

FORMAL ASPECTS OF NEURAL MODELING

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
CHARLES WAYNE JOHNSON
1971



This is to certify that the
thesis entitled

FORMAL ASPECTS OF NEURAL MODELING

presented by

Charles Wayne Johnson

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Philosophy

Richard J. Hall
Major professor

Date 13 May 1971

O-7639



ABSTRACT

FORMAL ASPECTS OF NEURAL MODELING

By

Charles Wayne Johnson

The purpose of this paper is to consider the question, "Is a sufficient neural model logically possible?" The consideration of this question is conducted in four stages.

First, some related concepts of "model" are characterized in terms of structural isomorphism. Neural models are specifically discussed, and a distinction between two sorts of "mechanism" is made, with respect to neural modeling. Mechanism₁ is designated as the view that a sufficient neurophysiological theory is logically possible. Mechanism₂ is designated as the view that a sufficient neural model is logically possible. The concept of "sufficiency," in the case of a sufficient neurophysiological theory, is characterized by a predictivity criterion, which is stated in terms of the covering-law model of scientific explanation. In the case of a sufficient neural model, the concept of "sufficiency" is characterized by a behavioral criterion.

Secondly, Mechanism₁ is defended against three arguments which are used to attempt to demonstrate that Mechanism₁ is self-refuting.

Thirdly, Mechanism₂ is defended against the argument that Gödel's 1931 findings show that a sufficient neural model is not logically possible.

Fourthly, the Brain-Model Problem is briefly discussed. The Brain-Model Problem is an epistemological problem that may be characterized by the question, "Can a sufficient neural model be simple enough to be understood?" The general conclusion is drawn that both Mechanism₁ and Mechanism₂ are viable. Finally, some suggestions are made for future research.

A comprehensive bibliography is included, together with two appendices, which contain some representative arguments against Mechanism₁ and Mechanism₂.

FORMAL ASPECTS OF NEURAL MODELING

By

Charles Wayne Johnson

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Philosophy

1971

67000

For Buffy

ACKNOWLEDGMENTS

I am fortunate in having many people to thank for their assistance and advice in various phases of this study. I am very grateful to Dr. Richard J. Hall for his perceptive and enthusiastic comments on this work; his remarks were instrumental in the elimination of several errors and ambiguities. To Dr. Edward M. Eisenstein, I owe thanks, not only for his advice in connection with this study, but also for providing the opportunity for me to actually do some science rather than merely study it. I must also thank Dr. Herbert E. Hendry and Dr. Peter D. Asquith for their helpful and illuminating comments. Additionally, I am indebted to Dr. William L. Kilmer whose timely advice regarding reference material greatly expedited my research for this paper. I must, of course, thank my wife, Patti, for, among other things, assisting me with typing and proof reading.

This study was supported by a Woodrow Wilson Dissertation Fellowship.

LIST

LIST

§1.0

§2.0

§2.1

§2.2

§2.3

§2.4

§3.0

§3.1

§3.2

§3.3

§3.4

§4.0

§4.1

§4.2

§4.3

§4.4

§5.0

§5.1

§5.2

BIBLIOG

APPENDI

APPENDI

TABLE OF CONTENTS

	Page
LIST OF TABLES	v
LIST OF FIGURES	vi
§1.0 INTRODUCTION	1
§2.0 PROBLEMS OF NEURAL MODELING	6
§2.1 THE EMPIRICAL SIGNIFICANCE OF FORMAL/ MATHEMATICAL NEURAL MODELS	16
§2.2 MECHANISM ₁	26
§2.3 MECHANISM ₂	28
§2.4 DISCUSSION: THE BRAIN-MODEL PROBLEM	32
§3.0 MECHANISM ₁ AND THE SELF-REFUTATION ARGUMENT	37
§3.1 THE INTENTIONAL PARADOX	44
§3.2 THE RATIONAL PARADOX	65
§3.3 THE DESCRIPTION PARADOX	79
§3.4 DISCUSSION	85
§4.0 MECHANISM ₂ AND THE GÖDEL ARGUMENT	87
§4.1 NEURAL ARTEFACT SUFFICIENCY AND THE CONCEPT OF "MACHINE"	89
§4.2 THE GÖDEL ARGUMENT	109
§4.3 ARTEFACTS, BRAINS, AND GÖDEL	119
§4.4 DISCUSSION	152
§5.0 MECHANISM ₂ AND THE BRAIN-MODEL PROBLEM	156
§5.1 THE BRAIN-MODEL PROBLEM	161
§5.2 SUGGESTIONS FOR FUTURE RESEARCH	177
BIBLIOGRAPHY	181
APPENDIX I	193
APPENDIX II	196

LIST OF FIGURES

	Page
Figure 2.0-1	8
Figure 2.0-2	13
Figure 2.0-3	13
Figure 2.1-1	18
Figure 2.1-2	19
Figure 2.1-3	19
Figure 2.1-4	19
Figure 2.1-5	20
Figure 3.2-1	77
Figure 4.1-1	97
Figure 4.1-2	100
Figure 5.1-1	166

LIST OF TABLES

	Page
Table 4.2-1	114
Table APP. II-1	207

§1.0

INTRODUCTION

The enterprise of modeling the human brain has met, as is to be expected, with enormous practical difficulties, but it has also resulted in the generation of problems of both a conceptual and logical nature. The conceptual difficulties have largely resulted from attempts to reconcile the mind-body dichotomy that is so entrenched in language and in thought. Logical considerations, on the other hand, have played a major role in both the formulation of the formal structure of neural models and in dealing with the general question of whether, in the final analysis, the construction of a sufficient neural model is logically possible.

The present study is primarily concerned with the formal aspects of neural modeling. The central question which we shall consider is: "Is it logically possible to construct a model of the human brain that can perform characteristically human (rational) activities and is also simple enough to be understood?"

As a preliminary measure, in the consideration of this question, it will be necessary to discuss the logical

structure of models in general, and of neural models in particular. Although it will not be possible to avoid conceptual matters entirely, these will not be stressed.

Considering the formal relationship that obtains between a model and the thing which it models, it would seem that advancements in the direction of constructing a sufficient neural model would be largely dependent upon the state of development of neurophysiology. The late Warren McCulloch, however, held that some significant advances can be made in this direction somewhat in advance of the corresponding discoveries in neurophysiology. This would seem to particularly be the case in studies wherein internal brain processes are ignored and a "black box" model is all that is desired. Such models may, in fact, aid in the development of neurophysiology. We may, therefore, expect considerations regarding neural modeling to provide a convenient way to study neural structures and neural functions. Neural models may also provide a new way of discussing classical problems associated with the alleged mind-body dichotomy.

At any rate, the present study is not adversely affected by the absence of a sufficient neural model; in fact, it is not obvious that the existence of a candidate model would resolve the question with which we are presently concerned, since the question regarding the logical

possibility of a sufficient neural model (together with, of course, considerations regarding sufficiency criteria) must be settled prior to the evaluation of a given candidate model.

The present study proceeds in four stages:

(1.) In §2, a working definition of 'model' is formulated in terms of structural similarities. The empirical value of logical and mathematical models of neural systems is assessed with particular emphasis placed upon network models. A distinction is then made between two senses of 'mechanism,' both of which are suitably restricted in meaning for the purposes of this study. Specifically, we take Mechanism₁ to be the proposition that a sufficient neurophysiological theory is logically possible. We take Mechanism₂, on the other hand, to be the proposition that a sufficient neural model is logically possible. Some problems that are peculiar to neural modeling are introduced; these problems are related to the two senses of 'mechanism.'

Notable among these problems are: (a.) the question of whether Mechanism₁ is self-refuting, (b.) the question regarding the possible implications for Mechanism₂ of Gödel's 1931 findings, and finally, (c.) the question of whether or not a neural model can both be sufficient and be simple enough to be understood. These questions are the topics of the remaining three sections.

(2.) In §3, we examine the self-refutation argument against Mechanism₁. This examination is done in terms of three paradoxes which allegedly prove that Mechanism₁ is self-refuting.

(3.) It has been argued that Gödel's 1931 findings provide the basis for an argument which refutes Mechanism₂ by showing that there will always be something that a human brain can do that cannot be done by a model (viz., a "machine" or "artefact"). In §4, this argument is considered in some detail.

(4.) Assuming that both a sufficient neurophysiological theory and a sufficient neural model are logically possible, the question remains of whether a model can be constructed which is both sufficient to model the human central nervous system (CNS) and is simple enough to be understood. This is an important feature of the more general Brain-Model Problem, and is the topic of §5.

The Brain-Model Problem indicates some intriguing quandries which are involved with the situation wherein one system attempts to both model and understand another system which is of its own level of sophistication.

In §5, we also summarize our findings and conclusions, and we briefly discuss modeling sufficiency criteria. Finally, we make some brief recommendations for future research.

which

brain v

to the

Neither

of a b

brain,

conside

As we shall indicate, a neural model of the sort which is desired may be a "black box" model of the human brain which has input-output correlations which correspond to the input-output correlations of a rational human brain. Neither the internal processes nor the internal structures of a black box model need be similar to those of the human brain, in order for the model to be sufficient. These considerations will be clarified later.

§2.0

PROBLEMS OF NEURAL MODELING

A considerable number of somewhat divergent views have historically come under the general heading of "mechanism." For our purposes, we shall use 'mechanism' to refer to the view that the question "Is it logically possible to construct a sufficient neural model?" should be answered affirmatively. There are, however, some preliminary considerations which must be discussed.

In common usage, 'model' is quite ambiguous; in scientific model theory, however, it is given a rather precise rendering. We shall, in accordance with current literature in scientific modeling, use the term in three distinct senses: (1) "logical model," (2) "theoretical model," and (3) "physical model" or "artefact."

Our use of the term 'logical model' coincides with the sense of 'model' that is employed by logicians. We shall say that an entity, M_0 , is a logical model of a set of sentences, T_0 , if and only if M_0 satisfies T_0 , where 'satisfaction' is defined in the usual manner (an excellent explication of '(logical) model' may be found in [MEN 1]).

which

to the

'mode

forma

by, fo

as [ME

cal mo

to and

model,

uninte

tion.

two th

(for a

[SWA 1]

rather

that it

a theor

say tha

pretati

physical

entity,

By 'theoretical model' we shall mean a theory which is isomorphic to another theory, which it is said to theoretically model. For the purpose of this sense of 'model,' one must consider the theories in question to be formalized, in the sense of 'formalization' which is used by, for example, P. Suppes (See [SUP 1], [SUP 2] as well as [MEN 1], [BYE 1], and [APO 1]).

In a closely related characterization of 'theoretical model,' one speaks of a theory, T_1 , as being related to another theory, T_0 , which it is said to theoretically model, in virtue of being an interpretation of the same uninterpreted calculus, C , of which T_0 is an interpretation. This is typically referred to as the "one calculus--two theories" characterization of theoretical modeling (for a detailed treatment of this characterization, see [SWA 1]).

If a theoretical model is a mathematical theory, rather than an empirical scientific theory, we shall say that it is a mathematical-theoretical model. Similarly, a theoretical model is a theory of formal logic, we shall say that it is a formal-theoretical model.

We shall use the terms 'isomorphism' and 'interpretation' as they are defined in [MEN 1].

By 'physical model' or 'artefact' we shall mean a physical entity which is isomorphic to another physical entity, which it is said to physically model. We shall

presu

a cor

repres

shoul

ended

Furthe

be a t

cal mo

tional

we hav

cal mod

ment of

they ar

explore

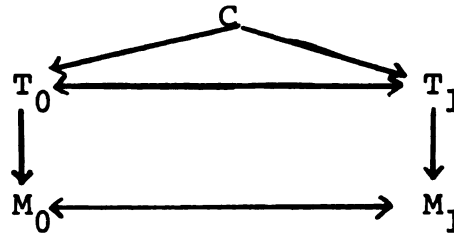
structu

well-de

deducti

presume that a physical model will be a logical model of a corresponding theoretical model.

The general modeling framework may be graphically represented by Figure (2.0-1); the single-ended arrows should be interpreted as 'is satisfied by,' and the double-ended arrows should be interpreted as 'is isomorphic to.' Furthermore, let C be an uninterpreted calculus; let T_0 be a theory, which is to be modeled; let T_1 be a theoretical model of T_0 ; finally, let M_0 and M_1 be physical relational systems which satisfy T_0 and T_1 respectively. Then we have:



(Figure 2.0-1)

Typically, one speaks of the purpose of theoretical modeling as being to expedite the deductive development of theories that are "rudimentary," in the sense that they are (1) theories in fields that have not been well-explored, and (2) they have "gaps" in their deductive structures. In the theoretical modeling procedure, a well-developed theory, T_1 , is used as a pattern for the deductive development of a rudimentary theory T_0 .

There are, in general, two ways in which physical modeling can proceed. Specifically, one way to proceed in physical modeling would be to attempt to have the artefact perform the relevant modeled operations of the original in the same way (or in a very similar way) that the original system performs these operations. For example, one way to approach the problem of modeling the human CNS is to attempt to produce an electronic artefact (for example, a computer) which, in virtue of its circuitry (or programming) effectively replicates the transmission structure, as well as general functions of the human CNS.

On the other hand, one may treat the system to be modeled as a "black box" (or, alternatively, one may treat discrete portions of the original system as black boxes). In this case, one would be interested simply in producing a physical model which has inputs and outputs that are in respective correspondence to the inputs and outputs of the original system. In this way, one may concentrate on, and take advantage of, the particular capabilities of the prospective artefact. For example, in modeling the human brain, one may choose to treat the entire brain as a black box and then attempt, in the most efficient and expedient way possible, to produce an (e.g., electrical) artefact, which has an input-output correspondence that is similar to that of the original system. In the use of this sort of technique, the circuitry (or pattern of transmission

paths) of the original system, i.e., the human brain, need have no structural correspondence to the circuitry of the artefact.

Neural systems are occasionally modeled by electrical systems, i.e., by fixed circuitry systems, or much more commonly, they are modeled by the more flexible computer techniques. Black box procedures may be used where only functional similarity between the prototype and model is desired. An intuitive sufficiency criterion for physical models of neural systems may be provided by, for example, Turing's Test.¹

Questions regarding the possibility of a sufficient neural model are, to a large extent, dependent upon an affirmative resolution to the question regarding the possibility of a sufficient neurophysiological theory. While, it is true that modeling developments may take place somewhat in advance of corresponding discoveries in neurophysiology,² we may expect questions regarding the possibility of a sufficient neurophysiological theory to be conceptually, epistemologically, and perhaps, logically prior to questions regarding the possibility of sufficient theoretical neural models or physical neural models.

¹See §4.

²See the views of W. S. McCulloch in his [McC 1], passim.

The situation in the case of neural modeling is somewhat unique in that mathematical models used in neural modeling are highly developed deductively but are poorly developed in another sense. They are deductively well-developed, but, to date, neither they nor their logical models have resulted in sufficient neural models. To a large extent, this situation is due to relatively low level of development of neurophysiology with respect to the explanation of total brain functions. Apparently, an empirical theory must achieve a certain degree of sophistication before mathematical modeling can be initiated on a large scale. It would seem that formal/mathematical neural modeling fits into the modeling framework rather well. The remaining question as to whether any new mathematical developments are required to facilitate sufficient neural modeling has been answered largely in the negative by, for example, M. A. Arbib,³ who sees the modeling problem as being one of developing computer techniques, and computer techniques are, for the most part, based on existing mathematics. He urges, as we have, a dropping of the distinction between "real mathematics" and "computer programming."

