a formal framework for the symbol grounding problem

benjamin johnston + mary-anne williams


mouse

wiffle_blag

Physica D 42 (1990) 335–346
North-Holland

THE SYMBOL GROUNDING PROBLEM

Stevan HARNAD
Department of Psychology, Princeton University, Princeton, NJ 08544, USA

There has been much discussion recently about the scope and limits of purely symbolic models of the mind and about the proper role of connectionism in cognitive modeling. This paper describes the "symbol grounding problem": How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols? The problem is analogous to trying to learn Chinese from a Chinese/Chinese dictionary alone. A candidate solution is sketched: Symbolic representations must be grounded bottom-up in nonsymbolic representations of two kinds: (1) *iconic representations*, which are analogs of the proximal sensory projections of distal objects and events, and (2) *categorical representations*, which are learned and innate feature detectors that pick out the invariant features of object and event categories from their sensory projections. Elementary symbols are the names of object and event categories, assigned on the basis of their (nonsymbolic) categorical representations. Higher-order (3) *symbolic representations*, grounded in these elementary symbols, consist of symbol strings describing categorical membership relations (e.g. "An X is a Y that is Z"). Connectionism is one natural candidate for the mechanism that learns the invariant features underlying categorical representations, thereby connecting names to the proximal projections of the distal objects they stand for. In this way connectionism can be seen as a complementary component in hybrid nonsymbolic/symbolic model of the mind, rather than a rival to purely symbolic modeling. Such a hybrid model would not have an autonomous symbolic "module," however, because the symbolic functions would depend on the nonsymbolic representations. Symbol manipulation would be governed not just by the arbitrary shapes of the symbol tokens, but by the nonarbitrary shapes of the icons and category invariants in which they are grounded.

