解　説

Grounding Symbols through Sensorimotor Integration

Karl F. MacDorman* *Osaka University, Department of Systems and Human Science

1. Grounding symbol systems

A prominent robotics professor surprised me at last year's RSJ conference: "There isn't really a symbol grounding problem for robotics, is there? I often ask people, 'Is symbol grounding a problem for your research?' and no one says, 'Yes.'" Sensing irony in his voice, I replied, "That's because no one is building systems with a human — or even vertebrate — level of competence. When they try to, they may respond to your question differently."

Harnad (1990) identified the symbol grounding problem as a problem of embodying symbol systems (Newell 1980). Symbol systems are usually comprised of a fixed set of elementary symbols and rules for combining them into more complex representations. Symbol systems are designed to manipulate these representations according to their constituent structure (i.e., syntax) in the hope that these transformation make sense semantically. The problem is: How do we causally connect a symbol system's internal symbols to the external objects, events, and relations they are supposed to represent?

Solving this problem is not simply a matter of learning *some* kind of grounded internal representation in a flexible and adaptable way. We are already building emergent systems that can do that (e.g., Uchibe et al. 1998), and artificial neural networks (Smolensky 1988), reinforcement learning (Sutton & Barto 1998), and behavior-based architectures (Brooks 1991) offer three popular approaches. But what the current generation of emergent systems generally lacks are three key aspects of thinking: its systematicity, productivity, and inferential coherence. They lack these aspects because their underlying methods have been unable to deal effectively with constituent structure, though more elaborate implementations should overcome this limitation (see, for example, Chalmers 1993).

We tend to think of these aspects of thinking as being typically human. This is probably because they were first studied in relation to language (Katz & Fodor 1963; Chomsky 1959, 1965). Human thinking is systematic insofar as people who can understand one sentence (e.g., *John loves Mary*) can, in general, understand structurally similar sentences (*Mary loves John*); it is productive insofar as people can understand and generate an unbounded number of sentences; and it is inferentially coherent insofar as people are capable of forming valid conclusions from valid premises. Nevertheless, Fodor and Pylyshyn (1988) have shown that the same arguments that have been applied to language understanding can be extended to belief in general and even nonverbal perception and cognition in animals. For example, it seems reasonable to assume that if a kitten can perceive that a ball of yarn is to the left of its mother (yLm), in general it can also perceive that its mother is to the left of a ball of yarn (mLy). But without constituent structure, a robot can't abstract the relation L. It can only represent these two different situations as two distinct states. Thus, a robot that can recognize the first situation may be blind to the second, simply because it had no prior experience of it.

If states lack constituent structure, situations cannot be decomposed into parts. Thus, to avoid perceptual aliasing, every behaviorally distinct situation must have its own state. In the above example, if we add a new relation (e.g., *above*), the number of relational states doubles: yLm, mLy, yAm, mAy. If we instead add a new object, it increases six fold: yLm, mLy, yLx, xLy, mLx, xLm, yLmLx, mLyLx, yLxLm, xLyLm, xLmLy, mLxLy. A linear increase in the number of basic categories our robot is trying to represent results in an exponential increase in the number of states. If we consider, for example, the reinforcement learning paradigm, this translates into an exponential increase in the time it takes for learning to converge, since convergence depends on actions being attempted repeatedly in each state. But in humans the relationship seems to be the opposite. People who know more don't take longer to think!

Fodor and Pylyshyn (1988) note two reasons why symbol systems are able to simulate systematic, productive, and inferentially coherent modes of thinking (pp. 28-30): *(1)* The syntax of a representation can encode its role in inference; and *(2)* computers can be programmed to manipulate representations according to their syntax.

The basic problem with this arrangement is that symbol manipulation depends solely on properties that are part of the system's *internal* workings: how com-

原稿受付　1998 年 9 月 31 日
キーワード：Affordances, Cognitive Robotics, Consciousness, Frame Problem, Machine Learning, Symbol Grounding
*Kisokogakubu, Toyonaka, Osaka 〒 560-8531

puter hardware implements the system's syntactic constraints. But the sensorimotor relation between a robot's body and the *external* environment must be able to influence the causal relation between its internal symbols and the external state of affairs they represent (see also Tani, 1996). Standard artificial approaches have failed because they try to set up symbol–object connections in advance — by means of a fixed set of feature detectors. But outside of sterile laboratory environments this doesn't work well. Real environments can change unpredictably as can real bodies. This alters what the robot can sense and do — and how it should 'think': its opportunities for interaction and the kinds of sensorimotor invariance available to recognize them. J. J. Gibson (1979) called these opportunities *affordances*. To illustrate how affordances vary according to body and environment, note that while a telephone pole serves nicely as a bird's perch (as long as it can fly), it may only be an obstacle to an earth-bound robot.

