought to approach a given system.

The Genesis of the Problem

That we have identified a problem in how we normally describe cognitive systems leads us to ask how the problem arises in the first place. We will not provide a causal explanation, nor an adequate historical account, but a series of more or less implicit factors that may play a role in how we think of systems and their descriptions. First, there is the main confusion that we pointed out in the Introduction: the confusion of cognitive description and specification of the conditions of observation. We have already dealt with this point in the previous section, so we will not address it further.

Second, there is the assumption of a static, human environment. We often think of other organisms as living in the same world we do, perhaps with more or less detail. This leads us to project our perspective onto other systems when they appear to successfully navigate a situation. But, as we have seen, the world in which the system lives, from its own perspective, can radically undermine these assumptions. And even more interestingly, that perspective can vary greatly dependent on the agent's state; we do not generally expect that changes of state in ourselves will radically alter our perception in such a manner. What is needed, then, is a more explicit appreciation of how an agent *brings forth* its world (Varela et al., 2017).

Third, and finally, there is too often a failure to appreciate the pragmatic roles descriptions can play. For example, the assumption of appropriate verb-mechanism correspondences reflects a tendency to view scientific propositions as somehow representing what the world objectively is (from a realist's perspective). But our demonstration of the shortcomings of this method does not reflect a failure to capture some ground truth, but a misapplication of a method to a context where it is less useful. The inadequacy of the natural language description, in this case, is not a matter of being *false*, but of *sense*. When we acknowledge the perspective of the agent we are describing, there is no matter of fact about whether it 'communicates' or 'recognises', but only a matter of whether these terms serve our purposes in elucidating the operation and perspective of the agent. To suppose there is such a correspondence is to assume we have well-defined notions of our cognitive terms, irrespective of the cognitive domain on which we impose them. But these terms originate in natural language, and it is the task of the cognitive scientist to determine when they are useful or not — it is not our job to assume their validity and search for the justification later. (See Wittgenstein (1953) for a fuller explication and defense of this view of natural language and description.[3])

Again, we wish to emphasise that we are not suggesting an abolition of natural language, but an appreciation of its role and what its limitations are; it is adequate to direct our attention to phenomena of interest, to raise preliminary questions, and for participation in a scientific community, but it is not a mirror of our world, let alone of other ones. Hence, though we have called the model presented above one of "referential communication" a number of times, we do not wish to say anything about whether the agents are actually communicative. We called it such because it is based on the conditions of observation that have to this point been associated with referential communication — we wanted to make clear how this model relates to the literature.

Method and Limitations

A number of issues regarding formalisation have arisen thus far. In particular, we have described a continuous system using discrete methods. While a continuous description would be, in principle, more valid, it is not very clear what the appropriate counterpart to an interaction graph (an interaction manifold) should look like, or whether such a construction would be useful. That the discretisation appears to work in our case seems to be a consequence of two factors. The first is that the perturbations observed in the task are clearly separable and the agents, in most cases, achieved some level of stability between interactions. The second is that we only cared about the cognition of the agent with respect to the task structure. It is therefore a very coarse-grained description with very limited applicability beyond the analysis presented here. The reason we use graphs is that we have tried to extend them from their origin in discrete systems (Beer, 2004, 2014) — in which they have proved very useful — into a continuous realm, while maintaining their utility.

Regardless, we do not see any reason why the particular circumstances of this model should be considered inherent limitations. In fact, there may already be tools available that could construct more detailed interaction graphs with fewer assumptions. For instance, the ϵ-machines of computational mechanics may offer a more algorithmic approach, if suitably modified to generate the construction appropriate for our needs (Crutchfield, 1994; Shalizi and Crutchfield, 2001; Nerukh et al., 2002). Given the symbolisation of a time series and a few parameters, one can organize the states[4] of a system into a graph with probabilistic transitions between states: this is the ϵ-machine describing the time series. The adaptation of this to our case would be an exchange of the probabilities for the perturbations that induce the transitions.

It would also be interesting to see these methods applied to simple models that exhibit more complicated continuous interaction, such as those of the perceptual crossing paradigm (Izquierdo et al., 2022; Severino et al., 2023; Merritt et al., 2024). There, agents are evolved to find and stay near each other while avoiding stationary blocks and each other's "shadows" (blocks attached to an agent that cannot sense it, but that the other agent can). In this case, there is a more direct pressure for agents to be able to distinguish between conspecifics and other objects in the environment.