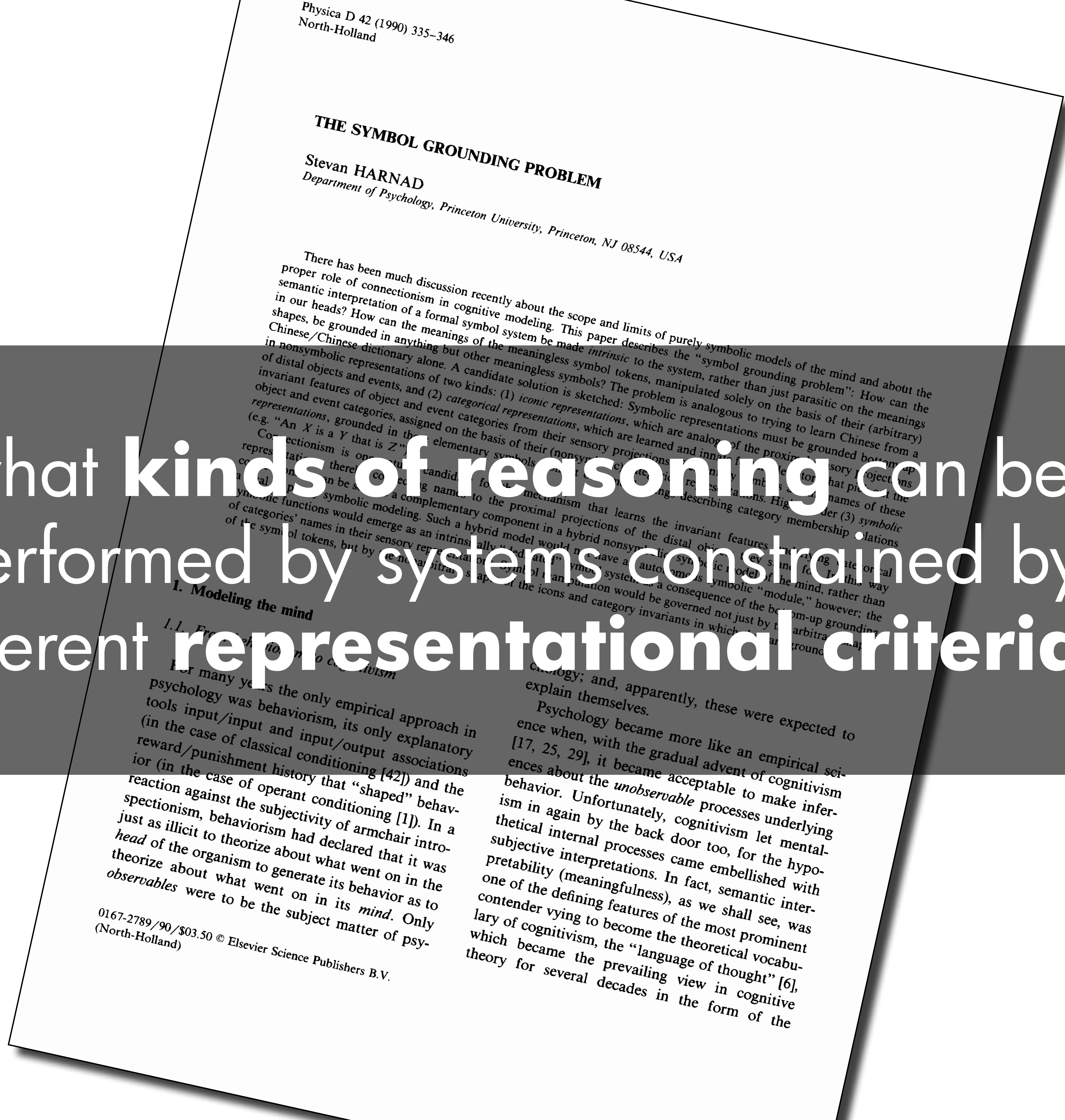
1. Modeling the mind

1.1. From behaviorism to cognitivism

For many years the only empirical approach in psychology was behaviorism, its only explanatory tools input/input and input/output associations (in the case of classical conditioning [42]) and the reward/punishment history that "shaped" behavior (in the case of operant conditioning [1]). In a reaction against the subjectivity of armchair introspectionism, behaviorism had declared that it was just as illicit to theorize about what went on in the *head* of the organism to generate its behavior as to theorize about what went on in its *mind*. Only *observables* were to be the subject matter of psychology; and, apparently, these were expected to explain themselves.

Psychology became more like an empirical science when, with the gradual advent of cognitivism [17, 25, 29], it became acceptable to make inferences about the *unobservable* processes underlying behavior. Unfortunately, cognitivism let mentalism in again by the back door too, for the hypothetical internal processes came embellished with subjective interpretations. In fact, semantic interpretability (meaningfulness), as we shall see, was one of the defining features of the most prominent contender vying to become the theoretical vocabulary of cognitivism, the "language of thought" [6], which became the prevailing view in cognitive theory for several decades in the form of the

"how can the semantic interpretation of a formal symbol system be made **intrinsic to the system**, rather than just parasitic on the meanings in our heads?"



Physica D 42 (1990) 335–346
North-Holland

THE SYMBOL GROUNDING PROBLEM

Stevan HARNAD
Department of Psychology, Princeton University, Princeton, NJ 08544, USA

There has been much discussion recently about the scope and limits of purely symbolic models of the mind and about the proper role of connectionism in cognitive modeling. This paper describes the "symbol grounding problem": How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols? The problem is analogous to trying to learn Chinese from a Chinese/Chinese dictionary alone. A candidate solution is sketched: Symbolic representations must be grounded bottom-up in nonsymbolic representations of two kinds: (1) *iconic representations*, which are analogs of the proximal sensory projections of distal objects and events, and (2) *categorical representations*, which are learned and innate [illegible]

1. Modeling the mind

1.1. [illegible]

For many years the only empirical approach in psychology was behaviorism, its only explanatory tools input/input and input/output associations (in the case of classical conditioning [42]) and the reward/punishment history that "shaped" behavior (in the case of operant conditioning [1]). In a reaction against the subjectivity of armchair introspectionism, behaviorism had declared that it was just as illicit to theorize about what went on in the *head* of the organism to generate its behavior as to theorize about what went on in its *mind*. Only *observables* were to be the subject matter of psychology; and, apparently, these were expected to explain themselves.