From a cognitive standpoint, traditional symbol systems are unsatisfying because they are nearly inexplicable in terms of evolution, neuroanatomy, and psychology (MacDorman 1997a, b; Schyns et al. 1998). In addition, they create a practical difficulty for their designers, who are required to anticipate all the elementary features the symbol systems will need to recognize and provide them with feature detectors to match. Even if this were possible, the computational demands of processing representations composed from elementary features may be prohibitively high (Janlert 1996). Psychological and neurophysiological evidence suggest that we can learn to recognize large coarse features as well as small detailed ones by means of feedback from miscategorization (Harnad 1987; MacDorman 1997b, 1998; Schyns et al. 1998). But the main point is that dynamically changing sensorimotor relations need to constrain abstract reasoning; syntactic constraints are not enough.

2. Parallel processing, conscious integration, and the frame problem

In robotics, a symbol system fits roughly in the middle of the sense-model-plan-act architecture — the part where there is symbolic modeling, reasoning, and planning (e.g., Nilsson 1984). Rodney Brooks (1991) took a radical position against this approach and especially against the centralized use of representations as typified by the symbol system. There seemed to be too many steps, and hence delays, between sensing and action. In addition, the placement of the symbol system created a bottleneck. All information must flow through it. Since the symbol system must represent everything that needs to be represented, processes can't run in parallel. This leads to behavior that is slow and deliberative — ill suited to dynamic environments where a quick response is needed.

In response, Brooks proposed the subsumption architecture, in which each processing layer constitutes a behavior (e.g., *wander, avoid obstacles, track ball, shoot goal*). Layers run in parallel with minimal interaction. They enjoy a tight coupling with sensing and action by directly using the robot's sensing of the environment as a point of reference instead of a centralized representation. This makes for fast, reactive behavior.

However, it is unclear how Brooks' purposes-build robots could adapt to changing affordances. To assert that the subsumption architecture doesn't suffer from the symbol grounding problem is like saying that invertebrates don't have backache. Purely behavior-based robots don't possess the same competencies that a symbol system does. Some hybrid approaches have tried to graft a symbol system layer on to a behavior-based architecture (e.g., Malcolm 1995). But so long as the symbol system operates only under internal syntactic constraints, hybrid architectures will run into all the usual problems involved with not letting sensorimotor constraints bear on abstract reasoning.

To deal effectively with new situations a robot needs to model its affordances so that it can test its actions against a model *before* testing them against the world. In this way, the robot doesn't have to jump off a cliff before discovering that this is dangerous; it can recognize the affordance and let its hypothesis about moving toward the cliff action die in its place (see Dennett 1996). A centralized representation may in fact form the core of a robot's affordance model, serving as a global conceptualization.

Some have argued that intelligence does not require a global conceptualization; it can just emerge from the separate interactions of simple units (e.g., neurons). But these two views do not conflict. Simple units together constitute a global conceptualization to the extent that their separate interactions foster global coherence and integration among separate bits of information. The conceptualization is centralized only in the sense that its parts are locked in mutual dependence. It is not a matter of mere proximity in space and time, since constraints can propagate in a distributed fashion.

Nevertheless, a global conceptualization exacts a high computational cost: the cost of integrating and maintaining the coherence of many different kinds of amodal and multimodal information while making them explicitly available to other processes including those involving abstract reasoning (MacDorman in press). This cost makes a global conceptualization an inherently limited resource. Implementing it with a traditional symbol system results in the *frame problem* (McCarthy & Hayes 1969). This is not only because it creates a processing bottleneck but because traditional symbol systems are underconstrained (Harnad 1993). Also the fact that they only have syntactic constraints means that they can represent anything that is logically possible including a limitless number of absurd concepts. Thus, time is wasted reasoning about events that can never occur (Fodor 1987). This problem results from an excess of freedom in the symbol system's representational form. A robot cannot overcome it by trying to figure out what not to reason about. If it must do this, it is already caught by the frame problem because it is already reasoning about things that don't change (Janlert 1996).

While biological systems can still be susceptible to the frame problem, they have finessed it in two ways.

First, reasoning in animals is not purely logical or syntactic. More importantly, the reasoning process itself depends on its relation to the world. The development of internal categories and transformations on those categories depends on external interaction. Animals are able to make grounded distinctions because evolution and learning have abstracted them in a bottom-up fashion from a flow of sensorimotor information spanning the life of the organism and its ancestors. In this sense all embodied reasoning is empirically constrained. It is also functionally constrained by the inherent characteristics of neurons, sense organs, biomechanics, and so on. If a robot's representational form can be made to include empirical constraints, it won't have to waste time reasoning about what is empirically impossible.

