MASTER COPY

Psychological Review
1979, Vol. ——

Further Arguments Concerning Representations for Mental Imagery: A Response to Hayes-Roth and Pylyshyn

John R. Anderson
Carnegie-Mellon University

Anderson advanced reasons for believing that it was not possible to discriminate between imagery and propositional representations. I provided a formal argument that behavioral data would not provide a basis for identifying the correct representation. I argue that it seemed unlikely that other criteria, such as physiological constraints, parsimony, efficiency, or optimality, would succeed in providing unique identifiability. Hayes-Roth and Pylyshyn have challenged these conclusions, alleging flaws in the formal argument, claiming potential for the other criteria I tried to discredit, and advancing new criteria. Each of their main points is carefully inspected, and it is found that either they are incorrect in their points or the points do not argue against the original conclusions.

In Anderson (1978) I explained my reasons for not believing that it would prove possible to discriminate between imaginal and propositional representations and argued that existing arguments for the imaginal and propositional points of view failed to make their case. There were two components to that article. First, I argued that on purely behavioral grounds, it is not possible to discriminate between the two types of representation. That is, for whatever one could do with a version of one type of representation, one could find a version of the other type of representation that did the same things. I presented a formal argument that established this nonidentifiability claim in the strongest possible terms. This point about behavioral data seemed very important because almost all of the discussion in the literature was of a behavioral variety. That is, this discussion presented arguments that one representation could lead to behavior observed of humans, but the other representation could not.

However, there are a number of ways to select among theories besides behavioral considerations. In the second component of my article, I considered a number of these—physiological factors, plausibility, parsimony, efficiency, and optimality. Here I was unable to make such strong conclusions. In part, this is because these are quite subjective criteria (with the exception of the physiological) relative to behavioral adequacy, and it is difficult to formalize arguments involving them. Although there is some potential for each of these criteria, there also are good reasons for doubting that the identifiability issue will be fully resolved by any combination of these considerations, at least in the foreseeable future. Although I remain pessimistic about these nonbehavioral criteria, I do not see them as hopeless. A certain amount of my research and thought is given to exploring their potential.

I had hoped that my article would have one of two effects. Either it would shift the focus from attempting to discriminate among different types of representation to trying to develop at least one truly adequate theory (of which there are none). Or it would shift the level of discussion from considering purely the behavioral questions, such as whether a particular type of representation can predict a given set of data, to questions concerning the nonbehavioral criteria that in conjunction with behavioral considerations, still offer some hope of providing a resolution to problems of representation.

In part, Hayes-Roth (1979) and Pylyshyn (1979) are trying to emphasize and illustrate the potential of the second response in their

Preparation of this article was supported by Grant BNS78-17463 from the National Science Foundation and Contract N00014-77-C-0242 from the Office of Naval Research. I am grateful to David Kieras, Paul Kline, Clayton Lewis, and Lynne Reder for their advice and comments on earlier drafts.

Requests for reprints and inquiries concerning this article should be sent to John R. Anderson, Department of Psychology, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

Copyright 1979 by the American Psychological Association, Inc. 0033-295X/79/0000-0000$00.75

articles. Although I find some of their analyses better than many of the earlier ones in the literature, they do not convince me to be optimistic about the alternative approaches. However, in part, they are also challenging how well I established the inadequacy of behavioral data for identifying internal representations. Here I simply disagree.

Their articles, like my original, are discursive. However, when we get to my rejoinders, I suspect the readers would like to focus on the central issues. So I will just respond to the main complaints and restrain my urge to respond to many of their other side remarks and criticisms. Fortunately, Hayes-Roth and Pylyshyn have facilitated by task by clearly identifying their main points. I will first summarize my response to each and then give my full response.

Hayes-Roth

1. Hayes-Roth claims that my behavioral argument does not extend to response-time data because one representation may support more efficient processors than another. However, my behavioral argument does extend to response-time data. Hayes-Roth's point about efficiency is irrelevant to a behavioral argument. We would need, but do not have, physiological knowledge to judge claims about the efficiency of processes associated with a representation.

2. Hayes-Roth claims I made the invalid assumption that brain processes are primitive recursive and therefore have computable inverses. The finite capacity of the brain guarantees that its processes are primitive recursive.

3. Hayes-Roth claims that imagery theories have made successful predictions, whereas propositional theories have not. This claim is incorrect in principle if we speak of possible theories, and it is incorrect in historical fact if we speak of the theories that have actually been proposed.

4. Hayes-Roth claims that I made exaggerated assertions about propositional models. It appears that he has misread the original article.

5. Hayes-Roth also claims that I created inaccurate caricatures of imagery models. Again this appears to be a misreading.

Pylyshyn

1. Pylyshyn claims that I have subjugated *explanatory adequacy* in favor of behavioral mimicry. However, his use of explanatory adequacy is basically vacuous and offers no basis for choosing among representational theories.

2. Pylyshyn claims that the intended interpretation of a representation places constraints on representation-process combinations. I agree that this point may prove to be an additional constraint similar to the informal constraints of parsimony, plausibility, efficiency, and optimally. However, this criterion as Pylyshyn has developed it is still too vague to permit careful evaluation. It seems doubtful that it will provide unique identifiability.

3. Pylyshyn claims that there is a mathematical error in the formal proof involving my condition of *preservation of internal distinctions.* He also claims that this condition would not be met in most interesting cases. In fact, this purported error is a misreading: It is also easy to show that the condition is met in the cases that have been of concern in the imagery–propositional debate.