Psychology became more like an empirical science when, with the gradual advent of cognitivism [17, 25, 29], it became acceptable to make inferences about the *unobservable* processes underlying behavior. Unfortunately, cognitivism let mentalism in again by the back door too, for the hypothetical internal processes came embellished with subjective interpretations. In fact, semantic interpretability (meaningfulness), as we shall see, was one of the defining features of the most prominent contender vying to become the theoretical vocabulary of cognitivism, the "language of thought" [6], which became the prevailing view in cognitive theory for several decades in the form of the

what **kinds of reasoning** can be performed by systems constrained by different **representational criteria**?

formal framework

method: assume formal primitives for
semantic interpretability and
the **universe of concepts**,
then **mathematically** model the SGP

not a **complete** solution,
but **simplification** + **real** progress

no math today: details in the proceedings

context-free semantically interpretable representation $\mathcal{CFSIR}$

semantically interpretable representation $\mathcal{SIR}$

iconic and symbolic representation $\mathcal{ISR}$

distributed representation $\mathcal{DR}$

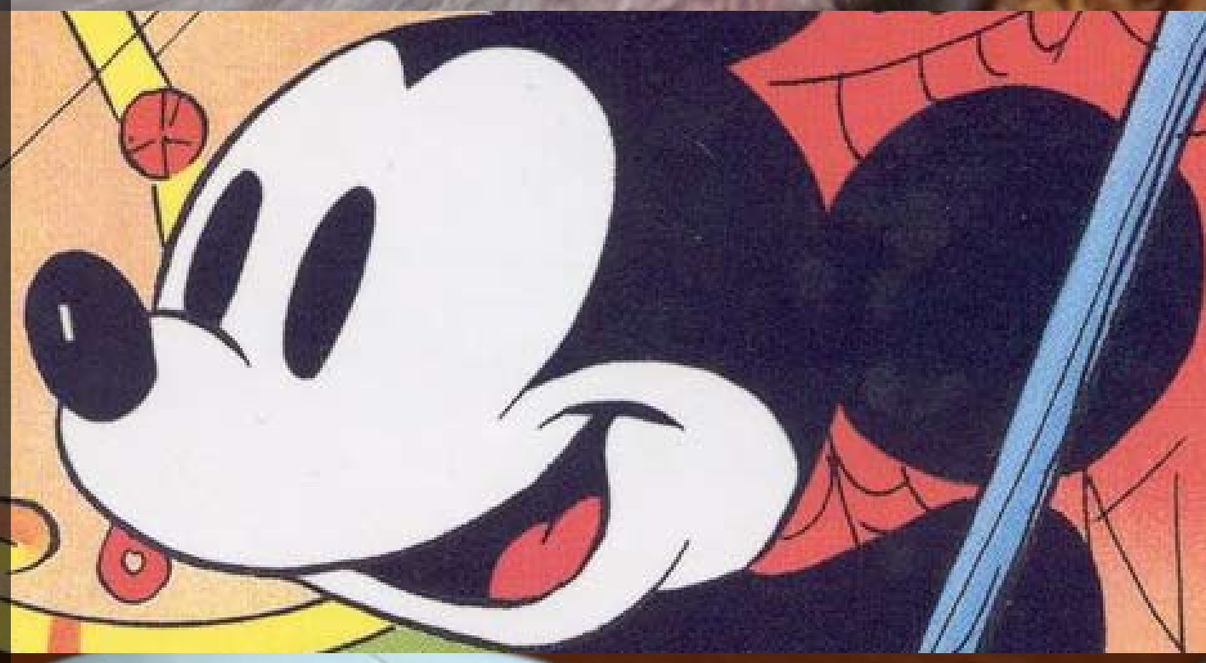
unconstrained representation $\mathcal{UR}$