We may generally state that neural modeling falls into the modeling framework, which is represented in

³See, for example, [STA 1].

(Figure 2.0-1). This point is partially substantiated by the following remark made by Perkel and Moore in their defense of neural modeling:

The most significant use of neural models, analog or digital, is a heuristic one: effectively used, they suggest possibilities; they encourage the enumeration of specific alternatives; they force the precise specification of alternative hypotheses; and they serve as an ideal tool for the elaboration of the consequences of hypotheses; in short, they furnish a vehicle for the application of the principle of strong inference so essential to sustained scientific progress.⁴

Note that Perkel and Moore regard neural models as being "tools." If they are correct, then we may conclude that neural models are of instrumental value; hence, they may be convenient to use, but they are not essential to neurophysiological theories.

Generally speaking, we may distinguish between two sorts of mechanism (and correspondingly, we may distinguish between two sorts of "mentalism"). We shall take Mechanism₁ to be the proposition that a sufficient neurophysiological theory is logically possible. Mechanism₂, on the other hand, we shall take to be the view that a sufficient neural artefact is logically possible. Accordingly, we take Mentalism₁ to be the denial of Mechanism₁, and Mentalism₂ to be the denial of Mechanism₂.

⁴[PER 1].

large

The mo

If we

cal the

$T_0/NT =$

then, w

While i

duced i

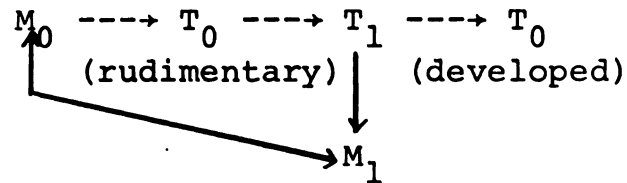
seems f

that th

against

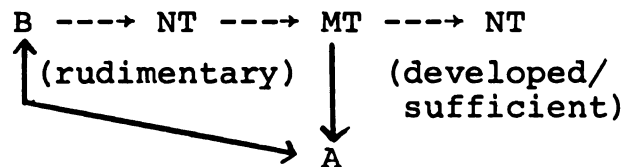
regardi

It is our contention that Mechanism₂ is, to a large extent, dependent upon an affirmation of Mechanism₁. The modeling procedure may go in the following direction:



(Figure 2.0-2)

If we let M_0 = a human brain = B; T_0 = a neurophysiological theory = NT; T_1 = a formal/mathematical model of T_0 /NT = MT; and M_1 = a neural artefact (or "machine")=A; then, we have:



(Figure 2.0-3)

While it seems possible that a sufficient A could be produced independently of a sufficient NT (i.e., an SNT), it seems for practical, heuristic, and other reasons unlikely that this will occur.

It seems apparent that many of the arguments against Mechanism₁ (especially, for example, arguments regarding intentions, consciousness, determinism, etc.)

could

argume

conclu

to Mech

truth o

questio

settled

thermon

upon an

at this

term 'n

in this

however

informa

commonl

involve

rather

and cat

units o

ted par

neural

could also be applied against Mechanism₂. Hence, if these arguments are successful in refuting Mechanism₁, it may be concluded that equal success could be expected with regard to Mechanism₂.

Our general conclusion is that the question of the truth of Mechanism₁ is, in several respects, prior to the question of the truth of Mechanism₂, and, hence, should be settled before considering Mechanism₂. We conclude, furthermore, that generally speaking, Mechanism₂ is based upon an affirmation of Mechanism₁.

One additional clarification seems to be required at this point. There are two prevalent senses of the term 'neural model.' We presume that the sense intended, in this study, is now clear. In an alternative sense, however, 'neural modeling' refers to the deciphering of information encoded in neural activity. This is more commonly called 'neural coding.'⁵ Neural coding does not involve the functional modeling of neural systems, but rather seeks to interpret the on-going activity of a CNS and categorize, insofar as is feasible, meaningful discrete units of neural activity with respect to previously selected parameters.

There is an overlap between neural modeling and neural coding which, as we shall indicate in §5, results

⁵See [PER 2].

in inte

descrip

practic

quiries

lated t

neural

nism,

finally

culties

sed in

three p

and will

in interesting problems of self-reference and self-description.

The problems of neural modeling, aside from purely practical problems that are common to all scientific inquiries, fall into roughly four areas: (1) problems related to the empirical significance of formal/mathematical neural models, (2) the questions of the truth of Mechanism₁, (3) the question of the truth of Mechanism₂, and finally (4) the self-reference and epistemological difficulties which are associated with the Brain-Model Problem.

The first of these problems, (1), will be discussed in the following section, §2.1, and the remaining three problems will be introduced in the remainder of §2, and will be discussed in detail in later sections.

biologi

neurons

is, by

of which

hope to

unit ne

indicat

ens

a p

wat

mol

our

in-

So, for

unit mo

to neur

system-

§2.1

THE EMPIRICAL SIGNIFICANCE OF FORMAL/MATHEMATICAL NEURAL MODELS

It is both unnatural and inaccurate to regard biological neurons as being discrete units.⁶ Biological neurons are best characterized by their histories, that is, by the history of their functions within the systems of which they are components. In addition, we cannot hope to be able to infer systemic functions from single-unit neural functions. As, for example, L. D. Harmon indicates,

In general, . . . we can no more easily deduce ensemble function from the single-unit function than a physicist can deduce the physical properties of water from his knowledge of hydrogen and oxygen molecules. Both difficulties are practical ones at our present levels of understanding; they may not be in-principle impossibilities.⁷

So, for the present, at least, it would seem that single-unit models would not be likely to be instructive relative to neural systems (or, for that matter, relative to system-models, viz., neural nets). As it turns out,

⁶See [MIN 5].

⁷[HAR₁ 1].

single

ticular

neurons

formal,

nal neu

the cor

(i.e.,

formal

McCullo

biologi

we shal

scribed

values,

an input

has n s

x_i term

$x_i > 0$.

integer

and

Having

this po

single-unit formal/mathematical models are not even particularly instructive with respect to the single-unit neurons which they model. The reason for this is that formal/mathematical single-unit neural models, viz. "formal neurons," are very simplified and idealized models of the corresponding portions of neurophysiological theory (i.e., they are highly incomplete).

To clarify this point, we shall briefly consider formal neurons and neural nets, as based upon the McCulloch-Pitts framework. We shall, then, note the biological inaccuracies of single-unit formal neurons and we shall indicate the heuristic value of neural nets.

Let N_ψ be a formal neuron. Then, N_ψ may be described using $n+1$ numbers. The first n numbers are input values, $v_1, v_2, \dots, v_n$, where for every v_i there is an input line (or path) x_i to which v_i is associated. N_ψ has n such inputs, $x_1, x_2, \dots, x_n$ ($n \geq 1$). For each x_i , x_i terminates in an excitatory endbulb if and only if $x_i > 0$. If $x_i > 0$, then $v_i = y$, where y is some positive integer. If $x_i < 0$, then, $v_i = -\infty$. Let

$$'N_\psi \text{ fires along } 0_\psi \text{ at time } t' =_{df} 'N_\psi(t) = 1'$$

and

$$'N_\psi \text{ does not fire along } 0_\psi \text{ at time } t' =_{df} 'N_\psi(t) = 0'$$

Having introduced these symbolic conventions, we may, at this point, posit the following three axioms:

(N-I)

(N-II)

(N-III)

duced in

neuron,

we shall

Let

$G=n$

the axio

sented a

(N-I) N_ψ functions on the basis of a sequence of discrete moments of time. The respective moments may be indexed by the integers:

$$t = 1, 2, 3, \dots$$

(N-II) For every N_ψ there are exactly two possible states:

A firing state: $N_\psi(t_i) = 1$, and

A resting state: $N_\psi(t_i) = 0$

Hence, $(N_\psi)(t_i)(N_\psi(t_i) = 1 \vee N_\psi(t_i) = 0)$

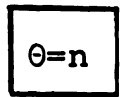
(N-III) $(N_\psi(t+1) = 1) \equiv \sum_i v_i x_i(t) \geq \theta$ [the dependency rule]

where ' θ ' =_{df} 'the threshold of N_ψ '

Considering axiom (N-I), a delay may be introduced in a transmission path by introducing a formal neuron, N_ψ , having $\theta = 1$. To graphically represent this, we shall make use of the following symbolic conventions:

Let $\text{---}\bullet$ represent an excitatory endbulb, and let

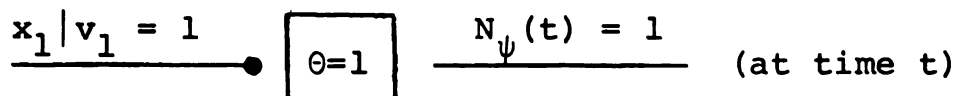
$\text{---}\circ$ represent an inhibitory endbulb. Finally, let



represent a formal neuron, as characterized by

the axioms (N-I), (N-II), and (N-III).

Using these conventions, a delay may be represented as:



(Figure 2.1-1)

excitatio

neuron,

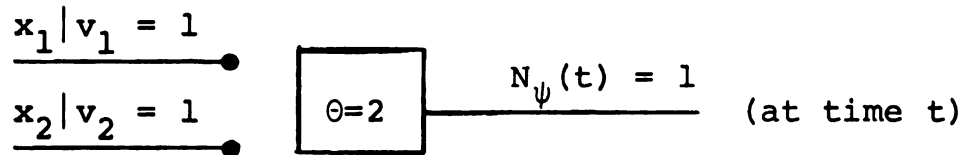
excitatio

neuron N

using an

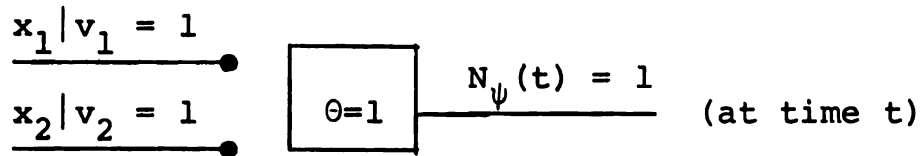
has a 6

A simple and-gate may be constructed using two excitatory inputs, each having values of 1, and a formal neuron, N_ψ , having $\Theta = 2$:



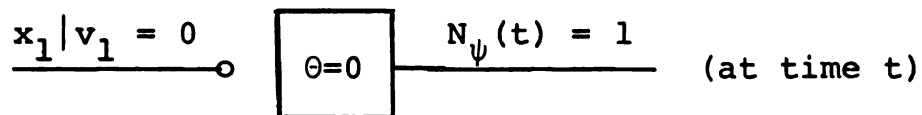
(Figure 2.1-2)

A simple or-gate may be constructed using two excitatory inputs, each having values of 1 and a formal neuron N_ψ , having a $\Theta = 1$:



(Figure 2.1-3)

Finally, a simple negation-gate may be constructed using an inhibitory input and a formal neuron, N_ψ , which has a $\Theta = 0$:



(Figure 2.1-4)

1).

defined

more adv

the pres

We may d

manner:

sati

may be e

culus.

graphica

threshol

restrict

N_1

N_2

N_3

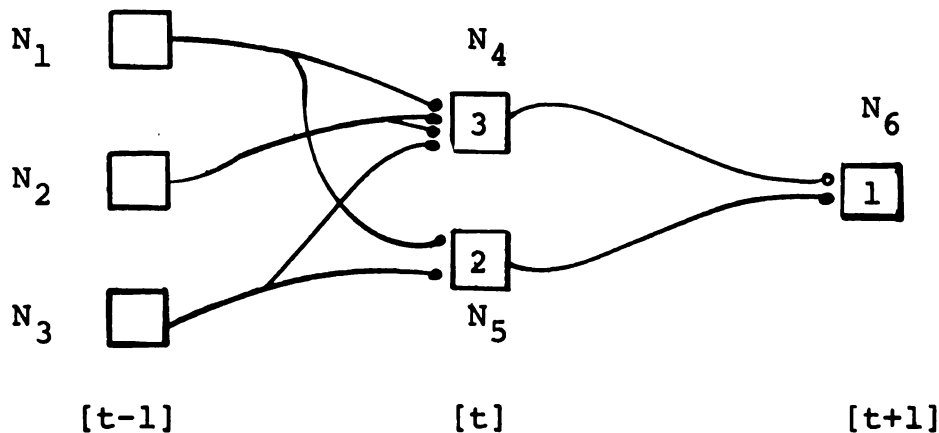
{

Clearly, if N_ψ is a negation-gate, then $(t_i)(N_\psi(t_i) = 1)$.

The term 'brain' (or, alternatively, 'CNS') may be defined as 'an aggregate of formal neurons.' It may be more advisable, however, to avoid cognitive issues, for the present, and speak instead in terms of 'neural nets.' We may define the term 'neural net' in the following manner:

A system S is a neural net if and only if it satisfies the axioms (N-I), (N-II), and (N-III). (2.1-1)

As McCulloch and Pitts demonstrated, neural nets may be effectively described using the propositional calculus. Consider, for example, the neural net which is graphically represented in (Figure 2.1-5). Clearly, the thresholds of N_1 , N_2 , and N_3 need not be specified in this restricted example.



(Figure 3.1-5)

We shall

$N_5 =$

sion:

sion:

Using (2

following

ure 2.1-

Theorem:

model of

ted, by,

formal n

specifie

We shall suppose, however, that we have $N_4|_{\theta_4} = 3$,
 $N_5|_{\theta_5} = 2$, and $N_6|_{\theta_6} = 1$.

N_4 may be characterized by the following expression:

$$N_4(t) \equiv N_2(t-1) \cdot (N_1(t-1) \vee N_3(t-1)) \quad (2.1-2)$$

N_5 may be characterized by the expression:

$$N_5(t) \equiv N_1(t-1) \cdot N_3(t-1) \quad (2.1-3)$$

Similarly, N_6 may be characterized by the expression:

$$N_6(t+1) \equiv \sim(N_4(t) \cdot \sim N_5(t)) \quad (2.1-4)$$

Using (2.1-2), (2.1-3) and (2.1-4), we may generate the following theorem for the neural net represented in (Figure 2.1-5):

Theorem: $N_6(t+1) \equiv \sim((N_2(t-1) \cdot (N_1(t-1) \vee N_3(t-1))) \cdot$

$$(N_1(t-1) \cdot N_3(t-1))). \quad (2.1-5)$$

Such a system as S is, clearly, only an incomplete₁ model of a neurophysiological theory. As has been indicated, by, for example, McCulloch, M. Minsky, and others, formal neural models, especially single-unit models (as specified, for example by (N-I) and (N-II)) are extremely

simplified

They poss

theory,

some of

simply fa

scribed b

there are

sociated

framework

system de

an "all-

While, i

more str

More imp

S fixes

formal ne

absolute

noted th

reliable

all time

neural s

the simp

which is

Form in

take use

pp. 3-7.

system, s

simplified mathematical models of biological neurons. They possess what is frequently called in scientific model theory, large "negative analogies." In addition, however, some of the assumptions made in formal neural modeling are simply false when compared to biological neurons as described by a neurophysiological theory. Specifically, there are five general ways in which the assumptions associated with formal neural modeling in the McCulloch-Pitts framework, may be shown to be false.⁸

First, it is assumed, in the construction of the system described above, i.e., S, that neural activity is an "all-or-none" process (viz., as specified by (N-II)). While, in general, this is true of biological neurons, a more strict biological account may not be so inflexible. More importantly, however, we should note the fact that S fixes the output values and thresholds of all of the formal neurons for all time; biological neurons are not absolutely stable in these respects. It should also be noted that, in this simple system, S, all elements are reliable and the structure of a neural net is fixed for all time.⁹ Here again, this is not true of biological neural systems.