4. Pylyshyn questions whether the mimicking model could be made to perform its computations as fast as the mimicked model. Under the constraints of current additive factors logic, the mimicking model could be so specified. However, Pylyshyn wants to constrain theories to involve steps that only take constant time, in contrast to additive factors models that can have stages with varying time. Pylyshyn's restriction is unacceptable because it would exclude many plausible psychological theories.

Hayes-Roth

Response Time

Hayes-Roth's concern under his Point 1 is that my theorem will not extend to response times: "The theorem entirely neglects other process-determined measures, such as the time (latency) required to compute a response to a particular stimulus" (1979, p. xxx). However, as he acknowledges in his Footnote 2, I do deal with response times, and the original article extended the formal argument to response times. However, he argues that "his approach assumes both (a) that two alternative mechanisms would perform the same computational steps and (b) that the two mechanisms could perform comparable steps in the same time .... These strong, questionable assumptions require more support than Anderson has provided"

(1979, p. xxx).[1] (For continued reference in this section, I refer to these as assumptions a and b.) I showed behavioral mimicry in my formal argument by in fact showing a. That is, the formal argument was that the mimicking theory produced the same behavior as the mimicked theory by going through the same steps (see Figure 3, Anderson, 1978). Therefore, if Hayes-Roth concedes I prove my theorem correctly, as he does, a is established. The points that he makes in detail seem aimed at b. That is, he questions whether a mimicking mechanism can always be made to perform corresponding steps in the same time as the mimicked mechanism. I claimed that we could adjust the basic speeds of the mimicking mechanisms to put the comparable steps into temporal correspondence.

So, the central question here is whether one is free to postulate that the speed of the steps in mimicking process match the speed of the corresponding steps in the mimicked process. The answer to this question is that without detailed knowledge about how the steps are physically implemented, there is no basis for objecting to assigning the needed temporal properties to the mimicking steps. Or put another way, the basis for Hayes-Roth's objecting to point b should be physiological; it cannot be in a behavioral or logical analysis. However, Hayes-Roth does not try to cast doubt on my assumption b by appeal to physiological facts. Indeed, I think it is clear that our current state of knowledge would not support such an appeal.[2]

When Hayes-Roth's remarks are examined in detail, I think it is clear that they are weak and need a physiological foundation. To illustrate this, I will consider one of the apparently stronger aspects of his arguments. This has to do with the problem of getting mimicry between graph-matching and template-matching processes. Hayes-Roth contrasts template models (imagery) that have *instantaneous recognition times* with general graph-matching algorithms (propositional) that are known to be nondeterministic polynomial-complete (NP-complete). As Hayes-Roth points out, it is generally suspected that NP-complete problems will have exponential complexity functions on many devices, including Turing machines and serial computers. An exponential complexity function means that computation time increases exponentially with problem complexity. (See Aho, Hopcroft, & Ullman, 1974, for a general discussion of the area of complexity.)

This example does not work very well for three reasons. First, as Hayes-Roth admits, there is no reason to limit the machinery for implementing propositional theories to those that result in exponential functions in pattern matching. If, rather than assuming our computing device is a serial computer or a Turing machine, we assume it is a Boolean network (see Meyer & Shamos, 1977), we would get linear time complexity functions. It only seems fair to allow propositional representations access to such machinery, since Hayes-Roth is allowing imagery representations access to such machinery when he assumes template matching. Second, it seems unlikely that templates, as we now know how to implement them, would yield constant complexity functions, let alone instantaneous functions. As more points have to be combined from a template, it seems almost certain that the physical architecture to realize this will have to grow at least linearly, yielding at least a logarithmic time function. Certainly this would be true if the templates were computed by Boolean nets, which are the obvious way to realize them.

The third and most important point is that the comparison between template matching and graph matching is unfair in that the graph matching is a more general computation. Graph-matching algorithms are intended to

---

[1] It should be clear that when I spoke of computational steps in my original articles, I was intending that rather large operations would count as steps—for instance, recognizing the identify of a stimulus might be a single step. My steps were intended to be whatever the basic units of analysis were in current propositional or imagery theories. Implemented in a physical system (e.g., a computer), such a step might be analyzable into many separate substeps. I think it is this substep analysis that Hayes-Roth means to comment on here, but my arguments apply at the higher level.

[2] The language I use here may lead to the erroneous impression that there is an extra burden of proof on the mimicking theory. The terms *minicked* and *mimicking* give the impression that the assumptions of the mimicked can be taken for granted, whereas the assumptions of the mimicking bear special scrutiny. Thus, one is naturally led to question whether it is physically possible for the mimicking theory to compute its steps in the proposed time. It is easy to overlook the fact that it is equally questionable whether it is physically possible for the mimicked theory to compute its steps in the proposed time. Thus, both theories are under equal clouds of uncertainty about their temporal assumptions, and as I argue, we are simply unable to address that uncertainty given current physiological knowledge.

apply to a far greater variety of matching problems than just those involving the recognition of points in a physical plane. It is easy to design a special-purpose graph-matching algorithm for recognizing propositional encodings of dot patterns whose computational time on serial computers will be an $n\ ln(n)$ function of problem complexity; that is, graph matching for picture encodings need not show exponential time complexity. Moreover, given the parallel machinery assumed in the template case it should only take $ln(n)$ time, which is the best templates can do.

So when we make a detailed analysis of Hayes-Roth's example, we see two things. First, it supports no general claims about whether a propositional model can mimic an imagery model. Second, every step in the argument is critically dependent on the physical machinery on which the processes are implemented. We are not in possession of such knowledge in cognitive psychology. Therefore, we cannot base our choice among representations on claims that one representation supports more efficient processes than others. More generally, this means that the constraints do not currently exist to prevent one from proposing that the steps of one process take the same time as the corresponding steps of another process. This is not to say that we will never be in possession of such constraints. As I admitted in my original article, physiological considerations hold a promise for the future, albeit a dim promise.