non-formal $\mathcal{NF}$

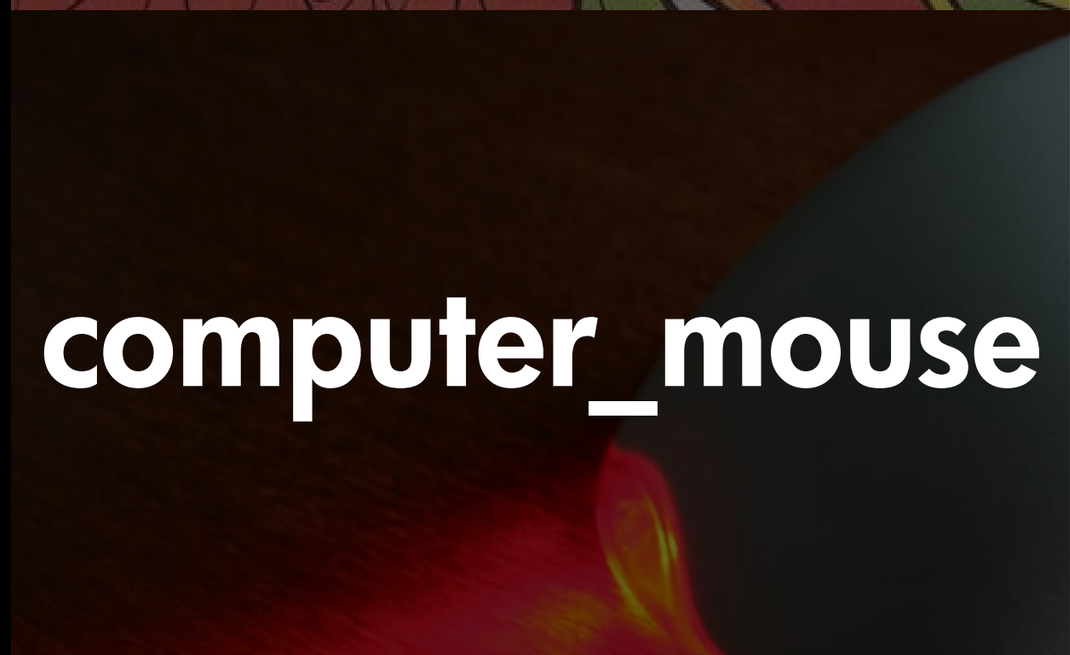
context-free semantically interpretable representation