⁸This discussion will be carried out in terms of the simplified, non-technical account of neural operations which is provided by W. S. McCulloch in his "Finality and Form in Nervous Activity." [McC 1], pp. 256-275. We also make use of some points made by M. A. Arbib; see [ARB 1], pp. 3-7.

⁹For a more detailed and biologically accurate system, see [RAS 1], passim.

is not t

synchron

moments

ity of a

that int

form; CN

function

delay pe

neurons

findings

other in

signific

CNS.

solutely

nected.

actual b

by syste

neurons

neurons,

tended t

[NEN 5],

Secondly, S treats time in an unnatural way. It is not true of biological neurons that they are completely synchronatated. In addition, the designation of discrete moments of time may involve dividing the on-going activity of a CNS in an arbitrary manner. (N-1) also requires that inter-neuronal transmission time be absolutely uniform; CNS delays, however, could, for example, also be a function of axonal diameter, whereas the only sort of delay permitted in system S is synaptic delay.¹⁰

Thirdly, S ignores all interaction between formal neurons except those that take place at synapses. Recent findings, for example, those of Lettvin, indicate that other interaction, such as inter-systemic noise, can be significant in the efficient information processing of a CNS.

Fourthly, in system S, an inhibitory endbulb absolutely inhibits the formal neuron to which it is connected. It is not clear that this is always the case in actual biological inhibition.

Fifthly, as Arbib indicates, no account is taken, by system S, of possible changes in behavior of formal neurons as the result of chemicals or hormones. Formal neurons, at least in the McCulloch-Pitts model, are intended to simulate only electrical activity.

¹⁰For an alternative treatment of time, See [MIN 5], pp. 2-7.

of the
exhaust
size th
incompl
simplif
taken t
such as
S is ob
neurons
Accordi
cal mod
model e
stitute
biologic
used as
prove to
theories

of forma
general
this res
neural n
theories
function
respondi

This brief survey of the biological inaccuracies of the formal neurons of a system such as S is by no means exhaustive. We mention these inaccuracies only to emphasize the sort of simplification that one might expect in incomplete single-unit neural models. The fact that the simplification is so drastic, however, should not be taken to reflect unfavorably on the value of neural nets, such as S. Considered as a source of single-unit models, S is obviously defective. But, as we noted, biological neurons are best understood as components of systems. Accordingly, we may expect that the most useful mathematical models of neurophysiological theories are those that model entire systems, viz., mathematical models that constitute neural nets. So, we may ask, considering the biological inaccuracies of formal neurons, can they be used as the basic constituents of neural nets which will prove to be valuable as models of neurophysiological theories regarding biological neural systems?

The consensus seems to be that nets constructed of formal neurons are of value in that they model the general logical functions of biological systems. In this respect, it is perhaps more transparent to regard neural nets as being formal models of neurophysiological theories. It is held, by Harmon, for example, that the functions of a neural model and the functions of a corresponding neurophysiological theory may be similar

without the

basic const

Alt.
in form
sentati
ceivabl
prototy
specifi

It should be

to distinguish

physical mod

productive o

sciences the

It s

models, such

were never r

physiological

theories. C

mathematical

plete (in B

models of th

formal/mathe

the logical

display disc

as well as b

11 [H

12 Fc

see [MIN 1],

without there being a corresponding similarity in their basic constituents. As Harmon expresses this point,

Although the "neurons" used in these models [viz., in formal models] are exceedingly simplified representations of biological neurons, it is entirely conceivable that the gross behavior in both models and prototype is essentially independent of finer unit specification.¹¹

It should be noted, to avoid confusion, that Harmon fails to distinguish between formal/mathematical models and physical models. This failure is a common one, and it is productive of confusions in both model theory and in the sciences themselves.

It seems clear that formal/mathematical neural models, such as the one formulated by McCulloch and Pitts, were never really intended to be complete, accurate, physiologically oriented models of neurophysiological theories. Consequently, the logical models of formal/mathematical neural models were not intended to be complete (in Brodbeck's sense of 'completeness') physical models of their original neural prototypes. Rather, formal/mathematical neural models are intended to model the logical structures of various sorts of systems which display discrete processes; such systems may be computers as well as brains.¹²

¹¹[HAR₁ 1], p. 24.

¹²For a more detailed discussion of this point, see [MIN 1], Chapter 3.

We s

tion that an

ments which

that there a

psyche/soul/

explanatory

that some as

be explained

explanations

on an obscur

related to c

Twen

gone a long

The more lin

debate, have

that portion

as being rat

tinctively h

Mechanism₁ i

§ 2.2

MECHANISM₁

We shall regard Mechanism₁ as being the proposition that an SNT is logically possible. The main arguments which are used by Mentalists₁ against this view is that there are some features of the human mind/spirit/psyche/soul/etc. which are emergent with respect to causal explanatory theories. In other words, Mentalists₁ contend that some aspects of distinctively human activity cannot be explained by means of contingent, causal, scientific explanations. Much of the debate, in this regard, has been on an obscure, metaphysical level, dealing with topics related to categories and substances.

Twentieth Century behaviorists, however, have gone a long way toward clarifying the metaphysical debate. The more linguistic questions which are related to the debate, have, however, remained. So, considering just that portion of human activities that may be characterized as being rational, intentional, or, more generally, "distinctively human," we may reformulate the proposition of Mechanism₁ in terms of the question of whether or not

such activi

a causal ex

The

been the so

Mentalist₁ a

that is pote

that Mechan

successful,

for its own

The

will be exan

such activities can be sufficiently explained by means of a causal explanation, i.e., by contingent causal laws.

There are many important side issues which have been the source of seemingly endless debate, but the one Mentalist₁ argument that is most directly to the point and that is potentially the most devastating is the argument that Mechanism₁ is self-refuting. If this argument is successful, then either Mechanism₁ constitutes evidence for its own falsehood, or it is self-contradictory.

The self-refutation argument against Mechanism₁ will be examined, in some detail, in §3.

We

it is logic

neural math

Most often,

or "machines

of words in

and misuse,

that 'model

(as is, for

the content

future, "thi

(2) 'thinkin

Many

Mechanism₂ H

outdated cha

discussing A

definition o

the possibili

§2.2

MECHANISM₂

We may take Mechanism₂ to be the proposition that it is logically possible for there to be both a sufficient neural mathematical model and a sufficient neural artefact. Most often, Mechanism₂ is discussed in terms of artefacts or "machines." It is unfortunate that there are a number of words in natural language that, as a result of long use and misuse, have become extremely vague. We have seen that 'model' is one such word; 'machine' is another one (as is, for example, 'life,' or 'intelligence'). It is the contention of Mechanists₂ that (1) within the near future, "thinking machines" will be developed¹³ and that (2) 'thinking' is not incompatible with 'machine.'

Many of the arguments which have been used against Mechanism₂ have been based upon an overly restricted or outdated characterization of "machine." Therefore, in discussing Mechanism₂, we shall be on guard against a definition of 'machine' that has a built-in bias against the possibility of there being an SNM. In addition, we

¹³Turing speculates that this will occur within fifty years. See [TUR 1], passim.

must watch for
confusion of
such as 'mat
tury counter
scientific r
extensive re
that a Sever
call a moder
tively argue
body problem
from Sevente

We s

problems fac
nor shall we
developed an
also attempt
classically
consider the
are possible

Mos

bility of "
viously pos
This proced

14 S

15 F

must watch for possible ambiguities which result from the confusion of Seventeenth Century definitions of terms such as 'matter' and 'machine' with their Twentieth Century counterparts. It has been pointed out, by several scientific methodologists, that these terms have undergone extensive revisions with regard to meaning, so much so that a Seventeenth Century dualist would very likely not call a modern computer a "machine."¹⁴ It can be effectively argued, in fact, that the entire classical mind-body problem may be based upon a misconception derived from Seventeenth Century mechanism.¹⁵

We shall not discuss, in this study, the practical problems facing the development of "thinking machines," nor shall we discuss the feasibility of their being developed and constructed in the near future. We shall also attempt to avoid, inasmuch as is possible, the more classically philosophical side-issues. We shall merely consider the question of whether or not "thinking machines" are possible in principle.

Most of the arguments against the logical possibility of "thinking machines" seem to depend upon previously posited definitions of 'mind' and 'machine.' This procedure seems to involve equivocation as well as

¹⁴See, for example, [TOU 3], [SMA 3], and [NAG 2].

¹⁵For example, see [RYL 1], [TOU 2], and [TOU 3].

involving t

avoid.

What

'artefact'))

to avoid th

term 'neura

A p

tion regard

one thing t

by an artef

this questi

"behavioral

biological

Give

is at least

artefact can

futed. If,

there is at

that cannot

Mechanism₂

It

one of its

always be so

artefact can

artefacts ar

program or

involving the sort of definitional bias which we wish to avoid.

What we mean by 'machine' is 'physical model' (or 'artefact'), as characterized in §2.0. We shall attempt to avoid the vague word 'machine' and use instead the term 'neural artefact' or simply, 'artefact.'

A possibly more effective way of posing the question regarding Mechanism₂ is to ask: Is there at least one thing that the human brain can do that cannot be done by an artefact? Of course, the sort of "things" to which this question refers should be restricted to rational "behavioral" outputs, and should exclude the specifically biological functions of the brain.

Given this restriction, it appears that, if there is at least one thing that a human brain can do that an artefact cannot do, then Mechanism₂ is effectively refuted. If, on the other hand, it is not the case that there is at least one thing that the human brain can do that cannot be done by an artefact, then, to this extent, Mechanism₂ is viable.

It has been argued that Gödel's Theorem has, as one of its philosophical consequences, that there will always be something that a human brain can do that an artefact cannot do. The reasoning, very briefly, is that artefacts are "programmed" (where, for our purposes, a program or "machine description" is a formal/mathematical-

theore

fixed

that a

mentar

can al

not pr

tion i

viz.,

in §4.

theoretical model) to do a finite set of operations by a fixed set of directions (or rules, or axioms). Assuming that a neural artefact is sophisticated enough to do elementary arithmetic, then it is argued that a human brain can always generate a formula that is true but that is not provable by the artefact (i.e., the formula in question is not derivable from the artefact's formal system, viz., the system which the artefact satisfies).

We shall consider this argument, in some detail, in §4.

in the
viron
This
it gi
cal e
secon
lems

serve
as we
percep
model
second
rary m
a some
The se

scienti

§2.4

DISCUSSION: THE BRAIN-MODEL PROBLEM

As has been pointed out by various investigators in the field of neural modeling, a CNS "models" its environment,¹⁶ i.e., a CNS maps its environment onto itself. This finding has two interrelated consequences. First, it gives rise to an alternative formulation of the classical epistemological questions posed by empiricists, and secondly, it results in intriguing self-reference problems in neural modeling.

Concerning the first consequence, we should observe that a distinction is occasionally made by scientists, as well as philosophers, between immediate and mediate perceptions.¹⁷ Correspondingly, we may say that a primary model is a direct representation of an environment. A secondary model, on the other hand, is a model of a primary model (we are, of course, using the term 'model' in a somewhat more general sense here than we did in §2.0. The sense of 'model' which is intended, in this instance,

¹⁶See [BRE 1], [HAR₁ 2], and [FOG 1].

¹⁷Different terminology is, of course, used by scientists; see, for example, [HAR₁ 2], pp. 2-3.

will become more clear in §5.). Considering these definitions, then, we may have the following result:

- (I) Only primary models are empirically significant or scientifically valuable.
- (II) All models in empirical science are secondary.
- (III) Therefore, it is not logically possible to produce a significant and valuable (scientific) model.

The rationale of this argument seems to be that primary models are immediate representations of the environment (for example, they are alleged to be immediate representations of a scientist's environment); primary models, therefore, are taken to be unmediated by neural/brain processes. The inference from this, is that the "models" used in empirical science are really "models" of the CNS's interpretation of its environment; these "models" are, therefore, secondary. Secondary models, however, not being direct representations of the environment, are not, it is presumed, scientifically significant or valuable.

A practical refutation of this argument would seem to be readily available. (I) can be denied for two reasons. First, apparently, there are no primary models; therefore, the primary/secondary distinction is not

meaningful. Secondly, the models used in science have proved to be useful; in addition, it seems to be somewhat spurious to deny scientific models "significance," assuming that it is even meaningful to speak of "significance" in a general, non-relative manner.

We contend that the primary/secondary distinction should be dropped entirely. Good evidence against the corresponding immediate/mediate distinction in epistemology has recently been presented.¹⁸ We feel that there is also good evidence against the primary/secondary model distinction.

The second consequence of the fact that a CNS "models" its own environment may be considered in terms of the Brain-Model Problem. The Brain-Model Problem involves a series of questions which fall, roughly, into two groups.

The main question of the first group is: What are the physiological processes by means of which animals (and especially human beings) are enabled to perform such activities as learning, intending, reasoning, etc.? This question is related to practical aspects of Mechanism₁; clearly, however, it is based on the presupposition that activities of the sort mentioned can, in principle, be explained by a neurophysiological theory, or by a theory in some similar science.

¹⁸See [HAL 1].

pract

are,

of wh

sciou

human

is a

littl

in th

possi

true,

Is it

to mo

enoug

such

sumin

possil

of, f

ment,

sumabl

things

of neu

fact,

model

The second group contains questions having both a practical as well as a conceptual/logical nature. There are, generally, two main questions in this group, the first of which is: How can the activities of rational, conscious organisms be modeled using systems or devices which human beings can actually construct? Here again, there is a presumption that Mechanism₂ holds. It would make little sense to ask how such systems could be constructed in the absence of an assumption that such systems are possible, in principle.

Assuming that both Mechanism₁ and Mechanism₂ are true, then there is one additional question remaining: Is it possible to construct an artefact that is sufficient to model human rational activity, but is still simple enough to be understood? It is clearly desirable that such an artefact be understandable by human beings, assuming, of course, that such an artefact is logically possible.

As we have indicated, one can consider the CNS of, for example, a human being as "modeling" its environment, i.e., as mapping its environment onto itself. Presumably (assuming the truth of Mechanism₂), among the things in the environment of an investigator in the field of neural modeling will be a neural artefact. This artefact, if it is sufficient, will be complex enough to model rational activities that are characteristic of

2

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

human brains, including those of the brain of the investigator in question. There are basically two problem areas associated with this situation.

Initially, there is a question regarding the possibility of there being a logical paradox associated with the concept of one system, for example, a human brain, constructing a model (for example, of itself) that is as complex as it is. This question will be briefly discussed, in §5, with regard to self-reproducing systems.

The remaining question is of both a conceptual and epistemological nature. At present, the human brain is not well understood with respect to its over-all functions. We may ask, then, can it be expected that a sufficient artefact, even if one were to be constructed, would be any more understandable than human brains are at present? One further question seems to be logically required at this point, namely, could an artefact which is sufficient to model a human brain understand itself any better than human beings presently understand themselves?