The second way biological systems have finessed the frame problem is by separating, automating, and parallelizing routine behavior patterns. These patterns are off-loaded to nonconscious processes, which may be partially encapsulated from one another. Thus, conscious integration is applied where it is needed most, in dealing with new and unexpected situations. Oddly enough, a global conceptualization appears to be key to learning and automating routine activities. Once automated, these activities may be performed in parallel with conscious processing.

The pros and cons of a global conceptualization are made apparent by considering a child learning to walk. Walking is a form of controlled falling, and it is not just the step movement that is important, but timing it to catch the fall while in other respects maintaining balance. At first, trying to walk occupies the child's full attention. The child makes each step with conscious and deliberate effort. The leg's current and target positions are explicitly represented as part of a global conceptualization. This representation is consciously accessible. We generally associate it with the child's experience. As the child becomes good at walking — as this behavior becomes automatic — the child's mind is freed up to think about other things, to play games, kick a ball, chase the cat, and to talk. In this way, precious cognitive resources are not squandered on the time-tested and habitual.

It is easy to extend this analysis of learning to walk to learning virtually anything: flying a kite, casting a net, riding a bike, typing, playing tennis, or driving a car. Brooks' subsumption architecture is like a robot driving a car by instinct. The behavior is impressive, but it did not learn it, and it cannot learn anything new. By contrast, the traditional sense-model-plan-act architecture is like a robot trying to steer a car, shift gears, watch the road, and talk to a passenger without having had any practice beforehand. It runs into the frame problem. It is not possible to do so many things at once consciously, given the high computational cost of global integration. We need a new approach which can combine the advantages of having a global conceptualization with those of having habitual behaviors, running in parallel, that are tightly coupled with sensing and action.

3. Predictions for affordances

Recognizing an affordance entails recognizing sensorimotor invariance that is correlated with that affordance. Gibson believed that we directly perceive invariance in the optical array (hence, his theory of *direct perception*). It is not surprising that Gibson underestimated the computational complexity of vision, since he wrote before researchers had begun to explore it seriously. We now know that invariance often exists only at a high degree of abstraction, far removed from raw sensorimotor data. Thus, the brain may need to process sensorimotor data extensively and to spend time learning what kinds of invariance are useful in recognizing an affordance. Intelligent creatures must also be able to discover new affordances, for example, by detecting the consequences of their actions or by generalizing from other similar experiences.

Gibson had little to say about the internal workings of a system that could model affordances. Let us roughly sketch what form this kind of system might take. If we build sufficient empirical constraints into our theory, it should be able to guide the construction of robots whose cognitive processes have sufficient empirical constraints. The aim is to build robots that can develop their own grounded symbols while avoiding the need to reason about stabilities.

An intelligent robot can discover the various interactions and effects that its environment affords by learning spatiotemporal correlations in its sensory projections, motor signals, and internal variables. These correlations are a kind of embodied prediction about the future. They integrate sensorimotor information from various sources and modalities and can be learned from experience. The consequences a robot predicts for its potential actions depend on its perceived circumstances. Embodied predictions are, in this sense, conditional. Thus, at any given moment, sensorimotor information (as well as internal processes like reasoning and remembering) will activate only a subset of the robot's learned predictions. Currently active predictions constitute the robot's affordance model (see Figure 1).

Active predictions expedite anticipatory responses to prepare the robot for what is likely to happen next. Should that event either fail to occur or fail to be per-

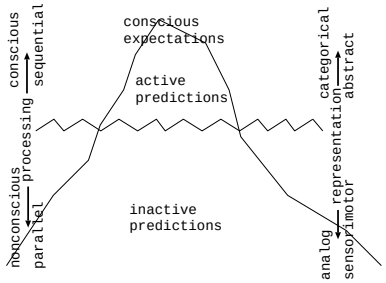

図 1 The robot's currently active predictions constitute its affordance model. Those predictions that enjoy a high degree of global integration are perhaps analogous to conscious expectations.

ceived, orienting responses direct attention to the source of error. The robot then revises old predictions and develops new ones to account for the unexpected occurrence. In this way, attention helps to feed information concerning miscategorization back into the process of learning to recognize affordances.