pet_mouse

mickey_mouse



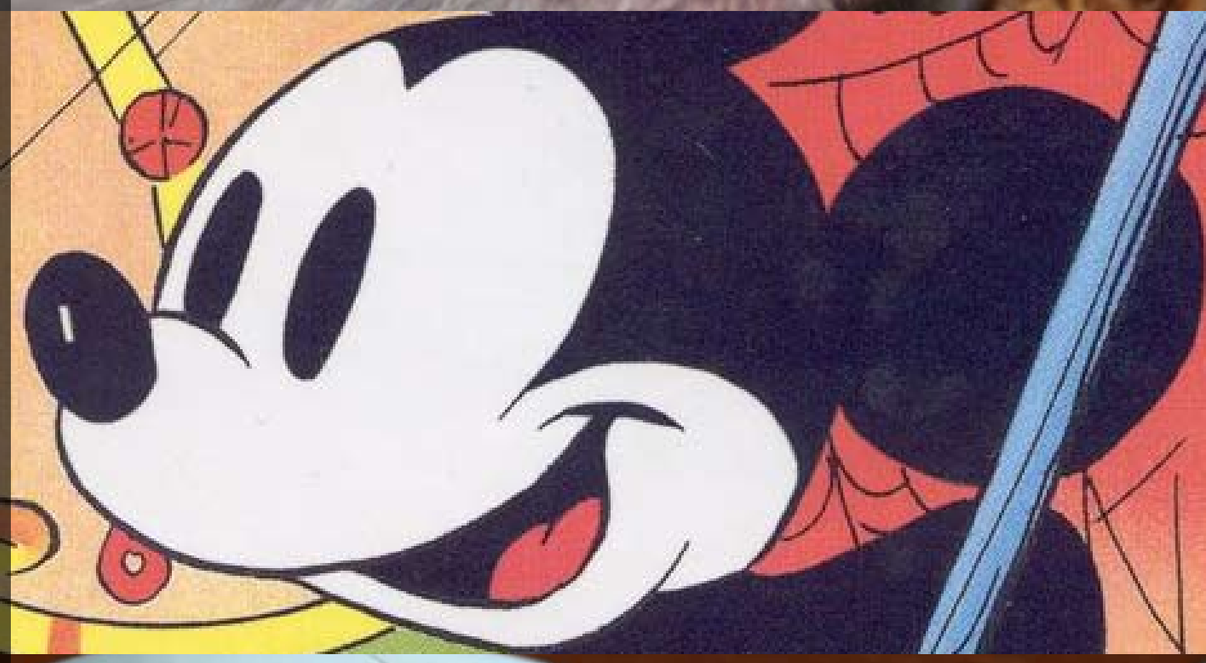
computer_mouse



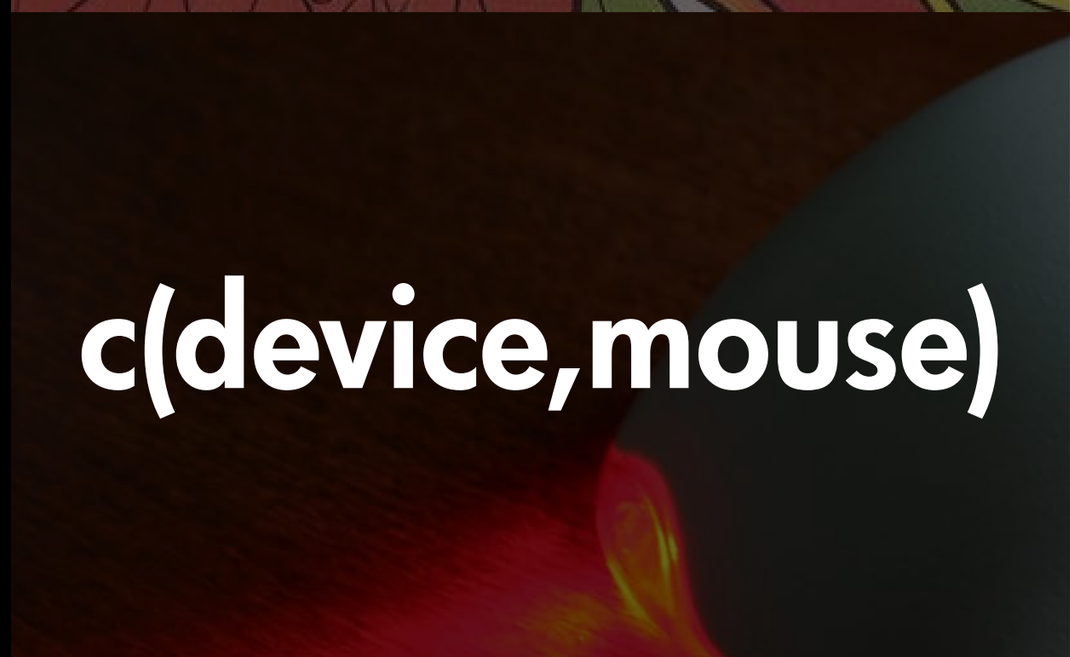