Inversion of Encoding Functions

To prove that one representation could be given process assumptions that enabled it to mimic another representation-process pair, I tried to establish the assertion that there existed a mapping from the mimicked to the mimicking representation that had a computable inverse.[3] A sufficient but not necessary condition for establishing the computability was that the encoding function, E, for the mimicked representation, the encoding function, E*, for the mimicking representation, and their inverses E^{-1} and E^{*-1} all be computable. There is no question about whether the encoding functions are computable, but Hayes-Roth questions how generally it would be the case that their inverses are computable.

I made the point that if the encoding functions E and E* are primitive recursive, then their inverses can be computed. Hayes-Roth argues that many cognitive processes are not primitive recursive. It seems to me that Hayes-Roth is wrong in this argument. The human brain is clearly a finite state device. This can be established, if in doubt, by considering the fact that it only contains a finite number of elemental units (such as atoms) in a finite number of discriminable configurations. Finite state machines can only compute finite state functions, which are a subset of primitive recursive functions and for which rather simple inversion routines can exist (Hennie, 1968). We often speak in theorizing as if humans computed functions that were not finite state, but this is only a convenient fiction. To compute the full range of nonfinite functions, the human brain would need an infinite memory, which it clearly does not have.[4]

Post Hoc Prediction

Hayes-Roth claims that there were results that were predicted by imagery theories, but that propositional theories had to be modified after the results to accommodate the data. Hayes-Roth argues that more credit should be given to a model that *predicts* a result than to one that *postdicts* the result. Hayes-Roth here seems to fall victim to the fallacy of thinking that there is a single propositional theory and a single imagery theory, that these theories make different predictions, and that a theory needs to be adjusted if it makes the wrong prediction. In fact, there are an infinite number of theories using either representation, and, as

---

[3] The reason for wanting to establish computability has sometimes been misunderstood. It is not that I wanted to show that the brain could compute these processes; rather, I wanted to show that the processes were capable of specification in a theory. It would not be enough to merely show that there is an appropriate mapping from g to g^* without considering whether it could be specified. There is an uncountably infinite number of mappings but only a countably infinite number of ways to specify a mapping in a finite theory. It would be little comfort to know an appropriate mapping existed if that mapping could not be specified in a finite theory. Thus, the purpose of computability was to guarantee rigorous specification in the theory, not to say something about what could be done physically. This is important. To establish behavioral mimicry, it should be sufficient to establish that the objects of the theory can be specified. There is no way to address the issue of what the neural mechanisms can in fact compute without understanding how they are put together.

[4] I am grateful to David Kieras for earlier conversations that helped clarify my thinking here.

I argued in my original article, for every theory using one representation there is another theory using the other representation that makes the same predictions in all situations. So, if it is true that an imagery theory predicted a result, then if follows that there is some propositional theory that predicts the same results. Conversely, if there is a propositional theory that mispredicts a result, it follows that there is an imagery theory that does the same.

Still if one accepts what Hayes-Roth asserts, it might seem that imagery theories have a certain "magic" associated with them. After all, he implies it was the imagery theories that people were working with that predicted the correct results and not some hypothetical theory as was the case with propositional theories. However, if we examine the true history of imagery theory, that magic rapidly dissipates. For instance, there was a time (c. 1970) when most working imagery theories would not have predicted effects of image complexity but propositional models would have. There turned out to be such complexity effects (Kosslyn, 1975). Now we have an imagery model that predicts such effects (Kosslyn & Schwartz, 1977). Other imagery conceptions (e.g., Cooper, 1975) would still seem not to predict this result. So exactly in contradiction to Hayes-Roth's claims, there is no connection in practice between imagery theories and correct predictions about effects of complexity.

Inaccurate Claims for Propositional Models

Hayes-Roth feels that I have made some inaccurate or misleading references to propositional models. I was able to identify two purported inaccuracies and will consider each:

1. One problem is that he thinks I treat propositional representations as theories. I quote below a number of statements of Hayes-Roth that seem to be variants on this theme:

> Anderson treats the propositional representation as though it were explanatory. (1979, p. xxx)

> In sum, propositional descriptions per se do not constitute models; they simply reexpress observations in a formal syntax. (p. xxx)

> Without some plausible mechanism for producin and modifying the propositions, the predicate calculu; is not a model, but simply a descriptive formalism. (p. xxx)

I am in perfect agreement with Hayes-Roth with respect to this claim about propositional representations. However, this is just the claim that I was making by insisting that one can not talk about representations in isolation, but needs to talk about representation-process pairs. There are numerous references to this point in the original article that I could quote but the following seems most apt:

> A representation without any process assumptions is not a theory. By making different process assumptions, it is possible to have quite different theories with the same propositional representation. We cannot test representations but only representation-process pairs. (Anderson, 1978, p. 257)

So with respect to Hayes-Roth's abovementioned fallacy, I do agree with Hayes-Roth that it is a fallacy. However, I do not think I am guilty of it. Moreover, the position he asserts, far from invalidating the thesis of the original article, is essential to it.

2. The other problem is that I claim that propositional representation makes certain kinds of computations easy. Although Hayes-Roth's discussion makes it seem that I made this claim generally, the fact is that this claim was only directed to inference making. Hayes-Roth points out that actual experience with theorem proving in artificial intelligence indicates that theorem-proving systems run up against real computational problems. However, I did not intend inference making with propositional systems to be equated with theorem proving. There are clear examples (e.g., Rieger, 1975) of inference systems that operate on propositional data bases but which use techniques that are clearly not theorem-proving techniques. Hence, computational problems met in theorem proving are not necessarily relevant.