These question, together with some general conclusions regarding Mechanism₁ and Mechanism₂ will be discussed in §5.

§3.0

MECHANISM₁ AND THE SELF-REFUTATION ARGUMENT

Generally speaking, we take 'Mechanism₁' to be the proposition that it is logically possible to formulate a sufficient neurophysiological theory. We must indicate, however, what is meant by 'sufficient neurophysiological theory' (hereafter, 'SNT'). Following N. Malcolm, we shall say that a neurophysiological theory is sufficient if and only if it can be used to both explain and predict all movements of human bodies, with the exception of movements which are caused by forces which are external to a given body.¹ Such a theory will also describe the movements in question. We may say that Mentalism₁ is the view that an SNT is not logically possible.

There are several qualifications which should be noted immediately.

First, Mechanism₁ clearly involves a promissory note. At their present stage of development, the neurosciences have not yet developed an SNT. The Mechanist₁, however, is optimistic that an SNT is forthcoming.

¹[MAL 2], p. 334. We shall use the pagination in [KRI 1] throughout.

to adve

a candi

whether

the cas

in 14,

depende

But, th

'suffic

logical

decisio

treat i-

with res

take Me

to an S

that an

for exa

result,

quate (a

i.e., an

well as

In gener

most suc

(x) (

The lack of an SNT, in any event, does not seem to adversely affect the present study. We need not have a candidate SNT in order to discuss the question of whether or not an SNT is possible, in principle. As in the case of a sufficient neural artefact, as we shall see in §4, the development of a neurophysiological theory is dependent upon both theoretical and empirical discoveries. But, the question of whether or not an NT could be termed 'sufficient' and the question of whether or not an SNT is logically possible, must be answered on the basis of a decision, not a discovery.

Secondly, we should note that an SNT, as we shall treat it at least, does not entail a reductionist view with respect to psychology. In other words, we shall not take Mechanism₁ to be the view that psychology is reducible to an SNT (for our purposes, it is not strictly essential that an SNT is a theory of neurophysiology rather than, for example, biology, chemistry, physics, etc.). As a result, we presume that there may be more than one adequate (and independent) explanation for the same event, i.e., an event E may be adequately explained by an SNT as well as by a sufficient theory of, for example, psychology. In general, our view regarding an SNT can be characterized most succinctly by the following formula:

(x) (if x is a behavior, then SNT explains x) (3.0-1)

'E

ti

be

or

th

do

as

pr

as

al

on

is

so

Sp

th

th

br

be

log

dis

cie

ti

ate

'Behavior' is an ambiguous term which includes both 'action' and 'motion.' We shall use 'action' to refer to behavior that is purposive, goal-directed, intentional, or "done for a reason" (hence, "rational"). 'Motion,' on the other hand, will be used to refer to behavior that is done indifferently or involuntarily.² We shall, of course, assume that if an SNT explains a behavior, x, it also may predict x and describe x.

It still may not be altogether clear what counts as "sufficiency" in a neurophysiological theory. Generally speaking, we take the sufficiency of an SNT to be characterized by a predictivity criterion. This criterion is largely in accordance with the covering-law model of scientific theories and with the structural identity thesis. Specifically, it seems that an SNT must be able to explain the occurrence of any output of a human brain; an SNT, therefore, will also have to be capable of predicting human brain outputs (viz., human rational behavior). It should be emphasized, however, that many current neurophysiological theories are sufficient to explain and predict discrete neural events. An SNT, however, must be sufficient to explain and predict total brain functions and

²Malcolm also makes the 'action'/'motion' distinction. His version of the distinction will be discussed later, viz. in § 3.1.

t

t

z

w

p

o

l

s

b

x

i

—

s

w

be

p

by

de

ci

pi

ve

ru

ea

i

in

in

in

in

ar

co

fl

so

the outputs of the human brain, regarded as a total system. This requirement is rather dramatically characterized by Bertrand Russell,³ where he suggests that an SNT would have to be sufficient to permit an investigator to predict verbal or written outputs of a person on the basis of given inputs, even if the investigator did not know the language of the subject who was being used in the investigation.

In general, then, the sufficiency of an SNT is based upon a predictivity criterion which is in accordance with the usual covering-law model of scientific explanation.⁴ In order to make the criterion intuitively clear,

³See [RUS 1], p. 274. A hypothetical situation is suggested by Bertrand Russell. Suppose that a person, P, was handed a telegram which read "All your property has been destroyed in an earthquake." and after reading it, P exclaimed, "Heavens, I am ruined!" Presumably, an SNT, by taking account of (1.) the shapes in the written statement "All your property has been destroyed in an earthquake." which were in P's visual field, together with (2.) P's brain structure, etc., could effectively predict the verbal utterance, which is written as "Heavens, I am ruined!" Furthermore, this prediction could have been made, using an SNT, without knowing the English language, i.e. without knowing the meaning of the message contained in the telegram, or the meaning of the verbal response.

Accordingly, we shall say that a theory is an SNT if and only if, in the situation described, the response, "Heavens, I am ruined!" is explicable by the theory, even in the absence of a knowledge of the English language.

⁴As we shall see, there are some problems which are associated with the concept of "explanation," in this context. We shall be forced to distinguish between explanations which cite causes and those which cite reasons.

how

if

the

she

su

be

and

any

giv

giv

's

tes

or

disc

howe

however, we shall generally say that an NT is sufficient if and only if for every stimulus (input) of a human brain, the NT can be used to predict the resultant output. We shall call this criterion "Russell's Test."

We may generalize this characterization of NT sufficiency in the following manner: Let Russell's Test be generalized so that:

An empirical theory, T, passes Russell's test if and only if for every overt behavioral event E enacted by any given subject P, T can predict E. (3.0-2)

As a corollary,

if T passes Russell's Test with respect to a given E of P, then T also explains E. (3.0-3)

Additionally,

if T passes Russell's Test with respect to a given E of P, then T describes E. (3.0-4)

With these results, we may proceed to describe 'SNT.' It should be noted that, of course, Russell's Test could be applied to a theory of, for example, physics or chemistry, etc. For convenience, we shall restrict our discussion to neurophysiological theories. In principle, however, our findings should be applicable to theories in

11

11

11

11

11

11

11

11

11

11

other fields. Given Russell's Test, therefore, we shall say that,

For any neurophysiological theory (NT),

NT is sufficient

if and only if NT passes Russell's Test. (3.0-5)

We have taken Mechanism₁ to be the proposition that an SNT is logically possible. We may rephrase this characterization in the following manner:

Mechanism₁ is the proposition that it is logically possible for there to be an NT that can pass Russell's Test. (3.0-6)

Alternatively,

Mechanism₁ is the proposition that it is logically possible for there to be an NT such that NT is an SNT. (3.0-6')

Perhaps the most potentially devastating argument against Mechanism₁ is the argument that it is self-refuting. We shall, however, immediately want to distinguish between two sorts of self-refutation, namely, pragmatic self-refutation and absolute self-refutation.⁵ A proposition

⁵We follow Passmore in making this distinction. See [PAS 1], Chapter 4.

13



is said to be pragmatically self-refuting if and only if an utterance of it provides evidence for its own falsehood; for example, the spoken statement "I cannot pronounce the word 'horseradish.'" is pragmatically self-refuting. On the other hand, a proposition is absolutely self-refuting if and only if it entails a contradiction; for example, the statement "There are no truths." is absolutely self-refuting.

We shall, in this section, examine three arguments to the effect that Mechanism₁ is self-refuting. The first, the Intentional Paradox, is used to argue against Mechanism₁ from the intentional nature of assertions; the second argument, the Rational Paradox, is used to argue against Mechanism₁ on the basis of an alleged incompatibility of reasons and causes; finally, the third argument, the Description Paradox, is used to argue against Mechanism₁ on the basis of the relation between a brain description and the composition of a brain's environment.

The Intentional Paradox and the Rational Paradox seem to be based on pragmatic self-refutation arguments, while the Description Paradox seems to be based on an absolute self-refutation argument.

We shall attempt to demonstrate that Mechanism₁ is not effectively refuted by these arguments.

§3.1

THE INTENTIONAL PARADOX

In his 1968 paper, Norman Malcolm argued against the conceivability of what we have called Mechanism₁. In the final analysis, he concluded that Mechanism₁ is self-refuting and, perhaps, self-contradictory. The basis of this conclusion was formed by two paradoxes, the first of which is the Intentional Paradox (hereafter, IP).⁶

The IP may be constructed in the following manner:

- (IP-1) Mechanism₁ is incompatible with the existence of any intentional behavior.
- (IP-2) Speech (viz., the making of assertions) is a form of intentional behavior.
- (IP-3) Assume that a person P asserts that Mechanism₁ is true.
- (IP-4) The making of this assertion is incompatible with its content, i.e., the proposition:
 "Mechanism₁ is true, and someone (P) asserts
 'Mechanism₁ is true.'
 is self-contradictory.
- (IP-5) Therefore, Mechanism₁ is self-refuting.

⁶Malcolm's second paradox constitutes one formulation of the Rational paradox; it will be discussed in §3.2.

For clarification, we might add:

(IP-4') If Mechanism₁ is true, then it cannot be asserted.
 (where "Mechanism₁" is the proposition "There is
 an SNT.")

It is not immediately clear whether the IP is based upon pragmatic or absolute self-refutation. As is stated in (IP-4), Malcolm holds that the proposition "Mechanism₁ is true and someone asserts it to be true." is self-contradictory.⁷ He considers, however, the possibility that the IP may be based upon "mere" pragmatic self-refutation. For example, he notes that,

The mere proposition that mechanism is true is not self-contradictory. But, the conjunctive proposition, "Mechanism is true and someone asserts it to be true," is self-contradictory.⁸

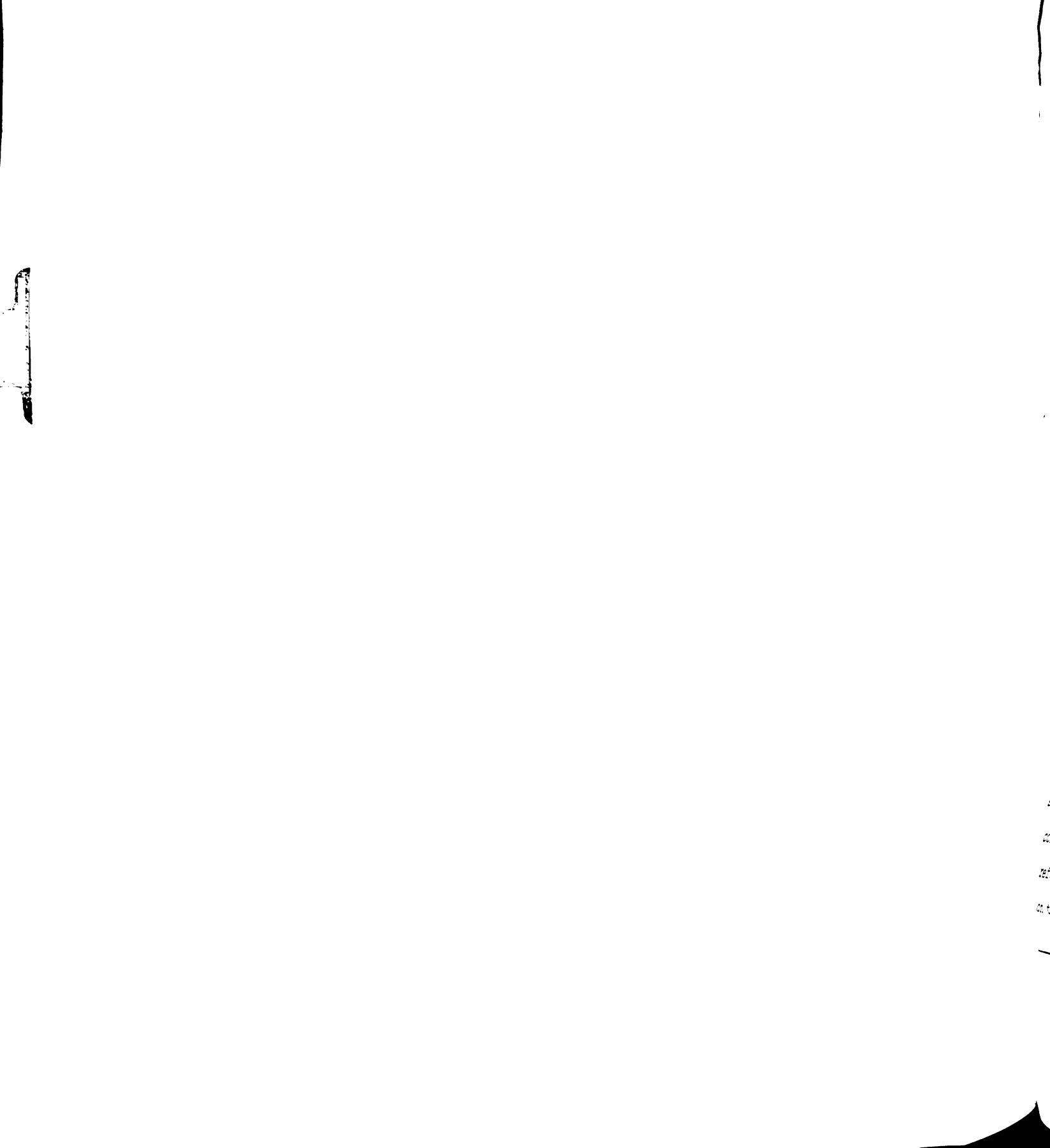
It does not seem obvious that a logical self-contradiction is present here, although substitution of suitable definitions of 'mechanism' and 'assertion' might turn it into one.

What Malcolm means, in effect, is that,

Mechanism₁ is true. (3.1-1)

⁷ [MAL 2], p. 346.

⁸ Ibid.



is pragmatically self-refuting, since, no person, P, can make a statement of the form (3.1-1) without pragmatically self-refuting himself.

Elsewhere,⁹ Malcolm describes a situation that is, in some respects, similar to this one. He asks us to consider, for example, statements of the form:

P is unconscious. (3.1-2)

Suppose that a person, P_1 , makes a statement which has the form (3.1-2), i.e., P_1 says, " P_1 is unconscious.", then, as Malcolm contends, P_1 has pragmatically refuted himself, by making that statement. Consider, on the other hand, a person other than P_1 , i.e., P_2 . P_2 can say " P_1 is unconscious." without pragmatically refuting himself. P_2 , of course, cannot state that " P_2 is unconscious." without pragmatically refuting himself. It is on this point that Malcolm takes statements of the form (3.1-1) to differ from statements which have the form (3.1-2). Malcolm contends that statements of the form (3.1-1) are self-refuting when spoken by any person, P, and he also contends that statements of the form (3.1-1) are self-refuting when spoken of any person, P. This is not true, on the other hand, of statements which have the form (3.1-2).

⁹See [MAL 1], passim. Also see [MAL 2], p. 347.

1

3
CC
St
W
W
W

The kind of pragmatic self-refutation which is used in the IP differs somewhat from the more particularized sort of pragmatic self-refutation which we characterized in §3.0. The IP is based upon what we might call "universalized pragmatic self-refutation."