semantically interpretable representation

c(animal,mouse)

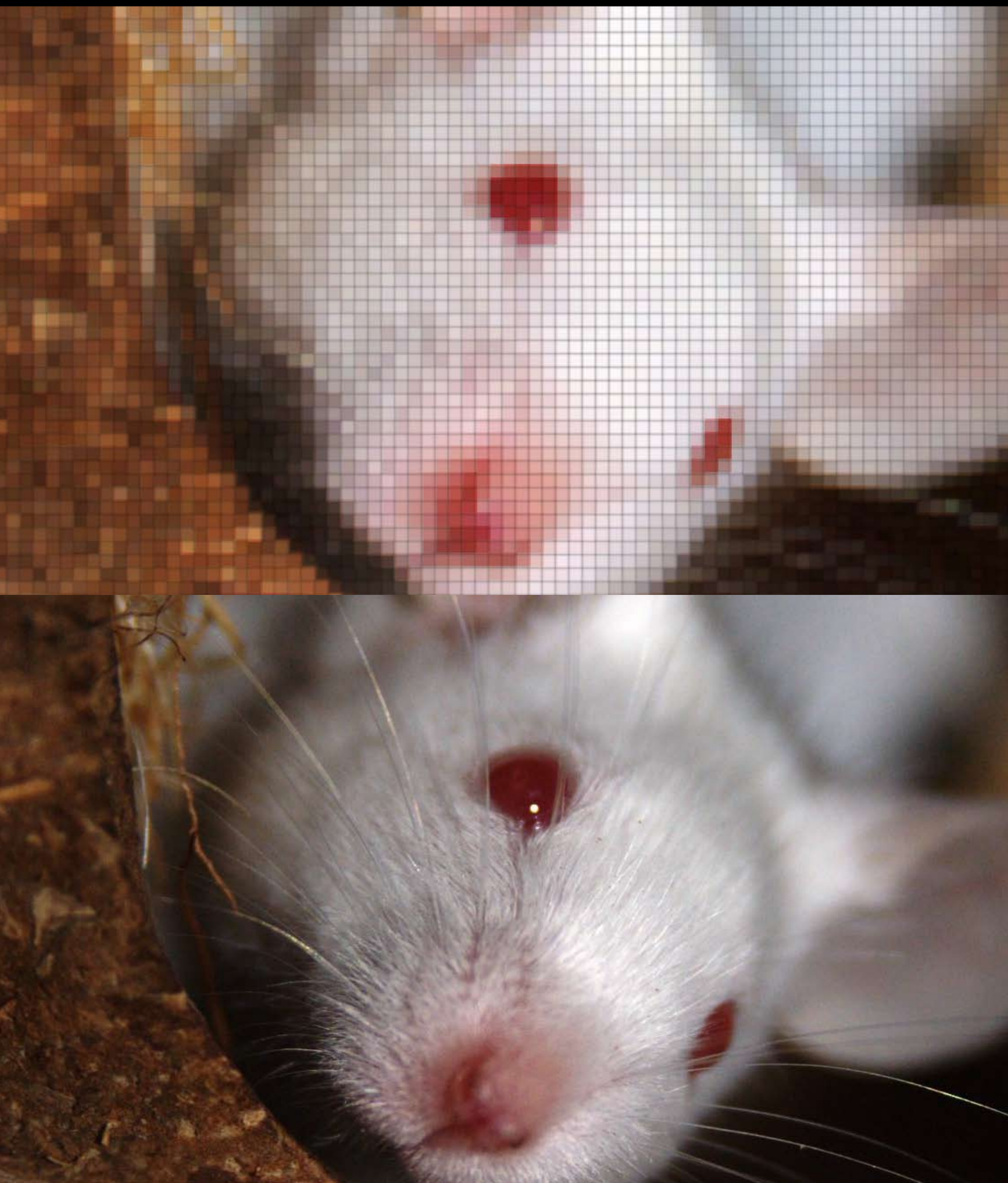
c(disney,mouse)