A second point is that it is irrelevant to the central thesis of the original article whether propositional representations do have any advantages for inference making. If we deny a special connection between propositional representations and inference making, we make it all the more improbable that there will be a way to discriminate between propositional and imaginal representations.

Fully Specified Models and the Picture Metaphor

Like his previously mentioned point, Hayes-Roth's point here seems to amount to two distinct subfallacies.

1. The first involves my claim that "one can test and evaluate only models with explicit process assumptions" (Hayes-Roth, 1979, p. xxx). This point seems so obvious to me that

I find it hard to understand how it can be questioned. A mental representation just "sits" inside the head. Some process has to interact with it before it can have behavioral consequences. As a counterexample Hayes-Roth presents Bohr's model of the atom, which assumed that the atom was largely empty space. This model of representation, claims Hayes-Roth, predicted that concrete objects would be transparent to small particles. I find the analogy dubious, but for argument's sake, I will pursue it. It seems to me that there was a *process assumption* in this model, even if it might have been implicit. This process assumption was that particles will move through space if there is no body to occlude their trajectory. This same remark applies to the other examples cited by Hayes-Roth—that is, the process assumptions are there, if implicit.

So the point still holds that a representation must have process assumptions to make predictions. It is true that sometimes the process assumptions are implicit, just as in other cases the representational assumptions are implicit. It is also a fact that science can progress some distance without explicitly stating assumptions. However, I hardly think it follows that it is desirable to have our assumptions implicit or that we should not strive to make our models as explicit as possible.

The above having been said, it is true that I do use the lack of a well-specified model to criticize much of current imagery theory, perhaps leaving the implication (as stated by Hayes-Roth, 1979) "that scientific models without explicit process assumptions have no value" (p. xxx). Such an implication would be incorrect and unfair to imagery theorists. Although I did not intend such a strong negative evaluation, I apologize for not taking more care to disavow it, which I do now. However, it should be noted that this unintended insult to imagery theorists (and indeed any other part of this subfallacy) is totally irrelevant to the identifiability issue.

2. Hayes-Roth's second subfallacy is that I only consider the *picture-metaphor* theory and overlook important alternatives in the set of analogue theories. I do not understand this criticism. My use of the picture metaphor was to defind imagery theory against the attacks from the propositional camp. If I could show that criticisms of imagery theory did not apply to this most extreme subclass of the imagery theories, then it would follow that they would not apply to the full class of imagery theories. Apparently Hayes-Roth thought I used this specific theory to criticize imagery theories. For example, consider this characerization of my argument:

> Moreover, by demonstrating the superiority of the propositional model to the only operationalized analog models (i.e., the picture-metaphor), we establish the propositional model's general superiority over that class of models. (1979, p. xxx)

This is almost the opposite of what I was trying to do with the picture theory.

In conclusion, Hayes-Roth's last point is a simple misinterpretation of my article.

Pylyshyn

Explanatory Adequacy

Pylyshyn's first point is that my arguments for indeterminacy are flawed because I have not considered explanatory adequacy. It seems to me that Pylyshyn has chosen to begin with his weakest point, although it is clear he thinks that it is his strongest. This is certainly symptomatic of the wide divergence of opinion about the force of explanatory adequacy. Chomsky (1964), who introduced the term, was worried about a similar indeterminacy problem in linguistics, in which different grammars could be formally equivalent. In his use a linguistic theory possessed explanatory adequacy if it prescribed a basis for selecting among those formally equivalent grammars.

Unfortunately, Pylyshyn is not clear about what he means by explanatory adequacy. The worst possible interpretation of Pylyshyn, which I suspect is accurate, is that he means to equate explanatory adequacy with "factors that serve to separate behaviorally equivalent models." Then, his first point becomes "Anderson has failed in his argument that one cannot distinguish between propositional and imagery models because he has failed to consider factors that serve to separate behaviorally equivalent models." If this is a correct interpretation of Pylyshyn, then it is clear why he is wrong in using this as a criticism of my article: I do consider a number of nonbehavioral criteria (e.g., physiological, parsimony, and efficiency) but conclude they do not work. If these are the criteria that he means to reference by explanatory adequacy, he should address my reasons for pessimism. If these are not the criteria he had in mind, then he should be more explicit about what other criteria he has in mind.

It is instructive to contrast Pylyshyn's use of explanatory adequacy with Chomsky's use.

Chomsky (1965) did attempt to provide some meat to this notion of explanatory adequacy by relating it to issues of language acquisition. An explanatorily adequate theory would specify the basis a child uses for selecting among equivalent grammars. Pylyshyn does not provide such a substantial interpretation for his use of explanatory adequacy in this context.

However, Pylyshyn does make some effort to indicate what he means in the discussion of his point. One suggestion is that an explanatorily adequate model will appeal to principles from outside the domain the model addresses:

> One of the things that makes a certain use of a model explanatory is that it appeals to general principles and mechanisms—ones that play a role outside of the narrow range of phenomena that that particular model addresses. (1979, p. xxx)

However, Pylyshyn seems to forget that the choice under consideration is the representational system (propositional vs. dual code) that will serve as a basis for all of human cognition. It is hard to imagine what more general constraints he might have in mind. Moreover, I see no reason why these more general constraints would solve the indeterminacy problems. It would seem that the same arguments could extend, mutatis mutandis, to the more general system in which the representation only played a part. To the extent Pylyshyn is arguing that physiological factors can serve as a constraint, I agree. However, I see no reason to retract the pessimism in my original article about whether these factors will in fact serve such a role.