More fundamentally, though, the concept of a cognitive distinction in itself need not be discrete if, for instance, continuous variation in a perturbation results in continuous variation in behaviour. It is just that this situation cannot be satisfactorily captured by the graph-theoretic formalism we have used here. One must also consider whether such variation is interesting, i.e., whether it meets the conditions of sufficient differentiation we define (and here we again see that the observer-community plays an essential role). Such conditions could be as simple as changing state in a high-resolution discretisation, or as coarse-grained and task-relative as the ones we have used. In any case, how we interpret the description we thus derive depends on our interest; higher resolution may result in simply a more redundant form

[3]There is certainly an interesting connection to be drawn between Maturana and Varela of the biology of cognition and the later Wittgenstein, especially in the way they come to apparently resonant epistemologies. Needless to say, exploring that possibility is well beyond the present scope (the interested reader should see Hutto (2013)).

[4]Really, the states are themselves time series indexed to indicate a past, present, and future. They are essentially sliding windows over the given time series.

that, for us, would be equivalent to the task-relative description (a long chain of states with the same perturbation collapsed into a single edge). Hence, more comprehensive descriptions do not necessarily provide us more information.

There are a still a number of other limitations to the model we have presented, however. For one, the dimensionality of our perturbation space is both extremely small and prohibitively large. It is small in the sense that, in any given instant of time, perturbations can only vary along one dimension. Considered in our cognitive domain, this is with respect to distance from an agent. Considered in the agent's cognitive domain, with respect to the sensor value. But if we consider perturbations as having temporal extension, as we have here, then the space of continuous perturbations becomes infinite.

Let us first consider the space as infinite-dimensional. Even in the case where perturbations cannot easily be separated to form a finite set of temporally distinct patterns, it is not as though the perturbations observed in any given task fully explore the full infinite-dimensional space. Thus, it may be possible to find simple dimensions of variation in the perturbation structure that can simplify the analysis. For example, one might use duration, magnitude, frequency of oscillation, or time-averages of these.

Considered small, the dimensionality of the perturbation space seems to significantly limit the scope of the cognitive phenomena we can explore, even in principle. For instance, it is unclear how the appearance of distinct objects of experience could be described in terms of cognitive distinctions. But perhaps this is more a consequence of our failure to imagine how a method in its infancy could generalise to the most complicated of cases. Moreover, how we conceptualise our own experience is certainly not an uncontroversial matter (Marr, 2010; Varela et al., 2017; Dennett, 1992; Sheets-Johnstone, 2011). In particular, if we take seriously some of the implications of cognitive distinctions as a *phenomenological* method of description, we shift our focus from objects as 'things for us to see' to an experience of variations in our senses that afford certain ways of acting; the apparent fixedness of objects then becomes a consequence of the regularity of our actions (and thus in the structure of our cognitive domain). This is, of course, largely speculative at this point, but it should be clear that it would be more profitable to continue developing the methods presented here before deciding once and for all on their ultimate fecundity.

Another limitation of the method is one we have emphasised a number of times: cognitive distinctions serve certain ends better than others. This is most evident when we try to derive conditions of observation from the interaction graph. This will almost certainly fail. The problem is that the interaction graph does not specify what a particular class of behaviour should look like in our cognitive domain, and all the more so if we use an intrinsic representation akin to the one in Figure 8b. How we relate the graph to a spatio-temporal description of the agent's behaviour is our decision. However, it is not as though such descriptions are completely unconstrained. If we want to create a task in which certain distinctions are necessary to the solution, then having a number of graphs derived from agents in different situations may provide insight into how we could facilitate those distinctions. We may find that certain graph motifs correlate with certain task structures (e.g., loops in the graph corresponding to repeated action).