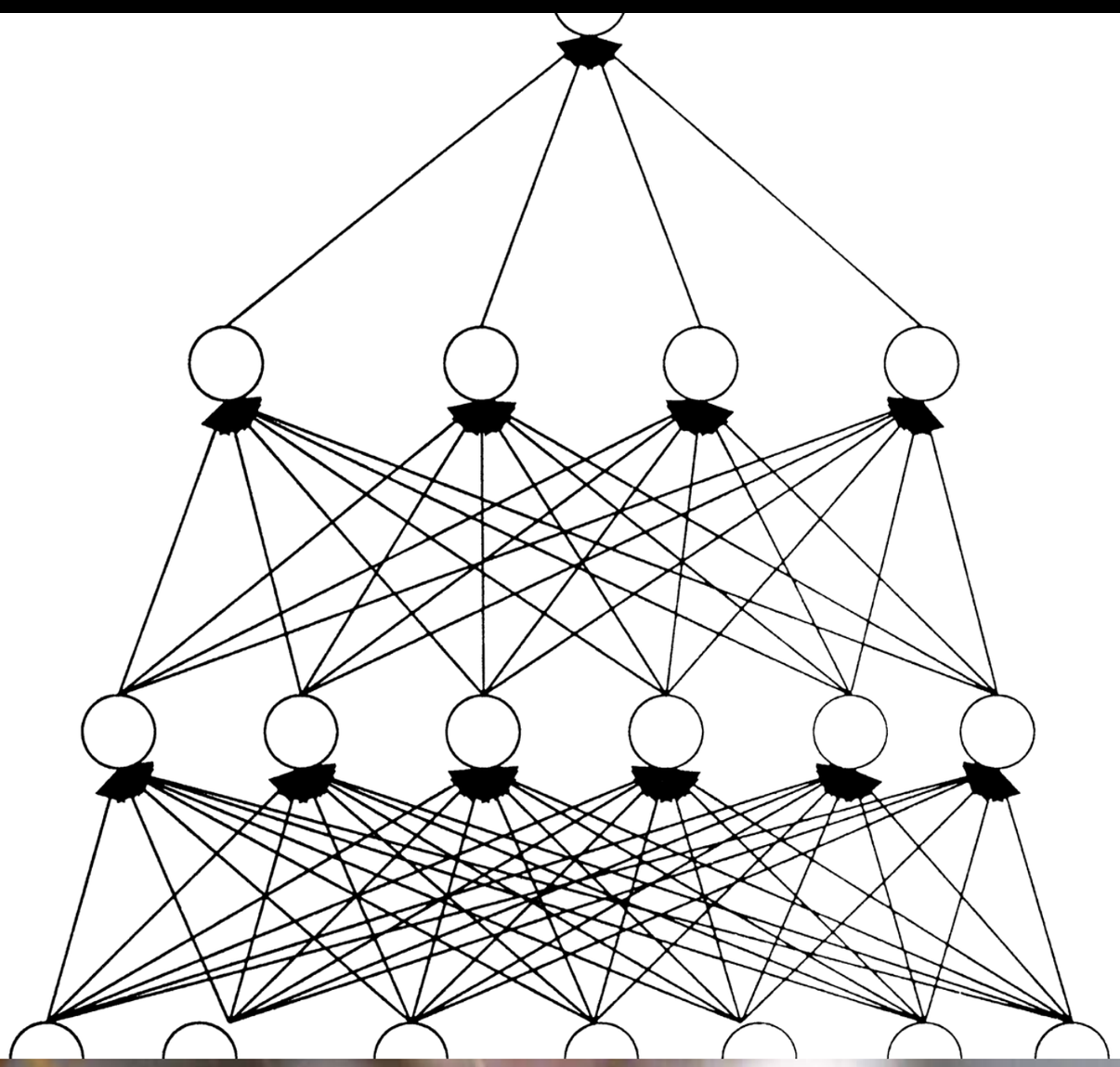
c(device,mouse)