At another point Pylyshyn says:

> The more explanatory theory is the one that captures the important generalizations and principles and that factors apart and independently justifies such things as the computational requisites of the task, the specific algorithms the organism uses in certain situations, and the presumably universal, biologically based, fixed mechanisms available to the organism for executing these algorithms. (1979, p. xxx)

It is unclear exactly what is being asserted here. The use of "important generalizations" suggests a standard appeal to parsimony. The reference to "the presumably universal, biologically based, fixed mechanisms available to the organism" suggests physiological considerations. The reference to "computational requisites" suggests efficiency. Thus, Pylyshyn is in part just using the term *explanatory* to refer to the nonbehavioral factors I acknowledge in my original discussion. He does not address my reasons for doubting their potential. The use of the phrase "independently justifies" suggests that he intends more than just this, but he no more makes explicit what he means by this phrase than he makes explicit what he means by explanatory.

Intended Interpretation of Representational Forms

Pylyshyn's second point is that representations have *intended interpretations* and that a constraint on the processes that operate on them is that they use the representations in the intended ways. Pylyshyn is onto something important here. However, this concept needs considerably more work, and even then I do not think it will resolve the identifiability issue.

Pylyshyn's remarks here are not relevant to my behavioral argument. Actually Pylyshyn is proposing a new criterion for theory evaluation of the status of parsimony, plausibility, efficiency, and optimality. As is the case with these considerations, this point could be developed to provide a basis for preferring some representations over other behaviorally equivalent ones. However, I do not think it nor any of the other criteria will give us unique identifiability. It is not possible to support this negative opinion with tight arguments because these criteria lack the rigor required to make formal analyses. All one can do is provide plausibility arguments for pessimistic or optimistic conclusions. I did this for my criteria in the original article.

Although Pylyshyn's example of the fork, bottle, and rose is both clear and compelling, there are many cases in which it simply is not clear whether a use of a representation violates its intended interpretation. For instance, philosophy and psychology have had a long history of debate about whether one can use a specific image, say of a triangle, to represent the general concept of a triangle. To take another example, is it permissible to use a picture of a church to represent the abstract concept of religion? As an example from the other side of the fence, is it permissible to use the number 7 in a representation to encode a shade of grey? In each of these cases there would not be consensus about whether the representation was being used as intended. If there is no consensus, then the notion of intended representation would fail in deciding representational issues.

Preservation of Semantic Distinctions

Pylyshyn's third point concerns my use of the condition called preservation of internal distinctions, which he chooses to relabel *preservation of semantic distinctions*. I made the formal argument that if this condition were satisfied between two representations, it would be possible to have one representation mimic the other. Pylyshyn takes two issues with my use of this condition. First, he claims that there is a mathematical error in my formal argument. Second, he claims that the condition of preservation of semantic distinctions would not be met between propositional and imaginal models. That is, he questions whether the formal argument applies to the case at hand. I will address his arguments in this order.

With respect to the purported error in the formal argument, I can only conclude that Pylyshyn has misread me. I will try to clarify the matter here. The formal argument was that if one representational theory $\langle \mathcal{I}^*, E^* \rangle$ preserves the internal distinctions of another theory $\langle \mathcal{I}, E \rangle$, then it is possible to construct some model M* involving $\langle \mathcal{I}^*, E^* \rangle$ that mimicks the behavior of any model M involving $\langle \mathcal{I}, E \rangle$.[5] (As pointed out by D. Kieras (Note 1), this is a sufficient, but not a necessary, condition for mimicry). The condition of preservation of internal distinctions involved three subconditions or assumptions:

1. There is a one-to-one mapping[6] f from $\mathcal{I}$ to $\mathcal{I}^*$,
2. f has a computable inverse, and
3. $E(S) = f^{-1}[E(S)]$.

For purposes of Pylyshyn's discussion, Subconditions 2 and 3 are the critical ones. After introducing the condition in the original article, I gave a discussion of the circumstances under which Subcondition 2 might hold:

> One might wonder under what conditions the mapping would have a computable inverse. If E and E* are primitive recursive (Minsky, 1967), a mapping f can be constructed with a computable inverse f^{-1}. One can simply make $f = E^* \cdot E^{-1}$, where $E^* \cdot E^{-1}$ denotes the combination of applying first the inverse of E and then E*. Similarly, $f^{-1} = E \cdot E^{*-1}$. If E and E* are primitive recursive, E^{-1} and E^{*-1} will be computable; and hence, f and f^{-1} will be computable. (Anderson, 1976, p. 25)

Note carefully that this discussion does not assert that this construction serves to guarantee Subcondition 3. Rather it was concerned with when 2 would be satisfied. (I have already explained in Footnote 3, with respect to Hayes-Roth's second point, why it was important to guarantee Subcondition 2. My interpretation of Pylyshyn indicates that he misread my article as asserting that this construction guaranteed 3. Perhaps I should explain why I brought up a special construction just to guarantee 2: It seemed to me that 2 was a more difficult condition to argue about than 3. I felt I could establish that 3 held simply through direct analysis of the properties of propositional and imaginal representations, but that I needed some other way of assuring 2. So, I turned to this construction. The argument was that if 2 were satisfied, this construction would assure that 3 was satisfied.