We now want to anticipate some potential criticism that threaten to make our argument trivial or irrelevant. One of these criticisms could be that "one just needs to design a better task, and the whole issue of failing to capture certain distinctions disappears." But let us ask in response: how does one evaluate a task in this respect? If we try to imagine how we would articulate an agent's failure to distinguish, without something akin to the method of description we propose, we seem to be stuck with, for instance, "It looks like it communicates, but in actuality, not quite." This is clearer when we cast this issue in the taxonomy of descriptions we have been using. Since designing a task, in this frame, is just generating a particular realization of some conditions of observation, there is nothing in it to directly specify the cognitive capacity of the agents. Thus, any modification to the task made in light of direct cognitive considerations is necessarily mediated by a change of descriptive method. We are then forced again to reckon with what methods we employ and how we mediate between them. Or, put simply, we cannot design a better task without some way to describe agents in cognitive terms, and so we should ensure that such descriptions, and changes thereof, are without basic epistemological error.

A related criticism is one asserting that "no fundamental change of language is necessary, so long as one is careful in analysis." While this point is stronger, it still fails for the same basic reasons as the previous one. For one, it ignores the fundamental role that language plays in guiding an investigation. Put another way, why would I look for whether particular distinctions are made? When we are left to look for verb-mechanism correspondences, or else to explain particular spatial trajectories, we have no reason to look for anything that challenges our intuitions. Further, both of these approaches fail to take the perspective of the agent into account. And again, if all we have are natural language terms and conditions of observation, the best that can be achieved is "this, but not quite." While neural and dynamical explanations may serve as the grounds on which we come to question the validity of a natural language description, they do not immediately suggest how we should articulate that subtly. When we lack a rigorous method of cognitive description, we sacrifice clarity in understanding the *cognitive* significance that our explanations have. This is not to say that previous models (or cognitive science in general) have said nothing useful about cognition, but rather that they lack a method in which to articulate this rigorously.

Conclusion

We end our discussion of cognitive distinctions by looking forward to what their actual implementation might look like. While we have mentioned their potential significance in broader domains, our particular concerns are with simple models of cognitive behaviour, as it is these that are most amenable to a formal treatment, in addition to their theoretical and practical utility (Beer, 1996, 1997, 2020). Thus, we envision a rough template that the construction and analysis of such models might follow. One general aspect of it would be the explicit mention of the descriptive methods employed, the point of using that method, and, most important, when changes in method are being made. To facilitate this, one might use a basic pattern to structure their investigation, in which ordinary language descriptions of cognitive behaviour are taken as guideposts to natural phenomena from which conditions of observation can be extracted. Then, after explaining how the system under investigation successfully satisfies those conditions (without reference to their origin), interaction graphs — or an alternative formalism — can be constructed and analysed. This would permit an evaluation of the significance of the operational explanations, as well as further comparisons among a population of cognitive domains to perhaps determine more general features of the cognitive structure inherent in the task.

We hope that the concept of cognitive distinctions, and the associated methods, can further enhance the theoretical power of these models to help us understand cognitive phenomena irrespective of our conceptual and linguistic prejudices — that is, to understand them *in their own terms*.

Data and Code Availability

All data and code for evolution, simulation, dynamical analysis, and graph generation for the results presented here can be found at `https://github.com/ThomasGaul/Cognitive-Distinctions-in-Referential-Communication`. Randall Beer's *Dynamica* package for Wolfram Mathematica was used for the dynamical analysis.

Acknowledgements

We would like to thank Haily Merritt, Gabriel Severino, and Randall Beer for their helpful comments and discussions.

References

Beer, R. D. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72(1–2):173–215.

Beer, R. D. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In Maes, P., Mataric, M. J., Meyer, J.-A., Pollack, J., and Wilson, S. W., editors, *Proceedings of the Fourth International Conference on the Simulation of Adaptive Behavior*, pages 421–429. MIT Press.

Beer, R. D. (1997). The dynamics of adaptive behavior: A research program. *Robotics and Autonomous Systems*, 20(2–4):257–289.

Beer, R. D. (2004). Autopoiesis and cognition in the Game of Life. *Artificial Life*, 10(3):309–326.

Beer, R. D. (2014). The cognitive domain of a glider in the Game of Life. *Artificial Life*, 20(2):183–206.

Beer, R. D. (2020). Lost in words. *Adaptive Behavior*, 28(1):19–21.

Campos, J. I. and Froese, T. (2017). Referential communication as a collective property of a brain-body-environment-body-brain system: A minimal cognitive model. In *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017 - Proceedings*, volume 2018-January, pages 1–8. IEEE.

Chittka, L. (2023). *The Mind of a Bee*. Princeton University Press.

Christensen, W. D. and Bickhard, M. H. (2002). The process dynamics of normative function. *The Monist*, 85(1):3–28.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1):7–19.