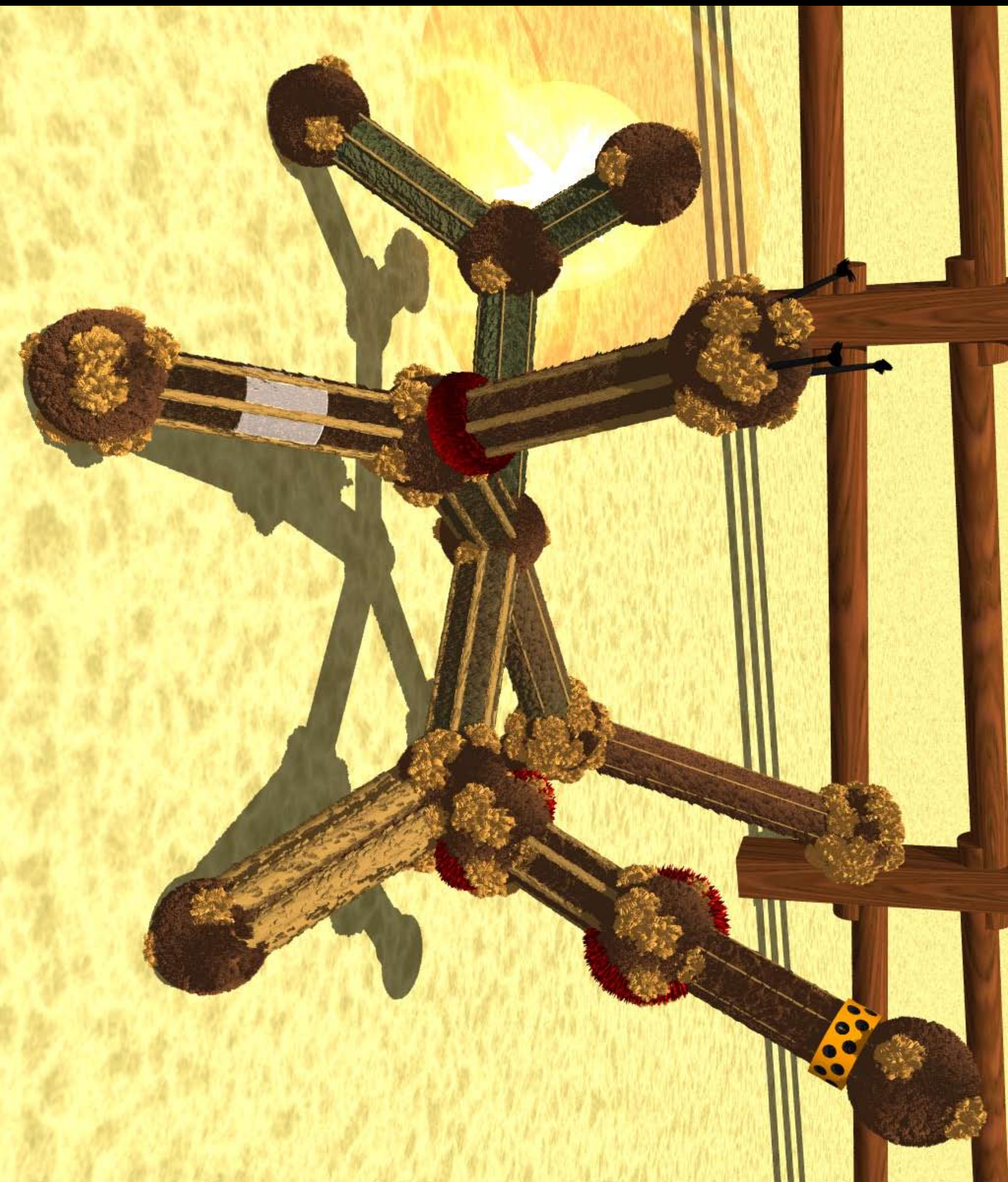
iconic and symbolic representation



distributed representation



unconstrained representation



non-formal

so, what's the point?

SGP is a major **challenge** of AI

the framework raises **new questions**

formalization opens the way for **proofs**

the framework aids **communication**

Physica D 42 (1990) 335–346
North-Holland

THE SYMBOL GROUNDING PROBLEM

Stevan HARNAD
Department of Psychology, Princeton University, Princeton, NJ 08544, USA

There has been much discussion recently about the scope and limits of purely symbolic models of the mind and about the proper role of connectionism in cognitive modeling. This paper describes the "symbol grounding problem": How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads? How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in anything but other meaningless symbols? The problem is analogous to trying to learn Chinese from a Chinese/Chinese dictionary alone. A candidate solution is sketched: Symbolic representations must be grounded bottom-up in nonsymbolic representations of two kinds: (1) *iconic representations*, which are analogs of the proximal sensory projections of distal objects and events, and (2) *categorical representations*, which are learned and innate feature detectors that pick out the invariant features of object and event categories from their sensory projections. Elementary symbols are the names of these object and event categories, assigned on the basis of their (nonsymbolic) categorical representations. Higher-order (3) *symbolic representations*, grounded in these elementary symbols, consist of symbol strings describing category membership relations (e.g. "An *X* is a *Y* that is *Z*").

Connectionism is one natural candidate for the mechanism that learns the invariant features underlying categorical representations, thereby connecting names to the proximal projections of the distal objects they stand for. In this way connectionism can be seen as a complementary component in a hybrid nonsymbolic/symbolic model of the mind, rather than a rival to purely symbolic modeling. Such a hybrid model would not have an autonomous symbolic "module," however; the symbolic functions would emerge as an intrinsically "dedicated" symbol system as a consequence of the bottom-up grounding of categories' names in their sensory representations. Symbol manipulation would be governed not just by the arbitrary shapes of the symbol tokens, but by the nonarbitrary shapes of the icons and category invariants in which they are grounded.

1. Modeling the mind

1.1. From behaviorism to cognitivism

For many years the only empirical approach in psychology was behaviorism, its only explanatory tools input/input and input/output associations (in the case of classical conditioning [42]) and the reward/punishment history that "shaped" behavior (in the case of operant conditioning [1]). In a reaction against the subjectivity of armchair introspectionism, behaviorism had declared that it was just as illicit to theorize about what went on in the *head* of the organism to generate its behavior as to theorize about what went on in its *mind*. Only *observables* were to be the subject matter of psychology; and, apparently, these were expected to explain themselves.

Psychology became more like an empirical science when, with the gradual advent of cognitivism [17, 25, 29], it became acceptable to make inferences about the *unobservable* processes underlying behavior. Unfortunately, cognitivism let mentalism in again by the back door too, for the hypothetical internal processes came embellished with subjective interpretations. In fact, semantic interpretability (meaningfulness), as we shall see, was one of the defining features of the most prominent contender vying to become the theoretical vocabulary of cognitivism, the "language of thought" [6], which became the prevailing view in cognitive theory for several decades in the form of the

$$\neg SIR \sim I \wedge ISR \sim I$$

some questions

which class is **sufficient** for AGI?

where does **your research** fit in?

which classes are pragmatically **useful**?

are any of these classes **equivalent**?