To show that Pylyshyn's purported mathematical error rests on this misreading, I will discuss in detail some of his points. Let us consider Pylyshyn's counterexample to the formal argument that is represented in Figure 1. There are three stimuli that are mapped into propositional and imagery representations. Note that the imagery representation collapses the distinction between S_1 and S_2 by assigning both to I_{12} but distinguishes both from S_3 by assigning it to I_3. In contrast, the propositional representation collapses the distinction between S_2 and S_3 by assigning them to P_{23} and distinguishes both from S_1 by assigning it

---

[5] To briefly review: $\mathcal{I}$ and $\mathcal{I}^*$ are the representationa systems in the mimicked and mimicking theory while E and E* are the encoding functions that map stimuli S into encodings I.

[6] My use of the term *mapping* to refer to f was confusing in two ways. The natural interpretation is that it is a function such that each element I in the domain $\mathcal{I}$ is assigned a unique element I* in the range $\mathcal{I}^*$. However, this is stronger than was intended or needed. Given that more than one element I* in $\mathcal{I}^*$ can correspond to an element I in $\mathcal{I}$, the intended interpretation is a more general relational system that can assign more than one element from the range to an element in the domain. What was actually required in the definition was that the *inverse* of f be a function, which was guaranteed by the constraint that f be one to one.

A second way f is confusing has to do with whether it is intended to be something computed by the mimicking model. The function T* was given as $f \cdot T \cdot F^{-1}$, and D* was given as $D \cdot f^{-1}$. Although in the original article there are discussions that consider the possibility of the mimicking model computing T* and D* through f, these discussions are tangential to the basic argument about behavioral mimicry. For the purposes of that argument we only have to think of f as being involved in the theoretical specification of D* and T*. Even here, f is used in existence proof—to show that specifications of D* and T* exist. As remarked in the original article, it is to be expected but not guaranteed that simpler specifications exist.

to P_1. (Actually Pylyshyn does not mention P_1, but I put it in for the sake of symmetry.) Pylyshyn argues that my above-mentioned construction when applied to this situation will create a mapping from the imagery representation to the propositional representation that is one-to-one but not distinction preserving. It definitely is not distinction preserving, as Pylyshyn notes. However, the formal argument was never intended to guarantee mimicry in cases like this. The condition in my formal argument was just meant to exclude examples like this. Hence, his counterexample is nothing more than an illustration of the importance of the conditions of my original argument.

Pylyshyn goes on to provide a definition of distinction preservation in terms of equivalence classes that will guarantee two aspects of the preservation of internal distinction—namely, 1 and 3. In fact, Pylyshyn's condition is equivalent to my 1 and 3. What Pylyshyn seems not to realize is that I used this equivalence-class formulation in an earlier rendition of the nonidentifiability argument (Anderson, 1976; pp. 10–12). Given that the two formulations are equivalent, in a certain sense it is arbitrary which one is used. However, I chose to use the mapping terminology rather than the equivalence class terminology because I judged that it facilitated certain aspects of the argument. It appears that the choice was not without an accompanying loss to the clarity of the discussion.

To summarize my points: There is no error; there is a misreading; in a certain sense Pylyshyn has only restated what I said; I have earlier used the same statement of the result that he offers.

Pylyshyn notes that one representational scheme (the mimicker) must impose as fine or finer a grain on the stimulus world than the other (the mimicked) and must not cross the boundaries of the other. This is exactly the image I was holding in mind as I wrote the article, but Pylyshyn writes:

> No one would bother to postulate two separate representational schemes and characterize them as being different schemes, if one were simply a refinement of the other, for in that case one set of codes could be viewed simply as a subset of the other. For coding schemes to be different in any interesting non-notational sense, they must cross-classify some of their semantic domain. (1979, p. xxx)

On the contrary, there are interesting contrasts between propositional and dual code representations in which one is a refinement (partition) of the other in Pylyshyn's sense. The contrast is probably interesting because that distinction has some physiological implications if not behavioral.[7]

This brings us to the second half of Pylyshyn's argument under his third point. He questions whether it is really possible to establish image representations as partitions of a propositional representation and vice versa. There are actually two questions here, depending on what we take as our imagery model. Can we find pure pictorial representations and propositional representations that will mimic (be partitions of) each other? Or can we find picture-plus-word (dual code representations; Paivio, 1971) representations and propositional representations that will mimic each other? It is not clear that anyone has espoused a pure pictorial theory.

Pylyshyn argues that there are certain kinds of distinctions in pictorial representations that propositional representations do not distinguish:

> For example, a propositional representation of two objects could assert that they were next to each other with no commitment as to whether one was to the left, right, above, or below the other, whereas it would violate the spirit of the imaginal view . . . to allow that form of indeterminacy to be depicted in a pictorial image. (1979, p. xxx)

Let us consider Pylyshyn's quoted point and see if it really creates any difficulties for the possibility of mutual mimicry. It is important to be clear as to why the *next-to* relation might be a problem. I suspect that Pylyshyn thinks it is a problem for imagery theory. The claim would be that humans show left–right confusions; a propositional theory would, but no image theory could. However, it would be

Figure 1. Representation of Pylyshyn's counterexample that demonstrates the purported error.

[7] It should be noted here that in other contexts, representational differences are considered interesting even if they are distinction preserving. For instance, a list for a sorting program can be represented as an array or as a linked list. These two representations clearly preserve each others' distinctions, but it can be an important issue which representation is used.

false to claim that no imagery theory could predict confusion. Since the image representation provides a finer partition than the propositional theory, the formal argument guarantees that it is possible to combine some process with the image representation to get a theory that mimics the propositional theory. Therefore, if the propositional theory predicts confusion, then some imagery theory does. Such an imagery theory would just refuse to report the left–right relation encoded. It can be reasonable to refrain from reporting left–right relations in some image systems. For instance, images might be vulnerable to left–right inversions. If so, the left–right relation in a memory image would not be a reliable indicator of what the original stimulus was. In conclusion, it is both formally possible and potentially plausible that an image theory representing right–left could mimic a propositional theory using next-to.