As we indicated in §3.0, the statement, ' P_1 cannot pronounce 'horseradish'.' is pragmatically self-refuting, when it is spoken by P_1 . Consider, however, the statement 'No one can pronounce 'horseradish'.' As it stands, this statement is simply false, but the conjunction, 'No one can pronounce 'horseradish,' and P_1 just pronounced 'horseradish'.' is self-contradictory. In this respect, we may say that the proposition 'No one can pronounce 'horseradish'.' is universally pragmatically self-refuting.

It is, apparently, Malcolm's view that the proposition (3.1-1) is universally pragmatically self-refuting, in the sense which we have just indicated. He certainly holds that (1.) the proposition (3.1-1) is false, and (2.) conjunctive statements of the form 'Mechanism₁ is true, and P_1 asserts that Mechanism₁ is true.' are self-contradictory. Malcolm would also subscribe to the stronger contention that propositions of the form 'Mechanism₁ is true, and P_1 asserts p.', where p is any proposition whatever, are self-contradictory. Furthermore, he holds that a statement of the form 'Mechanism₁ is true,

and P_1 asserts p .' is self-contradictory, even if it is made by someone other than P_1 .

Malcolm would allow that statements which have the form 'Mechanism₁ is true, and P_1 asserts p .' could be stated without constituting self-refutation, only if they were stated by a person who was a solipsist and who took Mechanism₁ to be true of "other organisms" but not of himself.¹⁰ Elsewhere, Malcolm finds solipsism to be inconceivable.¹¹ We may conclude, therefore, that he would consider Mechanism₁ to be inconceivable in any event, even if it were asserted by a solipsist who applied it only to others and not to himself.

If Malcolm is correct, then the IP may be based upon a "universally" pragmatic self-refutation argument.

Clearly, the success of the IP is dependent upon the truth of its first premise, (IP-1), since it is argued, in the IP, that the proposition 'Mechanism₁ is true' cannot be asserted, in virtue of the intentionality of asserting anything whatever. It is Malcolm's view that Mechanism₁ "rules out the possibility of" there being any intentional activity, viz., this is what he means by saying that Mechanism₁ is "incompatible" with the existence of intentional activity. But, Malcolm also indicates that

¹⁰[MAL 2], p. 347.

¹¹See [ROO 1] and [MAL 3].

making an assertion is an intentional activity, so the inference to (IP-4) and (IP-5) is obvious. It seems apparent, therefore, that if (IP-1) could be maintained, then the IP would be successful in showing that Mechanism₁ is self-refuting, when it is asserted.

Accordingly, we shall concentrate our attention upon Malcolm's argument for (IP-1). We shall then apply our findings to the IP itself. Malcolm's argument for the proposition (IP-1) may be constructed in the following manner:

- (I) Mechanism₁ is incompatible with purposive behavior.
- (II) If there could be no purposive behavior, then there could be no intentional behavior.
- (III) Therefore, Mechanism₁ is incompatible with intentional behavior. (IP-1)

This formulation of Malcolm's argument for (IP-1) is very over-simplified. As we shall see, Malcolm's arguments for (I) and (II) are rather complex. We shall have to examine them, in some detail, in order to be in a viable position to form a conclusion regarding his proposition (IP-1) and, more generally, regarding the IP itself.

Before presenting the argument for Malcolm's proposition (II), ". . . if there could be no purposive

behavior, there could be no intentional behavior."¹² it will be helpful to note two distinctions. Firstly, while Malcolm is somewhat unclear on this point, he seems to hold that there are two ways to characterize intentions. In the first way, it might be said that to intend to do an action, B, is to do B "on purpose," i.e., voluntarily or consciously. In the second characterization, on the other hand, to intend to do B is to do B with a purpose in mind, i.e., for the sake of some purpose, or to achieve some goal. Considered in the first sense, intentions are non-purposive, or not goal-directed. For example, a person, P₁, can intentionally raise his arm (i.e., he can "voluntarily" raise his arm), while having no purpose for doing so.

Considered in the second sense, they are purposive, or goal-directed. Malcolm seems to indicate this distinction in his statement that, "A man can do something intentionally, but with no further intention." as well as at various other points in his paper.¹³

Secondly, Malcolm explicitly makes a distinction between "simple" and "further" intentions. A simple intention is merely "The intention to do X,"¹⁴ where X is

¹²[MAL 2], p. 346.

¹³[MAL 32, p. 345.

¹⁴[MAL 3], p. 342.

some action. A further intention, on the other hand, is "The intention to do something else Y in or by doing X,"¹⁵ where both Y and X are actions. This distinction is a factor in the arguments that Malcolm presents for the proposition (II), 'If there could be no purposive behavior, then there could be no intentional behavior.' as well as being a factor in the subsequent proof of the proposition, (IP-1), 'Mechanism₁ is incompatible with intentional behavior.'

The proof for the proposition (II) proceeds in seven stages, in Malcolm's paper. The first stage is:

- (1) A man can do something intentionally, but with no further intention; his behavior is intentional but not purposive.¹⁶

This proposition calls upon both of the distinctions which we have just cited. An example of such an intention would be a non-purposive, simple intention.

The second step, in Malcolm's argument, employs a form that is common to his writings:

- (2) If some intentional actions are purposeless, it does not follow that all of them could be purposeless.¹⁷

¹⁵ Ibid.

¹⁶ [MAL 2], p. 345.

¹⁷ Ibid.

Malcolm's reasoning regarding this and similar matters is sufficiently well-known that a discussion of it here seems to be unnecessary.

Proceeding, then, to the next stage in Malcolm's argument, we have,

- (3) Any physical activity is analyzable into components.¹⁸

This proposition is assumed by Malcolm.

It should be noted that "activities," for Malcolm, may include both actions and motions. This step may, perhaps, be called into question, since Malcolm has nothing to say about how the analysis to which he refers could be accomplished. It would seem to be the case that the separation of components of an activity would be likely to be somewhat arbitrary. For example, the activity of a motorist applying the brake on his automobile would seem to be an example of an activity that could be analyzed into components: e.g., the components might be, (a) raise foot off of accelerator, (b) tilt toe down, heel up, (c) move foot to the left, and (d) lower foot onto the brake pedal. There is, firstly, the difficulty of dividing a smooth activity into components; we may ask, for example, could (a) and (b), in our example, be combined? Many activities,

¹⁸Ibid.

such as the one in our example, are not easily divisible into discrete components. In this respect, the selection and the division of components of an activity would be likely to be arbitrary.

Secondly, and more importantly, however, it is not clear which of (a), (b), (c), or (d), in our example, should be regarded as being an intentional component of the activity of a motorist applying the brake; but, according to the next step, in Malcolm's argument, at least one component of every activity is intentional:

- (4) If the activity is intentional, then at least some of its components will be intentional.¹⁹

It is a trivial consequence of (4) that,

- (5) . . . an intentional activity must have intentional components.²⁰

It seems to us that Malcolm provides no criteria for distinguishing between intentional components of activities and non-intentional components of activities, with the exception, perhaps, of definitional criteria. Definitional criteria, however, do not seem to be helpful in the actual carrying out of the analysis of an activity.

¹⁹ Ibid.

²⁰ Ibid.

resu

vol

th

a

Malcolm treats (4) as a definition. As well as resulting in some practical difficulties, this also involves a vicious circularity which unfavorably affects the IP as well as Malcolm's entire theory of intentionality.

We presume that the intentional components of an activity are actions [by definition]. Actions are characterized, by Malcolm, in the following manner:

Two things are implied by an ascription of an "action" to a person: first, that a certain state of affairs came into existence . . . , second that the person intended that this state of affairs should occur.²¹

One important factor, here, is that Malcolm makes a shift from talking about the purposive behavior of "organisms" to talking about the intentional actions of "persons." The shift is badly blurred, in his paper. One result is that intentions, since they are relevant only to actions (and not to mere motions) are defined, by Malcolm, in terms of actions. Actions, on the other hand, are defined in terms of intentions, resulting in a vicious circularity.

In the next stage of Malcolm's argument, an attempt is made to definitionally link the concepts of purpose and intention:

²¹[MAL 2], pp. 340-341.

- (6) The components [of an activity] will be purposive in relation to the whole activity.²²

This step, (6), clearly means that the given activity, taken as a whole, is done for the sake of some purpose or goal, since it is an "activity" and not a mere sequence of motions. The intentional components, according to (6), are actions, all of which are done for the same purpose. In other words, the purpose of all of the respective actions which are components of a given activity, is identical to the purpose of the activity taken as a whole.

This claim does not seem to be at all obvious. It appears that, considering Malcolm's definitions, an intentional component of an activity could have its own purpose independently of the purpose of the whole activity of which it is a component. For example, a motorist might intentionally move his foot in order to make it avoid an obstacle on its way to the brake pedal. The purpose of the avoidance action might be to prevent pain or to prevent a scuff on his shoe. This example may be made more feasible if one assumes that the obstacle in question could have been forcibly knocked aside and did not have to be avoided in order to perform the general activity of applying the brake. This seems to be a situation wherein an intentional component of an activity has a degree of autonomy, in that

²²[MAL 2], p. 345.

it ma
purp
pone
pos
a

a
L
6
J
e
C
M
S
a
i
P

it may be purposive in relation to itself and may not be purposive relative to the activity of which it is a component, since its purpose is not identical with the purpose of the activity as a whole. In fact, the avoidance action, in our example, may be termed a mere "motion" relative to the activity of which it is a component.

The small avoidance action, in our example, may also be considered as being an activity in its own right, since it has an intentional component. So, [by (3)] it may be divided into further components, some of which must be intentional [by (4)]. If one of these intentional components turns out to be purposively autonomous, in the sense that we have indicated, then it too may be an activity in its own right and would, therefore, have to be analyzed into discrete components, at least one of which must be intentional. This argument may be repeated at each level of the analysis, thus generating an infinite regress.

One additional difficulty, in Malcolm's proposed analysis of activities into discrete components, results from his distinction between intentions which are purposive and those which are non-purposive. Strictly interpreted, (5) requires that at least one component of every activity be an intentional action of some sort. Such an activity could have a purpose that is identical to the purpose of the activity of which it is a component only if

it co

own a

But

com

1

1

1

1

1

1

1

1

1

1

1

1

1

1

it constituted a purposive intentional action. By Malcolm's own account, however, some intentions are non-purposive. But a sequence of movements need only have one intentional component in order to be termed an activity. We may, therefore, draw two consequences. First, if the only intentional component of an activity is a non-purposive intentional action, then the activity would also be non-purposive. This contradicts Malcolm's stipulation that all activities are purposive. Secondly, suppose that a given activity does have a purpose. Suppose, further, that at least one of its intentional components is a non-purposive intentional action. Then, it follows that the given action could not have a purpose which is identical to the purpose of the activity of which it is a member, since it does not have a purpose at all; this result also contradicts (6).

We conclude, therefore, that Malcolm has not been successful in showing that the components of any given activity will be purposive in relation to the whole activity. We have shown three reasons for adopting this conclusion. First, the division of components of an activity is likely to be arbitrary; secondly, the "division of components" for the purpose of locating the intentional components of an activity leads to an infinite regress. Finally, we have seen that components of activities can have autonomous purposes, which are unrelated to the purpose of the larger activities of which they are members, or components of activities can be non-purposive.

posi

(7)

The conclusion of Malcolm's argument is the proposition, (II), which has been our present concern:

- (7) . . . if there could be no purposive behavior, there could be no intentional behavior.²³

Clearly, if we do not grant Malcolm step (6) in his argument, then the conclusion does not follow. It should be noted, additionally, that Malcolm uses 'intentional behavior' here. Given his own definition, the term 'behavior' is, at best, vague, and 'intentional behavior' may be inconsistent, since, for Malcolm, only actions (or "activities") are intentional.

In summary, it seems warranted for us to conclude that Malcolm's argument for (7), is unsound.

The next phase of our discussion is concerned with Malcolm's proposition ((III) = (IP-1)) that Mechanism is incompatible with intentional behavior.

Before entering into this discussion, however, note should be taken of why it is important to Malcolm that Mechanism₁ be incompatible with intentional behavior. He mentions at one point, that the two forms of "explanation," purposive explanation and neurophysiological explanation, are irrelevant to one another; his reasoning, here, is that they belong to different "language games": they make use of different concepts.²⁴ The possibility

²⁴See [MAL 2], p. 338.

rema:

plan

comp

be

y

remains, however, that even though the two forms of explanation are not relevant to one another, they may be compatible, i.e., an "explanation" of either form could be offered for one and the same sequence of movements. Malcolm hesitates to permit this sort of compatibility, however, because his aim is to show that Mechanism₁ rules out the possibility of intentional "explanations" of events. If he were successful in doing this, he would hope to infer that since we do "explain" certain events with the use of intentional "explanations," Mechanism₁ cannot be true. We shall discuss the alleged divergence between the two forms of explanation, PE and NE, in more detail in section 3.2.

The argument which is used by Malcolm to substantiate the proposition (IP-1) that Mechanism₁ is incompatible with intentional behavior begins, as we have indicated, with the conclusion, (7) (= (II)), of the last argument which we considered. We shall continue to number the steps, in sequence (beginning with (8)), in order to preserve continuity.

For the sake of discussion, we shall permit the use of (7) in order to examine the direction in which the remainder of Malcolm's argumentation proceeds.

While taking (7) as a (previously proved) premise, Malcolm adds a second premise, in his argument for (IP-1):

(8) b

Earl

is

?

A

2

W

7

4

2

7

7

2

- (8) . . . mechanism is incompatible with purposive behavior . . .²⁵

Earlier in his 1968 paper, he argues for this point. This is the central argument, for Malcolm, to the effect that purposive explanation and neurophysiological explanations are incompatible. To attempt to make his point, he uses an example, rather than a rigorous argument; he asks us, in particular, to consider a man who intentionally climbs up a ladder to a roof for the purpose of retrieving his hat. Suppose that this event were to be explained by an SNT. Then, Malcolm holds that,

. . . the movement of the man on the ladder would be completely accounted for in terms of electrical, chemical, and mechanical processes in his body. This would surely imply that his desire or intention to retrieve his hat had nothing to do with his movement up the ladder.²⁶

Clearly, however, the implication that Malcolm wishes to make, here, follows only if he presumes, in advance, that Mechanism₁ is incompatible with intentional behavior. This is the point upon which the argumentation of his entire 1968 paper turns. We find that Malcolm makes the assumption in question while the only thing that he provides as a justification for the required incompatibility presumption is the brief comment that the implication

²⁵[MAL 2], p. 346.

²⁶[MAL 2], p. 338. The italics are Malcolm's.

argt

conc.

He me

"would surely" follow. We must conclude, then, that his argument for (8) is invalid, or, more correctly, we should conclude that he has no argument for (8) at all.

Malcolm seems to recognize that this is the case. He mentions in a postscript to his 1968 paper that,

I must confess that I am not entirely convinced of the position I have taken in respect to the crux of this paper--namely, the problem of whether it is possible for there to be both a complete neurophysiological explanation and also a complete purposive explanation of one and the same sequence of movements. I do not believe I have really proved this to be impossible.²⁷

As we have concluded, he has not been successful in showing this to be impossible.

It should also be mentioned that, in a short restatement of his "proof" for (8) in his 1968 paper, Malcolm calls upon a concept of "explanation." He argues that if an SNT were to be taken to completely "explain" the event in question, i.e., the man climbing the ladder to get his hat, the mention of a man's intentions would not "explain" (or be a part of an "explanation" of) the event. Here Malcolm plays upon an ambiguity of the word 'explain.'