Embodied predictions that support routine patterns of behavior become parallelized, automatic, and nonconscious. This is necessary to free up conscious resources, but it means that it is easy for even humans to misapply them in new situations. You might, for example, walk to where your bicycle is usually parked as a matter of habit, only to discover that the bicycle rack is empty. As conscious resources become focused on the matter, you may recall that you had to park the bicycle somewhere else because you arrived late and the gate was locked. Violated predictions bring the expressive freedom and representational power of conscious processing to bear on a previously unexpected occurrence, such as the discovery of the empty bicycle rack. In this way, they can make sense of the occurrence and embed it in a new context, so it can be reautomated.

The few mistakes we make because we must do most things nonconsciously are a small price to pay for overcoming the frame problem. Otherwise, we too might waste most of our time reasoning about things that normally do not change. Perhaps Alfred Whitehead was right to say that civilization advances by increasing the number of tasks that can be performed without thinking about them.

4. Planned implementation of Ψro

We now introduce the mobile robot Ψro, which we are designing to recognize affordances and whose separate modules we have tested independently. Exploiting an affordance provides internal and external feedback. This feedback helps Ψro learn to discriminate affordances by developing predictions concerning how its interactions transform physiological variables and sensory projections from distal objects.

4.1 Segmentation and tracking

Ψro's goal is survival. It must intercept tasty robots and avoid poisonous and dangerous robots in a cluttered dynamic environment. Ψro uses motion information to segment and track potential sources of invariance. Isard and Blake's (1998) method of conditional density propagation for visual tracking is particularly useful, since it can clearly segment a moving object while tracking it.

4.2 Learning affordance categories

Preprocessing. Once the robot has segmented the potential source of invariance, it converts it to a canonical form. This highlights invariance and facilitates comparison between different segmented images. The process involves *(1)* removing the background, *(2)* scaling the segmented image to fit on a 64-by-64 grid, *(3)* recoding color information in terms of an intensity, red-versus-green, and blue-versus-yellow channel, *(4)* decomposing the recoded image into a set of wavelet coefficients, and *(5)* quantizing the coefficients, retaining only the largest in absolute magnitude, to form a compact signature for each segmented image.

Neurophysiological evidence supports opponent process recoding in the brain, and empirical research in image querying suggests that it is more useful for categorization than other coding schemes. The wavelet transform and other multiresolution techniques are useful because, at any given scale, it is often hard to find invariant features. The wavelet transform, when performed with a parameterized family of two dimensional Gabor filters, is also neurophysiologically plausible: these filters match the receptive field profiles of 97% of simple cells in the cat visual cortex.

Learning categorical representations. While Ψro is tracking a potential source of invariance, it is calculating and accumulating image signatures. Internal feedback gives Ψro the affordance when it makes contact with it. The robot then creates a *categorical representation* (Harnad 1987) by statistically filtering out all signature values except those that tend not to vary among signatures of the same affordance category but vary among signatures of different affordance categories. Once Ψro has learned some categorical representations, it predicts the affordance from the representation that best matches the image signatures. If Ψro miscategorizes, it refines its categorical representations accordingly and may learn several representations in order to discriminate the same affordance.

4.3 Learning a sensorimotor model

Although we may conceive of certain concepts in purely abstract terms, it is unlikely that the brain represents anything in a way that is completely free of empirical, sensorimotor constraints. Learning these constraints is especially important in making rapid, graceful, well-coordinated movements. This is because error signals often become available only after a movement has completed (for example, when trying to score a goal in soccer) and the physical dynamics of a body sometimes change unpredictably.

Ψro learns its sensorimotor model by developing predictions concerning how motor signals transform sensory projections. Ψro's predictions are generalized from its past sensorimotor interactions. The robot uses a k-D tree (Sproull 1991) to represent these experiences as points in a multidimensional phase space. Ψro predicts how a new visual location maps onto its motor subspace by local linear interpolation: New visual locations are projected onto the motor subspace by means a coordinate system determined by closest points in the image plane subspace. Learning occurs when predictions fail. New points are then added to the phase space and, if Ψro's sensorimotor dynamics have changed, old points are updated or discarded.

4.4 Navigation

Ψro uses its learned sensorimotor model to plan paths to potential affordances in a cluttered environment. The sensorimotor effects of chains of actions may be described by a phase space similarly to how we described those of single actions. Ψro's search is highly constrained because the robot quantizes the phase space into distinct hypercubes and then uses dynamic programming to avoid recomputing subpaths between hy-

percubes.

The robot attempts actions 'mentally' beginning from its starting position. It repeats this process among those hypercubes it has examined according to their priority. Hypercubes that are closest to the goal in distance and closest to the start in time have the highest priority. If Ψro has very little time to plan, its path somewhat resembles one calculated by local hill climbing. If the robot has a little more time, the path becomes smoother and approaches a coarsely optimal solution, even though its calculation takes just a fraction of the time.