Crutchfield, J. P. (1994). The calculi of emergence: Computation, dynamics and induction. *Physica D: Nonlinear Phenomena*, 75(1–3):11–54.

Dennett, D. C. (1992). *Consciousness Explained*. Back Bay Books.

Fox, R. and Bullock, S. (2023). Nectar of the bots: Evolving bidirectional referential communication. *Adaptive Behavior*, 31(1):65–86.

Frisch, K. v. (1967). *The Dance Language and Orientation of Bees*. Harvard University Press.

Garfinkel, A., Shevtsov, J., and Guo, Y. (2017). *Modeling Life: The Mathematics of Biological Systems*. Springer.

Hutto, D. D. (2013). Enactivism, from a Wittgensteinian point of view. *American Philosophical Quarterly*, 50(3):281–302.

Izquierdo, E. J., Severino, G. J., and Merritt, H. (2022). Perpetual crossers without sensory delay: Revisiting the perceptual crossing simulation studies. In Holler, S., Löffler, R., and Bartlett, S., editors, *ALIFE 2022: 2022 Conference on Artificial Life*, pages 27–36. MIT Press.

Manicka, S. (2012). Analysis of evolved agents performing referential communication. In Adami, C., Bryson, D. M., Ofria, C., and Pennock, R. T., editors, *ALIFE 2012: The Thirteenth International Conference on the Synthesis and Simulation of Living Systems*, pages 393–400. MIT Press.

Marr, D. (2010). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. MIT Press.

Maturana, H. R. and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. D. Riedel Publishing.

Maturana, H. R. and Varela, F. J. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston: Shambala.

Merritt, H., Severino, G. J., and Izquierdo, E. J. (2024). The Dynamics of Social Interaction Among Evolved Model Agents. *Artificial Life*, 30(2):216–239.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories: New Foundations For Realism*. MIT Press.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of science*, 56(2):288–302.

Nerukh, D., Karvounis, G., and Glen, R. C. (2002). Complexity of classical dynamics of molecular systems. i. methodology. *Journal of Chemical Physics*, 117(21):9611–9617.

Newell, A. and Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Commun. ACM*, 19(3):113–126.

Oyama, S. (2000). *The Ontogeny of Information: Developmental Systems and Evolution*. Duke University Press.

Severino, G. J., Merritt, H., and Izquierdo, E. J. (2023). Between you and me: A systematic analysis of mutual social interaction in perceptual crossing agents. In Iizuka, H., Suzuki, K., Uno, R., Damiano, L., Spychala, N., Aguilera, M., Izquierdo, E., Suzuki, R., and Baltieri, M., editors, *ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference*, pages 176–184.

Shalizi, C. R. and Crutchfield, J. P. (2001). Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of statistical physics*, 104(314):817–879.

Sheets-Johnstone, M. (2011). *The Primacy of Movement*. John Benjamins Publishing Company.

Strogatz, S. H. (2018). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. CRC Press, second edition.

Van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21(5):615–628.

Varela, F. J. (1979). *Principles of Biological Autonomy*. New York: North Holland.

Varela, F. J., Thompson, E., and Rosch, E. (2017). *The Embodied Mind, Revised Edition: Cognitive Science and Human Experience*. MIT Press, revised edition.

Von Uexküll, J. (1992). A stroll through the worlds of animals and men: A picture book of invisible worlds. *Semiotica*, 89(4):319–391.

Williams, P. L., Beer, R. D., and Gasser, M. (2008). Evolving referential communication in embodied dynamical agents. In Bullock, S., Noble, J., Watson, R., and Bedau, M., editors, *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 702–709. MIT Press.

Winograd, T. and Flores, F. (1986). *Understanding Computers and Cognition: A New Foundation for Design*. Norwood, NJ: Ablex Publishing.

Wittgenstein, L. (1953). *Philosophical Investigations.* Oxford: Blackwell, fourth edition.

Yao, S., Nunley, J., and Izquierdo, E. J. (2023). Go by its name: Evolution and analysis of conceptual referential communication. In Iizuka, H., Suzuki, K., Uno, R., Damiano, L., Spychala, N., Aguilera, M., Izquierdo, E., Suzuki, R., and Baltieri, M., editors, *ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference*, pages 457–465.