On the other hand, Pylyshyn might think that this example poses a problem for the propositional representation's ability to mimic the image representation. One might argue that it is possible for the picture representation to discriminate in its behavior the left–right relation, whereas the propositional representation could not. However, we can use another propositional representation, one that uses relations like *above*, *below*, *right*, *left* that would preserve the pictorial representation. The identifiability problem is hard because we have to distinguish between a class of propositional representations and a class of imagery representations. It is not enough to show that one propositional representation cannot mimic one image representation.

Pylyshyn also tries to argue that propositional representations preserve distinctions not preserved in any picture or dual-code representations. In the case of dual-code representations with their verbal traces, it seems clear that some dual-code theory can preserve the internal distinctions of any propositional theory. This derives from the close relation between sentences and propositions. He makes his strong argument with respect to a pure picture theory:

> Conversely, the propositional representation would invariably make a commitment with respect to some particular interpretation of a perceived three-dimensional figure whose two-dimensional projection was ambiguous (e.g., a wire outline of a cube), whereas the pictorial representation would not distinguish between the two possible interpretations (i.e., it would encompass both). (1979; p. xxx)

This example is particularly noncompelling, since certainly nobody intended pictorial representation in image theory to be restricted to two dimensions. Indeed, Metzler and Shepard (1974) try to demonstrate three-dimensional representations. In general I think there is enough room in the world of pictures to represent directly any interpretation Pylyshyn might think of. It is also the case that a propositional representation could code a two-dimensional dot pattern without a commitment to whether the encoding should be interpreted in three dimensions. Thus, it is also possible for propositional representation to blur the distinction.

In conclusion, it seems clear all distinctions among well-specified pictures can be preserved in some propositional representations. Also because of the verbal code, it is clear that all propositional distinctions can be preserved in some dual-code theory. Therefore, the assumptions of my formal argument are satisfied. Pylyshyn is wrong both in his assertion about an internal flaw in that argument and about whether the argument applies to the case at hand.

Strong Equivalence of Process

Pylyshyn questions whether my argument about behavioral mimicry extends to issues of response time. In a way this is similar to Hayes-Roth's first point. However, Pylyshyn has different reasons for questioning whether the argument extends to reaction times. Therefore, it requires an independent reply. I stated in the original article that my argument extended to response times because the mimicking model would go through the same steps of information processing as did the mimicked model (Figure 3, Anderson, 1978). Since the two models go through the same steps, we can guarantee that they take the same time by having the mimicking model perform its steps at the same pace as the mimicked model. Pylyshyn objects to this, arguing that it is necessary to analyze the computation underlying each large step into its primitive steps to assure that the mechanism can compute each step according to its specified time parameter.

In fact, what Pylyshyn insists on is not usually done in cognitive psychology. The typical additive factors logic (Sternberg, 1969) analyzes information processing into a sequence of stages, specifies which factors affect the time for each stage, specifies the time

parameters for each stage, but does not analyze in computational detail why each stage takes as long as it does or why it is affected by factors in the way that it is. Reaction-time data have been used in developing propositional or imagery theories according to this additive factors logic. But Pylyshyn is calling for a more demanding discipline on how behavioral data are to be interpreted. He wants theories to be specified down to the level of *primitive steps*. In his view, the defining feature of a primitive step is that it takes a constant amount of times. Although it is reasonable to aspire to as detailed an analysis of information processing as possible, it seems unreasonable to require constant time steps. It is perfectly reasonable, for instance, to propose a theory in which the time for a stage or a step to be completed depends on the energy available to the stage. For example, in this way Townsend (1974) suggests that stages can speed up in a serial model, but Pylyshyn feels that such assumptions reflect a misunderstanding of "the special status accorded to reaction-time measures in information-processing psychology" (1979, p. xxx). Rather than a misunderstanding, it is standard practice (e.g., Atkinson, Holmgren, & Juola, 1969). Indeed, there is ample evidence that basic neuronal units can vary in their rate of responding (e.g., Thorson & Biederman-Thorson, 1974). Therefore, Pylyshyn's proposed regimen for reaction-time analysis is not acceptable.

Pylyshyn's point about the difficulty in getting the mimicking process to be as fast as the mimicked may have some force through considerations of complexity of computation, even if we do not accept his regimen. I too have been associated (Anderson & Kline, Note 2) with proposals to use complexity analyses as constraints on psychological models. The problem with such complexity analyses is that they require knowledge about the hardware on which the computation is occurring. (This was the problem with Hayes-Roth's proposal, too.) Still, let us consider how Pylyshyn plans to sue complexity analyses to resolve the identifiability problem.

Pylyshyn's particular objection to my claim about mimicry of time relations is that the mimicking model may require processes that are combinatorial in a way that the mimicked processes are not. For instance, the mimicked process may take a constant time, whereas the mimicking process's time increases linearly with the complexity of the structure it operates on. Or to take another example, the mimicked process may be linear in complexity of structure, whereas the mimicking process may be exponential. As Pylyshyn points out, if the mimicking process displays a complexity function greater than the mimicked, then no finite amount of speedup for the mimicking process will allow it to compute as fast on all possible problems as the mimicked process. No matter how much we speed up the minicking process, there will always be problems of enough complexity that the mimicking process will take longer than the mimicked. Thus, if the mimicking process has a greater complexity function than the mimicked and if problem complexity can be increased without bound, it will not be possible to always achieve mimicry even with a speed-up proposal.