Like Ψro, potential affordances are often moving. Yasushi Yagi has studied the problem of learning a predictive model that could provide our planner with a probability field of locations that moving objects are likely to occupy. Integrating such a model into our robot results in planning that is proactive. A further benefit is that the robot does not have to recalculate paths as frequently.

References

[1] Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*, 139-159.

[2] Chalmers, D. J. (1993). Connectionism and compositionality: Why Fodor and Pylyshyn were wrong. *Philosophical Psychology, 6*(3), 305-319.

[3] Chomsky, N. (1959). Review of B. F. Skinner's Verbal Behavior. *Language, 35*, 26-58.

[4] Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

[5] Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America, 2*(7), 1160-1169.

[6] Dennett, D. C. (1996). *Kinds of minds: Towards an understanding of consciousness.* London: Weidenfeld & Nicolson.

[7] Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1-2), 3-71.

[8] Fodor, J. A. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres. In Z. W. Pylyshyn (Ed.), *The robot's dilemma: The frame problem in artificial intelligence.* Norwood, NJ: Ablex.

[9] Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston, MA: Houghton Mifflin.

[10] Harnad, S. (1987). Category induction and representation. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition.* Cambridge: Cambridge University Press.

[11] Harnad, S. (1990). The symbol grounding problem. *Physica D, 42*(1-3), 335-346.

[12] Harnad, S. (1993). Problems, problems: The frame problem as a symptom of the symbol grounding problem. *Psycholoquy, 4*(34). frame-problem.¶11.

[13] Isard, M. & Blake, A. (1998). Condensation: Conditional density propagation for visual tracking. *International Journal of Computer Vision, 29*(1), 5–28.

[14] Janlert, L.-E. (1996). The frame problem: Freedom of stability? With pictures we can have both. In K. M. Ford & Z. W. Pylyshyn, *The robot's dilemma revisited: The frame problem in artificial intelligence.* Norwood, NJ: Ablex.

[15] Katz, J. J. & Fodor, J. A. (1963). The structure of semantic theory. *Language, 39*(2), 170-211.

[16] MacDorman, K. F. (1997a). How to ground symbols adaptively. In S. O'Nuallain, P. McKevitt & E. MacAogain, *Readings in computation, content and consciousness.* Amsterdam: John Benjamins.

[17] MacDorman, K. F. (1997b). *Symbol grounding: Learning categorical and sensorimotor predictions for coordination in autonomous robots.* Technical Report No. 423. Computer Laboratory, Cambridge (e-mail librarian@cl.cam.ac.uk).

[18] MacDorman, K. F. (1998). Feature learning, multiresolution analysis, and symbol grounding. *Behavioral and Brain Sciences, 21*(1), 32.

[19] MacDorman, K. F. (in press). Extending the medium hypothesis: The Dennett-Mangan controversy and beyond. (See http://www.me.es.osaka-u.ac.jp/∼kfm)

[20] Malcolm, C. M. (1995). The SOMASS system: A hybrid symbolic and behaviour-based system to plan and execute assemblies by robot. In J. Hallam, et al. (Eds.), *Hybrid Problems, Hybrid Solutions*, pp. 157-168. Oxford: ISO Press.

[21] McCarthy, J. & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence* (vol. 4, pp. 463-502). Edinburgh: University of Edinburgh Press.

[22] Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135-183.

[23] Nilsson, N. J. (1984). Shakey the robot. Technical Report No. 323. SRI AI Center, Menlo Park, CA.

[24] Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences, 21*(1), 1-17.

[25] Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*(1), 1-74.

[26] Sproull, R. F. (1991). Refinements to nearest-neighbor searching in k-d trees. *Algorithmica, 6*(4), 579-589.

[27] Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

[28] Tani, J. (1996). Model-based learning for mobile robot navigation from the dynamical systems perspective. *IEEE Transactions on System, Man and Cybernetics*, Part B (Special Issue on Robot Learning), *26*(3), 421-436.

[29] Uchibe, E., Asada, M. & Hosoda, K. (1998). State space construction for behavior acquisition in multi-agent environments with vision and action, *Proceedings of the International Conference on Computer Vision,* pp. 870-875.

Karl F. MacDorman

Presently a Lecturer at Osaka University, Karl MacDorman received his Bachelor's in computer science at the U. C., Berkeley and his Ph.D. in machine learning and robotics at Cambridge. His research interests include sensorimotor representation, robot communication, and computational neuroscience. Further details including publications are available from his website: http://www.me.es.osaka-u.ac.jp/∼kfm