In one sense of the term, to "explain" an event E, when, e.g., E is a behavioral event, is to give a causal account of the occurrence of E; in another sense,

²⁷[MAL 2], p. 349.

1

to "explain" E is to justify the performance of E (by some person, P).

When one recognizes these two divergent senses of this term, then it seems clear that there is no inconsistency in saying that an SNT (causally) explains an activity while a mention of intentions justifies it. We shall discuss the ambiguity of 'explain' in more detail in section 3.2.

We shall continue with Malcolm's argumentation, in order to bring our recent findings into perspective with the central concern of the present study. Malcolm's next step, in his argument for (IP-1) is:

- (9) . . . mechanism is incompatible with intentional activities.²⁸

Malcolm expects this step, (9), to follow from (7) and (8). Aside from other difficulties which we have already mentioned related to (7) and (8), it should be noted that Malcolm has once again made an innocuous, but possibly defective shift in terminology. (7) and (8) speak only of "behavior," while (9) involves a shift in terminology to the term 'activities.' If the motion/action distinction were rigorously observed, this small defect could likely be corrected.

²⁸[MAL 2], p. 346.

Malcolm's final conclusion is:

- (10) [Therefore] . . . [Mechanism₁] is incompatible with all intentional behavior.²⁹₁

Clearly, (10) is equivalent to (IP-1). Note, however, that Malcolm has again made a shift to 'behavior,' in (10).

The effect of our findings upon the IP should now be clear. It is evident that the IP depends upon its first premise, (IP-1), for its success. As we indicated earlier, Malcolm attempts to support the proposition, (IP-1), by using the proposition (II), 'If there could be no purposive behavior, there could be no intentional behavior.' and (I), 'Mechanism₁ is incompatible with purposive behavior.'. If Malcolm's defense of (IP-1) had been successful, then the IP itself would have been successful.

We have found that there are two points which are critical to the success of the IP. The first point is Malcolm's defense of steps (4), (5), and (6) in the argument which we considered. We have found these steps to be defective, and we have found that (7) is not well-founded. Without (7), of course, the remainder of Malcolm's argument is not sound.

The second point that is critical to the success of the IP, in any event, is Malcolm's non-defense of (8).

²⁹ Ibid.

We have found (8) to be an unwarranted, unproved proposition; Malcolm has, himself, expressed doubts about it. Without (8), however, (10) cannot be established. Since (10) is the first premise of the IP, viz. (IP-1), and since (IP-1) forms the basis of the entire IP, we must conclude that the IP fails.

Hence, considering that the IP is unsuccessful, we conclude that Mechanism₁ has not been shown to be self-refuting. Consequently, we conclude that an SNT still appears to be logically possible.

§3.2

THE RATIONAL PARADOX

The Rational Paradox (RP) is based upon an alleged incompatibility of rationality and causality. Specifically, if the RP is sound, then Mechanism₁ is self-refuting, since, according to the RP, if there is an SNT, then there can be no reason for believing it.

In what is probably its most simple form, the RP may be constructed as follows:

- (RP-1) Assume that reasons can be given for believing Mechanism₁ (or, assume that a (rational) proof can be cited, the conclusion of which is a proposition expressing Mechanism₁).
- (RP-2) Then, Mechanism₁ can be held for a reason.
- (RP-3) But, if Mechanism₁ is true, then reasoning (and inference) is a strictly causally determined process which occurs in human brains.
- (RP-4) Therefore, reasoning (and inference) is a non-rational process.
- (RP-5) Hence, belief in Mechanism₁ is arrived at by causal, non-rational processes.

The conclusion may be restated as:

(RP-5') Consequently, if Mechanism₁ is true, there is no reason to believe that it is true.

Alternatively, the conclusion may also be stated as:

(RP-5'') Mechanism₁ both can and cannot be held for a reason.

The form of self-refutation argument which is used in the RP has a long history. It is usually used to combat theories that are similar, in certain respects, to Mechanism₁, the essential similarity being the relevance of doctrines such as Mechanism₁ to a certain set of issues in the philosophy of mind. Recently this form of self-refutation argument has been used to criticize materialism (as the view is related to the philosophy of mind), determinism,³⁰ and versions of mechanism that are similar to our Mechanism₁.³¹ The RP which we are presently considering, has been extrapolated from these recent arguments. In logical form, the RP closely resembles the self-refutation argument in J. N. Jordan's "Determinism's

³⁰See [JOR 1], passim.

³¹See "Townes' Paradox" in [TOU 5], p. 5.

Dilemma."³² (See Appendix I(A)). In terminology and intent, the RP closely follows "Townes' Paradox" which is presented in S. Toulmin's recent paper, "Reasons and Causes."³³ (See Appendix I(C)). In some respects, the RP also resembles Malcolm's Rational Paradox³⁴ (see Appendix I(B)), although, Malcolm's RP is based upon considerations regarding intentionality, which are the same as those which are discussed in §3.1.

The argument of the RP, as it stands, is clearly fallacious. (RP-4) is an unwarranted inference unless it is presumed that reasons are incompatible with causes. It seems, then that the RP has a suppressed premise that may be expressed as:

(RP-4') Causality is incompatible with rationality.

The defense of (RP-4'), however, is seriously hampered by the fact that the terms 'cause' and 'reason' may be used in various ways. Questions regarding causes, such as, for example, the question, "What caused P to do A?," can result in quite different sorts of replies, depending upon the circumstances in which they are asked. The

³²See [JOR 1], passim.

³³[TOU 5]. See especially pp. 5 and 23-34.

³⁴See [MAL 2], p. 348.

question, "What caused P to do A?" could illicit a response in which P's reasons for doing A were cited, while in other circumstances, physical conditions or empirical laws could be mentioned as being "causes" for P's doing A.

It seems to us that, in the final analysis, the RP is not based simply upon an incompatibility of causality and rationality, but rather it is based upon an alleged incompatibility of giving a causal account of an event, and giving a "rational" account of the same event. So, it appears that the RP is based upon an alleged incompatibility of causal explanations and rational "explanations."

Accordingly, we may further clarify the suppressed premise of the RP by reformulating it as follows:

(RP-4'') Causal explanations are incompatible with rational explanations.

It seems to be clear that, if (RP-4'') is not equivalent to (RP-4'), it is at least implied by it. Hence, if we can refute (RP-4''), i.e., if we can prove that it is not the case that (RP-4''), then we can obtain a denial of (RP-4'), by modus tollens.

This alleged incompatibility between causal explanation and rational explanation is, as we have indicated, not proved in the RP, and, for this reason, the RP is invalid. We must, however, consider the possibility that (RP-4'') is true even if it is not proved in the RP.

It is our contention that (RP-4'') is false. This contention can be most readily defended by pointing out that (RP-4'') could be plausibly maintained only if 'explanation' was being used unambiguously in (RP-4''), i.e., if both occurrences of the term 'explanation' were being used in the same sense. Close scrutiny reveals that 'explanation' is ambiguous and that Mentalists₁ play upon the ambiguity in the RP, and specifically in [RP-4'').

One can be said to "explain" a behavioral event E in two senses. In one sense, to "explain" E is to give a causal account of it. Presumably if an adequate causal account of E were to be given, it would be a covering-law explanation, and it could be used to explain events which are similar to E, and it could have been used to predict E. In another sense, however, to "explain" E is to justify E or to give an indication of one's intentions in performing E. In this sense, one does not seek to be able to infer E from a set of laws and a set of statements of conditions, nor does one seek to predict E. Rather, one seeks, generally speaking, to "justify" the performance of E and thereby attempt to illicit praise and dispel blame.

It appears, therefore, that 'explain' may be used in two divergent senses. In one sense, to "explain" an event E, is to causally explain the occurrence of E. In a second sense, to "explain" an event E is to justify one's

doing of E. Considering this ambiguity, we may once again reformulate the suppressed premise of the RP as:

(RP-4''') Causal explanations are incompatible with
rational justifications.

It is, specifically, this proposition, (RP-4'''), that we claim to be false.

(RP-4''') could be successfully defended (as being true) only if it could be demonstrated that a causal explanation of an event E (in Malcolm's words) "rules out the possibility" of there being a rational/purposive justification of E. So, we must ask if this can be demonstrated.

One approach, in the attempt to demonstrate that causal explanations rule out the possibility of there being rational justifications, would seem to be to argue that there are not two senses of 'explanation,' but rather there are two "kinds" of explanation, and these two kinds of explanation are incompatible. Malcolm takes this approach with regard to purposive explanations. His view is that purposive explanations are a subset of rational explanations, since one way in which one can rationally explain one's actions is to indicate one's purpose or intention in performing one's actions. For Malcolm, as we have seen, purposive explanations explain actions, whereas,

neurophysiological explanations explain movements. Both explain behavior. Malcolm, however, notes that,

. . . we can say this only because we use the latter word [i.e., 'behavior'] ambiguously to cover both actions and movements.³⁵

To generalize, Malcolm subscribes to the proposition:

Purposive explanations differ in kind from neurophysiological explanations. (3.2-1)

As Malcolm states this contention, in one paper,

If thought is identical with a brain process, it does not follow that to explain the occurrence of the brain process is to explain the occurrence of the thought. And, in fact, an explanation of the one differs in kind from an explanation of the other. The explanation of why someone thought such and such involves different assumptions and principles and is guided by different interests than is an explanation of why this or that process occurred in his brain. These explanations belong to different systems of explanation.³⁶

Elsewhere, he clarifies what he means by different "systems" of explanation; Malcolm holds, to be specific, that any neurophysiological explanation (NE) is of the following form:

(NE-1) "Whenever an organism of structure S in neurophysiological state q, it will emit movement m.

³⁵[MAL 2], p. 338.

³⁶[MAL 4], p. 179. The italics are Malcolm's.

(NE-2) Organism 0 of structure S was in Neurophysiological state q.

(NE-3) Therefore, 0 emitted m.³⁷

Any purposive explanation (PE) is considered, by Malcolm, to be of the following form:

(PE-1) Whenever an organism 0 has goal G and believes that behavior B is required to bring about G. 0 will emit B.

(PE-2) 0 had G and believed B was required of G.

(PE-3) Therefore, 0 emitted B.³⁸

Potential difficulties regarding tense shifts, in both NE and PE, could be easily avoided by a simple restatement of both of the explanation forms in terms of an arbitrary time "t." Hence, we shall not comment on difficulties regarding tense.

Other difficulties, however, remain. In particular, it does not follow, from there being two kinds of events, that there are two kinds of explanations. Specifically, Malcolm argues that, where 'PE' and 'NE' are defined as above,

(1.) PE's explains actions.

³⁷[MAL 2], p. 335.

³⁸Ibid.

- (2.) ME's explain motions.
- (3.) A's differ in kind from M's.
- (4.) Therefore, PE's differ in kind from NE's.

This argument form can be shown to be not valid.

We may generalize the argument form as:

- (1.) E_1 explains X.
- (2.) E_2 explains Y.
- (3.) X's differ in kind from Y's.
- (4.) Therefore, E_1 's differ in kind from E_2 's.

where E_1 and E_2 are explanations and X and Y are events. Specifically, let X designate the rising of the ocean tides, and let Y designate the revolution of the moon around the earth. Also, let E_1 be a Newtonian explanation of X, and let E_2 be a Newtonian explanation of Y. Then, certainly it is clear that X and Y can be said to differ in kind, but it cannot be said that E_1 and E_2 differ in kind. So we may conclude that Malcolm cannot validly infer that PE's differ in kind from NE's from the fact that they are used to explain different kinds of events.

Malcolm may still be able, however, to demonstrate that PE's differ in kind from NE's by some other means. Such a demonstration, however, will have to contend with the views of E. Nagel, M. Schlick, C. Hempel and others; these methodologists have convincingly argued that

teleological and functional explanations do not differ in kind from other covering-law explanations.

Malcolm argues that the PE differs from the NE in three essential respects.³⁹ First, purposive explanations are ordinarily not organized into theories whereas neurophysiological explanations are. Secondly, purposive explanations employ the concept of "purpose or intention" whereas Neurophysiological explanations do not. Finally, purposive explanations do not make use of contingent, causal laws whereas neurophysiological explanations do.

The first difference does not seem to be a crucial one. We may argue that while ordinarily PE's are not axiomatized, there does not seem to be a reason for saying that they cannot, in principle, be so organized. Malcolm's construction of a formal PE seems to support this contention.

Regarding the third difference, Malcolm holds that specification of an "implicitly understood" clause regarding the absence of "countervailing circumstances" makes (PE-1) an "a priori" principle, while (NE-1) is, at best, a contingent, causal law.⁴⁰ This, presumably, is one basic reason for holding a PE is different "in kind" from an NE. Malcolm's terminology, on this point, however, is open to suspicion. For example, he regards (PE-1) as being

³⁹See [MAL 2], p. 335.

⁴⁰See [MAL 2], pp. 335-337.

"a priori" rather than using the, in some respects, stronger term 'necessary.' Yet, he contrasts (NE-1) with (PE-1) by noting that (NE-1) is "contingent," rather than using the weaker term 'a posteriori.' If one grants any merit to W. V. O. Quine's remarks regarding this entire set of distinctions,⁴¹ then Malcolm's argument, in this instance, is doubly suspect.

In any event, it seems unlikely that the introduction of a provision regarding countervailing circumstances could adequately distinguish between the PE and the NE. It seems apparent that the NE could likewise be supplemented with an appropriate ceterus parabus clause which would seem to put (NE-1) and (PE-1) on the same footing.

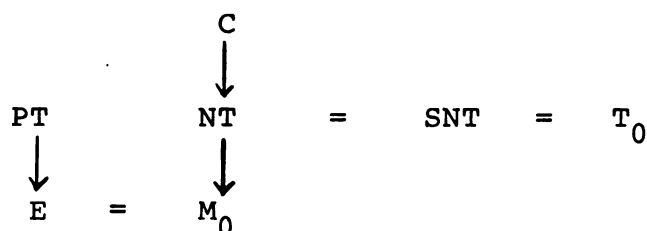
It seems clear that, for Malcolm, the major difference between the PE and the NE is that the PE uses the concepts of "purpose" and "intention" whereas the NE does not.

It is apparent that, for Malcolm, an explanation of purposive/intentional activity is emergent with respect to a neurophysiological explanation. This aspect of Malcolm's viewpoint is, in fact, generalized for any instance of PE, not just those involving uniquely human purposive/intentional activities. Specifically, note that the PE is worded in terms of "organisms" in general and is thereby not restricted to only explanations of human activities.

⁴¹See [QUI 1], passim.

We take reductionism, and correspondingly, emergentism, to be characterized in terms of certain formal relationships between theories. We presume that a PE or an NE would include the laws of some respective theories. In the case of a PE, the laws of a purposive/psychological theory (PT) would be used, which employs concepts such as "intention," "purpose," "rationality," "intelligence," etc. Similarly, an NE would be based upon the laws of a neurophysiological theory. Hence, a PT would be reduced to an NT if and only if all of the primitive terms of PT were definable by means of the primitive terms of NT and the theorems of PT were derivable from the axioms of NT. Malcolm, however, holds that no PT is reducible to any NT. Specifically, he indicates that terms such as 'purpose' or 'intention' could not be defined by means of the terminology of any NT. It follows, then, that some of the vocabulary of any PT is emergent with respect to any NT.