The mistake in this argument is the assumption that problem complexity can be arbitrarily increased. There are clear bounds on the complexity of the problem that can be presented to the human system. For instance, if we are concerned with visual pattern recognition, there are limits on pattern complexity posed by such factors as size of short-term memory and discriminative capacity of the visual system. As long as there are complexity bounds on the possible problems, the mimicking process can be sped up (with a finite boundary on the amount of speedup) to be able to mimic the other processes on all problems that are actually possible. Pylyshyn objects to such speed-up proposals as being misunderstandings, but as noted above, this objection is incorrect. Therefore, complexity arguments of this variety are simply inappropriate to choosing among psychological theories.

Conclusions

In this article, I have concentrated on defending the earlier article and not on advancing new lines of arguments. This is what seemed appropriate in a reply to criticisms. That original article made a number of important points, and I wanted to make sure they were properly understood. I hope this exchange has accomplished this function of clarification.

Although the Hayes-Roth and Pylyshyn articles provide no solid bases for identifiability, they do reinforce the notion that there might be nonbehavioral bases for establishing unique identifiability. In my mind there are three open questions. It seems that if any of these questions can be answered in the appropriate way, we might be able to crack the identifiability problem. (a) Is there some ob-

jective interpretation of intended meaning that will yield high consensus? (b) Is there some appropriate way to apply ideas about computational complexity without a clear understanding of our neural machinery? (c) Is there any useful physiological constraint on what cognitive processes we can propose? I doubt that any of these questions will yield answers leading to unique identifiability, but they seem the best hopes that we have.

I would like to conclude this article by sharing my earliest reason for doubting the possibility of unique identifiability. I did not make much of this reason in the original article. It is an inductive argument of sorts. We have observed a great deal of research and argument aimed at deciding this image–proposition issue. Although this effort has been richly successful in identifying new, interesting phenomena and developing new, important concepts, it seems to have been totally unsuccessful at moving us closer to deciding whether the underlying representation is imaginal or propositional. I simply cannot find one compelling reason to decide in favor of imagery or propositional representations (beyond personal bias), and many other researchers have made similar conclusions. Although this failure does not provide a rigorous basis, it does make one very suspicious that there is something basically wrong with the goal of deciding between the representational classes. In my 1978 *Psychological Review* article, I tried to present solid reasons for the difficulty. Despite the efforts of Hayes-Roth and Pylyshyn, those reasons still appear solid.

Reference Notes

1. Kieras, D. Personal communication, November 1978.
2. Anderson, J. R., & Kline, P. J. *Design of a production system.* Paper presented at the Workshop on Pattern-Directed Inference Systems, Honolulu, Hawaii, 1977.

References

Aho, A. V., Hopcroff, J. E., & Ullman, J. D. *The design and analysis of computer algorithms.* Reading, Mass.: Addison-Wesley, 1974.

Anderson, J. R. *Language, memory, and thought.* Hillsidate, N.J.: Erlbaum, 1976.

Anderson, J. R. Arguments concerning representations for mental imagery. *Psychological Review*, 1978, *85*, 249–277.

Atkinson, R. C., Holmgren, J. E., & Juola, J. F. Processing time as influenced by numbers of items in a visual display. *Perception & Psychophysics*, 1969, *6*, 321–326.

Chomsky, N. *Current issues in linguistics theory.* The Hague, Holland: Mouton, 1964.

Chomsky, N. *Aspects of the theory of syntax.* Cambridge, Mass.: MIT Press, 1965.

Cooper, L. A. Mental transformation of random two-dimensional shapes. *Cognitive Psychology*, 1975, *7*, 20–43.

Hayes-Roth, F. Distinguishing theories of representation: A critique of Anderson's "Arguments concerning mental imagery." *Psychological Review*, 1979, *86*, xxx–xxx.

Hennie, F. C. *Finite-state models for logical machines.* New York: Wiley, 1968.

Kosslyn, S. M. Information representation in visual images. *Cognitive Psychology*, 1975, *7*, 341–370.

Kosslyn, S. M., & Shwartz, S. M. A data-driven simulation of visual imagery. *Cognitive Science*, 1977, *1*, 265–295.

Metzler J., & Shepard, R. N. Transformational studies of the internal representations of three dimensional objects. In R. L. Solso (Ed.), *Theories of cognitive psychology: The Loyola symposium.* Hillsdale, N.J.: Erlbaum, 1974.

Meyer, A. R., & Shamos, M. I. Time and space. In A. K. Jones (Ed.), *Perspectives on computer science.* New York: Academic Press, 1977.

Minsky, M. L. *Computation: Finite and infinite machines.* Englewood Cliffs, N.J.: Prentice-Hall, 1967.

Paivio, A. *Imagery and verbal processes.* New York: Holt, Rinehart & Winston, 1971.

cyshyn, Z. W. Validating computational models: A Pylritique of Anderson's indeterminacy of representation claim. *Psychological Review*, 1979, *86*, xxx–xxx.

Rieger, C. Conceptual memory. In R. Schank (Ed.), *Conceptual information processing.* Amsterdam, Holland: North-Holland, 1975.

Sternberg, S. Memory scanning: Mental processes revealed by reaction time experiments. *American Scientist*, 1969, *57*, 421–457.

Thorson, J., & Biederman-Thorson, M. Distributed relaxation processes in sensory adaptation. *Science*, 1974, *183*, 161–172.

Townsend, J. T. Issues and models concerning the processing of a finite number of inputs. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition.* Hillsdale, N.J.: Erlbaum, 1974.

Received April 14, 1979 ■