We are in essential agreement with Malcolm on this point. We differ with him, however, in holding that both a PE and an NE could explain the same event, and explain it sufficiently. Where the event in question is E, we have the result that PT and NT may be satisfied by the same relational system and that E is an (output) state (or internal state) of that relational system. Hence,



(Figure 3.2-1)

So, our position is that PE and NE could independently and sufficiently "explain" the same event, while PT is emergent with respect to NT.

Hence, in view of our observations that the term 'explanation' is ambiguous, and in view of our observation that a rational, purposive explanation (or, more, properly, "justification") will be emergent with regard to an SNT, we conclude that an SNT, which causally explains a given event E, will not rule out the possibility of a rational justification of the same event E. Therefore, we conclude that an SNT is not incompatible with the existence of rational "justifications;" consequently, we conclude that the proposition (RP-4''') is false.

Since (RP-4''') has been found to be false, it is apparent that the RP cannot be maintained. To this extent, we conclude that Mechanism₁ has not been shown to be self-refuting, and hence is still viable.

It should be recalled that Mechanism₁ is really a quite weak claim, namely that every human bodily motion is (causally explicable. As we have attempted to

demonstrate, Mechanism₁ does not rule out the possibility that some of these motions may also be rationally justifiable. Therefore, there may be two "explanations" for the same set of movements, one a causal SNT and the other a (rational, intentional, or purposive) justification.

§3.3

THE DESCRIPTION PARADOX

One of the most frequently mentioned distinctions between social science and natural science is that the laws of social science may become either invalid or valid if they become a matter of public knowledge. E. Nagel suggests two sorts of instances in which this might be the case.⁴²

First, a prediction in the social sciences, which is valid at the time at which it is proposed, may be falsified at a future time, if future actions are made by agents who have knowledge of the social science theory which is used in making the prediction. Nagel provides the example of the 1947 recession which was predicted by economists, but failed to materialize because of the common knowledge of the prediction.

Secondly, there are self-fulfilling prophecies which might occur in the social sciences. A prediction which is not valid at the time at which it is proposed may be made valid, in a particular instance, if actions of agents who have knowledge of the prediction make it

⁴²See [NAG 2], pp. 466-473.

valid. For example, the United States Bank, as Nagel notes, was in no financial trouble in 1928, but trouble was falsely predicted. Depositors reacted on the basis of the false prediction, thus making it true.

It may be the case, that an SNT is in a similar situation to a theory in the social sciences in that it may become invalid, if it becomes known. The features of an SNT that might make it invalid if known are, however, somewhat different from the features of social science theories that have this result. Theories, as well as being used in explanations and predictions of events, may also be used to describe them. Also, as we indicated in §2, a brain can be regarded as "modeling" its own environment. If an SNT were in existence, it would constitute a portion of the environment of some brain, and hence, would be "modeled" by some brain. This situation prompts the Description Paradox (DP) which may be formulated in the following manner:

Let 'BS' abbreviate 'Brain state' and, accordingly, let 'BS_n' abbreviate 'Brain state n', where BS_n occurs at some specifiable time t. Then, we have:

- (DP-1) Mechanism₁ is true. [Assumption]
- (DP-2) Then, it is logically possible for there to be an SNT, such that for every BS_n, SNT can be used to explain BS_n.

- (DP-3) Let B_1 be a brain which is in BS_n , and assume that SNT explains B_1S_n .
- (DP-4) Hence, SNT also described B_1S_n .
- (DP-5) But, a brain which has received a description of itself cannot be in the brain state described.⁴⁵
- (DP-6) Assume that B_1 has received SNT.
- (DP-7) Then, it is not the case that SNT describes B_1S_n .
[From (DP-5) and (DP-6)]
- (DP-8) SNT describes B_1S_n and it is not the case that SNT describes B_1S_n . [From (DP-4) and (DP-7)]
- (DP-9) Therefore, not (DP-1), i.e., Mechanism₁ is false.

Clearly, the conclusion is supposed to follow by indirect argument.

It should be immediately noted that, in our weak version of mechanism, namely, Mechanism₁, we supposed that an SNT would have to pass Russell's Test with respect to overt bodily movements. The motion in question would presumably be the final stage of a causal chain having as one of its members a corresponding brain state. So, for clarification, and to bring the DP into perspective with our limited conception of Mechanism₁, we should amend the DP with:

⁴⁵For a detailed discussion of this point, see [MAC 1], p. 395.

(DP-3') SNT predicts E_n , where E_n is an overt bodily motion, and E_n is the last member of the causal chain of which B_1S_n is a member.

If (DP-3') is accepted, it seems trivial that (DP-3).

One initial approach to the DP might be taken in terms of chronological factors. Let 'D' abbreviate 'Self-descriptive data provided by an SNT.' Suppose that, at a given time t , SNT describes B_1S_n . Then, at $t + 1$, B_1 receives D, which results in B_1S_{n+1} . Hence, SNT does not describe B_1S_{n+1} . So, suppose that SNT is up-dated to form SNT'. Then, at $t + 1$, we have SNT' describes B_1S_{n+1} . But, suppose that, at $t + 2$, B_1 receives D'. Then, B_1 's reception of D' results in B_1S_{n+2} . Hence, SNT' does not describe B_1S_{n+2} . We can again up-date SNT' to form SNT'', but this seems to result in an infinite regress.

As we see it, the central question is whether B_1 could be in a given brain state, B_1S_n , and receive D at the same time. There are three things to consider here: first, D will undoubtedly be quite complex; secondly, if an NT is sufficient, and it explained a given B_iS_n , then it could predict B_iS_{n+1} . So, we must ask whether a given B_i could "receive" a given D within a single B_iS_n . Thirdly, we must ask whether a given B_i could keep up with an SNT in generating the series SNT', SNT'', SNT''', . . . since, for any B_iS_n , an SNT could predict B_iS_{n+1} . These

questions are relevant to §4 and §5, as well as being relevant to the present section.

D. M. MacKay argues⁴⁶ that an SNT can describe a B_1S_n only if B_1 does not receive D. We shall argue that B_1 cannot "receive" D.

First, it should be emphasized what constitutes D. Assuming, for the sake of simplicity, that SNT conforms to the D-N model of explanation,⁴⁷ the explanans of SNT will consist of a set of law-like generalizations, together with a set of explicitly specified conditions, the conjunction of which effectively describes B_1S_n . The explanandum of SNT would be a proposition which describes B_1S_n (as we have indicated, an SNT, in the extended sense, would have an overt bodily motion as the object of its explanandum).

As we shall demonstrate in §5, it is questionable whether a sufficient neural artefact can be constructed that is simple enough to be understood. Correspondingly, the theoretical model which a sufficient artefact satisfies would be of such a complexity that it may not be understandable, if by 'understandable' one means that the theoretical model could constitute the substance of a single thought (viz. a single B_1S_n). Clearly, an SNT would be as

⁴⁶See [MAC 1], especially p. 395. MacKay's terminology, it will be noticed, is quite different from ours.

⁴⁷See [HEM 1], p. 336.

complex, structurally, as its theoretical model. So, correspondingly, it is questionable whether an SNT (considered, e.g., as the description of a given $B_i S_n$) could constitute the substance of a single thought (viz., a single $B_i S_n$). If this could not be the case, then a B_i could not "receive" a relevant D in the sense that is required by the DP. This general problem will be discussed in detail in §5; our present argument against the DP will, accordingly, be completed in that section. For the present, however, a few brief comments should suffice.

Let $L_1, L_2, \dots, L_n$ be the laws of an SNT, and let $C_1, C_2, \dots, C_i$ be the statements of conditions required to deduce a statement describing $B_i S_n$. We presume that B_i could know or "receive" $L_1, L_2, \dots, L_n$. Hence, a person could have knowledge of the principles of his own neural operations. What the Brain-Model Problem questions, with regard to an SNT, is whether or not B_i could know or "receive" $C_1, C_2, \dots, C_i$ within a single $B_i S_n$.

Therefore, we deny that B_i could "receive" D, where D is a total description of $B_i S_n$ of the sort indicated, and where D is "received" by B_i within a single $B_i S_n$. This conclusion constitutes a denial of (DP-6). Without (DP-6), however, the DP fails. Consequently, if we are correct, the DP is not successful in refuting Mechanism₁.

§3.4

DISCUSSION

We have attempted to demonstrate, in this section, that Mechanism₁ is not self-refuting. In order to do this, we have considered three arguments to the contrary.

The Intentional Paradox was shown to be defective, because it is based upon the unproved and illicit assumption that Mechanism₁ is incompatible with the existence of intentional behavior.

The Rational Paradox is mistakenly based upon the assumption that causality and rationality are incompatible. We also demonstrated that both the Intentional Paradox and the Rational Paradox fail to take account of the distinction between 'justification' and 'explanation.' Considering the ambiguity of 'explanation,' confusions seem to be inevitable.

Finally, the Description Paradox is based upon the apparently fallacious assumption that an SNT, considered as a total description of a given brain state, can be "received" by a given brain, from moment to moment and from brain state to brain state. The argument which we have presented against the Description Paradox will be completed in §5.

Since Mechanism₁ has not been shown to be self-refuting by the arguments which we have considered, we may conclude that, to this extent, at least, an SNT is logically possible.

Granted this conclusion, we may proceed to consider the question of whether a sufficient neural artefact is logically possible.

MECHANISM₂ AND THE GÖDEL ARGUMENT

Gilbert Ryle mentioned, at one point, ". . . a tautology which is sometimes worth remembering." namely, "Men are not machines, . . . They are men" ¹ Certainly, it is clear that the proposition 'men are men' is a tautology, as is 'machines are machines.' It is not clear, however, that 'men are not machines' is a tautology, and it is even less clear that (as is often inferred from Ryle's tautology) 'men cannot be modeled (e.g., by "machines")' is a tautology. Since tautologies imply only tautologies, the inference from 'men are men' to 'men cannot be modeled' seems to be immediately suspect.

Probably the strongest non-metaphysical argument that has been presented in favor of this inference, and against Mechanism₂, is the argument based upon Proposition VI of Gödel (1931). ²

As was the case with Mechanism₁, we somewhat restrict the general doctrine of mechanism to Mechanism₂.

¹[RYL 1], p. 81.

²See [GÖD 1], p. 57.

For our purposes, we take Mechanism₂ to be the proposition that it is logically possible to construct a sufficient (physical) neural model (SNM), viz., an artefact that is sufficient to simulate all characteristically rational human brain functions. If there were an SNM, then, it would, presumably, be able to perform the verbal, mathematical, and other "rational" behaviorisms that are usually associated with rational human beings.

What the Gödel Argument allegedly proves is that there is at least one thing that a human brain can do (aside from strictly biological functions) that no artefact can do. The conclusion which is drawn is that, while there may be an incomplete (physical) neural model, a sufficient complete neural model is not logically possible. We shall examine this argument in the present section.

We shall begin by discussing neural artefact sufficiency. Sufficiency in an SNM has generally been regarded as being a more difficult concept than sufficiency in an SNT, but there seems to be no inherent reason why this should be the case. We shall, then, present the Mentalist₂, Gödel Argument against Mechanism₂, while citing some of the variations which have been made on the central argument. We shall, at that point, raise some objections to the Gödel Argument. Finally, we draw some general conclusions which are based upon the objections raised.

§4.1

NEURAL ARTEFACT SUFFICIENCY AND THE CONCEPT OF "MACHINE"

The problem of establishing sufficiency criteria for a candidate SNM has received considerable attention. It is not at all surprising that this should be the case, since these criteria should closely parallel the criteria for human intelligence. A definitive account of human intelligence has not yet been produced, and accordingly, there are no universally accepted criteria for "machine" intelligence. Investigations of this dual criteria problem would very likely have ramifications in both psychology and in the philosophy of mind.

For our purposes, we shall treat the concept of "intelligence" in a rather cavalier manner, as seems to be the common practice in the sciences. We shall take (human) 'intelligence' to be operationally defined on the basis of stimulus-response correlations that are characteristic of normal, adult humans.

Unfortunately, there is, at present, no such thing as a "normal adult artefact," so criteria that are less intuitively satisfying must be used to provide an account of what an intelligent artefact might be like. (It should,

perhaps, be mentioned that some investigators in the field of neural modeling, are involved with evolutionary models, e.g., as is M. Minsky; these models are used, at least initially, to simulate "immature" humans). We, of course, must be on guard against an account of artefact intelligence which would definitionally rule out the possibility of there being an SNM.

Probably the most famous criterion for an (intelligent) SNM is Turing's Test.³ Turing suggests a hypothetical situation in which a person, viz., an "interrogator," is placed in an enclosed room. In an adjoining room, there may be either another person or a "machine" (i.e., a candidate SNM). Communication between the two rooms is facilitated by two console typewriters. The interrogator is challenged to determine, by typing out questions and receiving printed out replies, whether the replies which he receives are made by a person or by a "machine." If the interrogator is unable to distinguish between human-answers and machine-answers, then the "machine" is said to have passed the test, and in our terminology, constitutes an SNM.

More precisely, we shall say that,

where 'NM' =_{df} 'neural model' and 'SNM' =_{df} 'Sufficient neural model,' we have:

³See [TUR 1], passim. See especially pp. 4-5. (We shall use the pagination in [AND 1] throughout.)

For every NM, NM is an SNM if and only if NM
passes Turing's Test. (4.1-1)

We shall, shortly, restrict Turing's Test situation to
strictly computational matters.

The mechanics of the testing situation, viz., the
room, the typewriters, etc. are non-essential. As far as
intellectual capabilities are concerned, it does not mat-
ter if the "machine" is in humanoid form⁴ viz., is an
android, or if it has an appearance that is similar to
that of a present-day digital computer. For Turing, if a
"machine," regardless of its physical appearance or mode
of communication, were able to pass for a human being
(any human being), that is to say, if it could pass Tur-
ing's Test, then it could be considered to be an SNM.
Turing, in fact, suggests that a question such as "Can a
machine pass Turing's Test?" is more coherent than the
question "Can machines think?" and, hence, should replace
it.⁵

Turing's Test, however, does have some defects.⁶
One of the initial difficulties is the presence of a

⁴A Wittgensteinian might disagree: see [WIT 1],
§§ 281-283.

⁵See [TUR 1], pp. 5-6.

⁶A critical commentary of Turing's Test may be
found in [GUN 1]. A somewhat qualified defense may be
found in [SCR 1].

confusion which results from an inadequate specification of what it is that an SNM models. As with the SNT, we somewhat restrict what would be minimally required of an SNM. Specifically, it seems enough that an SNM be able to simulate the verbal/written behavior that is characteristically generated by normal human brains. This is simply another way of saying that the input-output correlation in an SNM should correspond to the input-output correlation of a normal human brain with respect to verbal/written responses.

Some of the most often used criticisms of this and similar criteria are to the effect that an SNM could not be expected to exhibit certain esthetic sensitivities, e.g., an SNM could allegedly neither write nor appreciate Poetry. Of course, there are many people who can neither write nor appreciate poetry. Hence, it does not seem that the charge that artefacts cannot write or appreciate Poetry adversely affects the possibility of an SNM. Minimally (at least according to Turing's Test), an SNM must be sufficient to model some person. An "uncultured" SNM may model only the CNS of uncultured people, but it seems clear that an "uncultured" SNM is still an SNM, since it could pass Turing's Test. Also, recent findings seem to indicate that this charge is ill-founded, but human intellectual behavior that might be termed "esthetic" or that would fall under various similar designations, need not



be a part of our present discussion. The reason for this is that the Gödel Argument will allegedly be effective even if extremely severe limitations are placed on the sort of behavior that is required of an SNM.

Specifically, we may restrict the verbal behavior that is required of an SNM to computational matters. Let us say that an NM is sufficient if and only if it displays what might be called "mathematical intelligence." Turing machines, for example, are often said to display such intelligence. So, we may further clarify our sufficiency criterion in the following manner: Where A is an NM and P is a person (or more specifically, P is a human CNS), then we may say that A sufficiently models P if and only if A can compute everything that P can compute.

This criterion may be expressed by the following formulae, where 'S(a)' is interpreted as 'a is sufficient,' 'M(a,b)' is interpreted as 'a is modeled by b' and 'C(a,x)' is interpreted as 'a can compute x'. Using these conventions, we have (where A is an artefact and P is a person):

$$M(P,A) \equiv (x) (C(P,x) \supset C(A,x)) \quad (4.1-2)$$

and,

$$S(A) \equiv (\exists P) (M(P,A)) \quad (4.1-3)$$

Intuitively, these formulae mean that, in Turing's Test situation, A (the "machine") passes the test if and only

if there is at least one person such that the interrogator is unable to distinguish between that person and A. For our purposes, of course, the person, P, would minimally be a normal human adult and the interrogator would be restricted to questions of a computational nature.

It may still be asked, however, whether or not A "understands" the questions and answers. In a more generalized Turing Test situation, it might be argued that A is comparable to a very sophisticated phonograph but that A no more understands its outputs than a phonograph does.⁷

If the "understanding" in question is taken, in some obscure sense, to be a distinction that is characteristic only of human behavioral outputs, then this might arbitrarily rule out the possibility of an SNM "understanding" anything. It should be noted that it is tautologous that human outputs are human outputs, and likewise, SNM outputs are SNM outputs. The most that can reasonably be expected of an SNM is that its outputs be similar to human outputs.

A strong argument may be raised against Turing's Test on the basis of the question of whether or not an artefact can meaningfully be said to "understand" anything, or more generally, be said to be "intelligent."

⁷This seems to be the sort of objection that would be raised by Malcolm and others who share his views. See [MAL 3], passim, and [ROO 1], passim.

A Malcolman, for example, may be unwilling to permit Turing to exchange the question "Can artefacts think?" with the question "Can any artefact pass Turing's Test?" It may be objected, in particular, that this latter question illicitly avoids the problem of deciding whether or not an artefact can be said to "think" or be "intelligent." If a Mechanist₂ answers the question, "Can an artefact pass Turing's Test?," affirmatively, he would, presumably, want to hold that this affirmative answer implies that artefacts can, in some sense, be said to be "intelligent." A Malcolman, however, may well object that this move begs the question. Specifically, it might be objected⁸ that one could "be mistaken" in Turing's Test situation.

It is alright, in Turing's view, to speak of the interrogator as being mistaken merely about whether a person or a "machine" was in the other room, but it is potentially devastating to Turing's Test and to Mechanism₂, to speak of the interrogator as being mistaken about whether or not the object in the next room was thinking (Or, alternatively, was intelligent). In other words, where I_n is the interrogator and the other symbolism remains the same, we have two possible situations, one is permissible and the other is not. If I_n remarks, following the test, "I was mistaken, it was A in the other room,

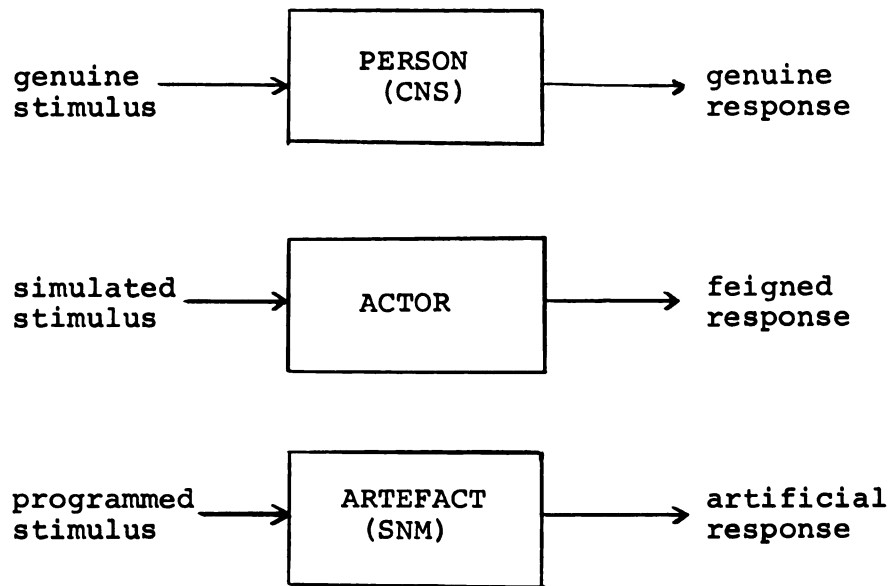
⁸ Again, this objection might be raised by one who shares Malcolm's views.

not P." then, A has passed Turing's Test. But, if I_n remarks "I was mistaken, it was A in the other room, not an intelligent/thinking thing." then the entire point of Turing's Test is destroyed.

So, we have two ways in which I_n may be said to have been mistaken; in the first way, I_n is mistaken in believing that a person rather than an artefact was in the adjoining room. The point of Turing's Test is to show that an artefact can make an interrogator be mistaken in this way. In the second way, however, I_n is "mistaken" if he believes that something in the adjoining room was thinking; this way of being mistaken makes Turing's Test not only unsuccessful, but irrelevant.

In many respects, an SNM's situation is similar to that of an actor, who is good enough at his art to make people mistakenly interpret his behavior, e.g., his pain behavior. There seems to be a high degree of intuitive correlation between an actor's feigned behavior and an SNM's artificial behavior. (See Figure 4.1-1.)

There is, however, one important difference. When an actor pretends, for example, that he is in pain, and when one knows both that he is an actor and that he is presently acting, then one can deny that he is in pain, but one would not ordinarily deny that he is intelligent, nor, for that matter, would one deny that he is behaving intelligently. If, on the other hand, an artefact (say,



(Figure 4.1-1)

an android) were pretending to be in pain, one might deny, both that it is in pain (or could be in pain) and that it is intelligent, without expecting to be contradicted. The point that we wish to make here is that there is a difference between the artificiality of an actor's performance (when he is acting, of course) and the artificiality of an artefact's performance. An actor can be said to be performing both artificially and intelligently, while an artefact may be regarded as performing artificially and non-intelligently.

There are two basic reasons why this would be the case: one reason is related to the criteria for understanding/intelligence/thinking/etc., and the other reason

is concerned with what we shall call the "anti-machine bias."

It is held, by various investigators in mathematical neural modeling, that the distinction between human beings and present-day neural models (for example, appropriately programmed digital computers) in terms of "intelligence," is a matter of quantity, not of quality.⁹ So, if this is correct, a sufficiently sophisticated artefact, if one were to be produced, would presumably be termed "intelligent."

What is at issue, however, is whether or not an artefact can meaningfully be called "intelligent." This question may be formulated in various ways, for example, "Assuming that artefacts are (capable of being) intelligent, are they (capable of being) intelligent in the same sense that humans are intelligent?" or "Is anything permitted to count as an (intelligent) SNM?" Questions of this sort require a decision, not a discovery.¹⁰ The decision which is made will inevitably depend upon intelligence criteria and modeling sufficiency criteria.

⁹See, for example, Turing's "Critical Mass Thesis" in [TUR 1], pp. 25-26. See also [ARB 1], p. 140 and [MIN 4], passim.

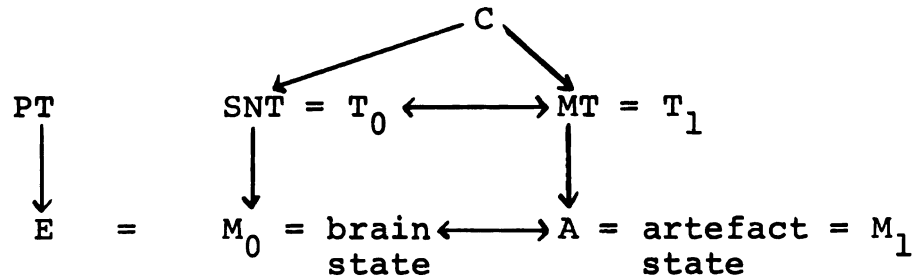
¹⁰Authors on both sides of the Gödel Problem issue agree on this point. See, for example, [PUT 1], and [KEM 1].

At present, the most widely accepted criteria for intelligence/understanding/etc. are overt, public, behavioral criteria. Since, given our restrictions, the output of an SNM would be entirely overt, this sort of criteria should be acceptable and readily applicable. This should especially be the case when the outputs in question are restricted to computational matters.

The relationship which obtains between "intelligent" human behavior and "intelligent" artefact behavior is indicated in (Figure 4.1-2). In (Figure 4.1-2), let 'PT' and 'E' be interpreted as in §3. We take E and M_0 to be the same bodily event, explained respectively by PT and SNT, and we take A to be the model-event corresponding to $E = M_0$. Admittedly, there is a potential confusion on this point since explanation instances explain discrete events which occur in physical relational systems; the relational systems, in this instance, are a brain and a neural artefact respectively. For present purposes, using as we have, Russell's Test and Turing's Test, we take P, M_0 and A to denote specific events (in (Figure 4.1-2)) which are explained respectively by PT, SNT, and MT.

Normally, in Malcolm's view,¹¹ for example, a person is said to understand (what he is saying) only if he can,

¹¹See [MAL 3] and [ROO 1].



(Figure 4.1-2)

by publically observable behaviorisms, for example, pointing, demonstrate that he understands. Presumably, he would have to point at things and name these things consistently, over a period of time (or, at least, until all doubt was erased on the part of observers), in order to successfully demonstrate his understanding. The problem of humanoid androids has been somewhat disconcerting to people who share Malcolm's viewpoint, since a "cleverly" designed android could conceivably fool anyone but its manufacturer for a long period of time, or at least until, for example, it was dismantled in the presence of observers who had previously called it "intelligent." At that point, like the I_n mentioned above, the observers, if they shared Malcolm's view, would say that they had been mistaken (in the second sense which we indicated).

It is not clear to us what conclusion should be drawn from this. It seems that the whole question of sufficiency and intelligence criteria is in great need of

further research, as we shall urge in §5, both in terms of human intelligence and in terms of artefact sufficiency. One conclusion that could be drawn from the difficulty which we have been discussing is that artefacts may be "intelligent," but the intelligence in question may not be identical to human intelligence.¹²

The distinction we should make then, is between "real" human, intelligent behavior and artefact modeled behavior. It should be clear that an M_1 may model an M_0 without being identical to it; indeed, a large part of the value of an M_1 is that it is not identical to M_0 (if it were, it would be trivial in the worst sense, since every relational system is isomorphic to itself).

If, for example, an adequate mechanical model were constructed to model an electromagnetic system (as was, for example, attempted by Maxwell), it would not be meaningful or constructive to criticize the mechanical model by saying that its "behavior" was not identical to the behavior of the electromagnetic system which it models. It is our contention that claiming that an SNM is not "intelligent" is similar to claiming that a mechanical model of an electromagnetic system is not "electromagnetic."

We conclude that the outputs of a model are not necessarily identical to the outputs of the system modeled.

¹²To be more "philosophical," one might want to say that artefact "intelligence" is in a different category from human intelligence.

What is at issue, then, is not whether the distinctive attributes of the system modeled, e.g., "intelligence," should be applicable to the respective model, but, rather, whether a given model is heuristically valuable or instrumentally valuable.

Turing suggested, as we have noted, that the question "Can artefacts think?" should be replaced by the question "Can an artefact pass Turing's Test?" The Malcolman objection, however, showed that this question is also inadequate. This objection, however, was based upon confusions regarding model theory. As we see it, artefacts cannot think, if one definitionally restricts 'thought' to 'human thought.' Artefacts (if they are SNM's) do not "think," in this sense [by definition], they model thought. What is minimally required of an SNM is that its modeled thought converge upon human thought.

Hence, we suggest that Turing's question "Can an artefact pass Turing's Test?" should be replaced by a question that is phrased in terms of modeling. The question that we submit is: "Can an artefact be constructed which can model human thought processes?"

Ideally, we would like to simply speak of the modeling sufficiency of an SNM in terms of Turing's Test. As we have seen, however, Malcolman objections seem to indicate that the designation 'intelligent' may be withheld from an SNM, even if it had passed Turing's Test.

Basically, a Mentalist₂ may argue for the contention that no artefact is capable of being "sufficient," in the sense that it can be "intelligent" or can "think," in the following manner:

- (1.) If A is an SNM, then A is "thinking" (or "intelligent").
- (2.) But, only humans are "thinking" (or "intelligent").
- (3.) A is not a human.
- (4.) Therefore, A is not "thinking" (or "intelligent").
- (5.) Therefore, it is not the case that A is an SNM.

Clearly, this argument may be reiterated, for any A (where A is a candidate SNM). We have attempted to refute this argument by denying step (2.), i.e., we deny that only humans are (or can be) "intelligent." We observe, specifically, that (2.) holds only if 'intelligence' or 'thinking' is defined to apply only to humans; this definitional restriction results from a confusion regarding modeling relations.

If we take 'thinking₁' to designate human thought processes and 'thinking₂' to designate neural artefact processes, where 'thinking₂' =_{df} 'modeled (human) thought,'

then the interrogator in Turing's Test situation can claim to be mistaken about there being a thinking₁ thing in the other room without destroying the intent of Turing's Test. We may rephrase his remark as "I was mistaken, there was not a thinking₁ thing in the other room, but there was a thinking₂ thing, viz., an SNM." Since 'thinking₁ thing' is synonymous with 'person' and 'thinking₂ thing' is synonymous with 'SNM,' we may conclude that our question "Can an artefact be constructed which can model human thought processes?" may be equivalently phrased as "Can a thinking₂ artefact be constructed?" or, alternatively, "Is a thinking₂ artefact logically possible?"

When "thought" is restricted to computational intelligent behavior, then it appears that the convergence between thinking₁ and thinking₂ is nearly complete. It, in fact, does not seem immediately clear that Malcolmian criteria would be adequate to distinguish between humans and artefacts when only computational behavior is used as a basis for comparison. Both humans and artefacts would presumably have quite similar modes of communication, viz., for the most part, computational communication of any appreciable level of sophistication would be made by means of symbols written on paper, a scratch pad for a human and a print-out for a machine. Pointing and naming, as overt criteria for "understanding" and intelligence, would seem to be less relevant here than it would be in determining

"understanding" of a more general sort. What is important here, as a criterion for understanding, is whether or not computational principles (e.g., axioms and postulates) can be applied in various circumstances and, accordingly, in various ways. It is argued by proponents of Mechanism₂ that, at present, machines can be programmed to apply principles in various ways, and therefore "understand" the principles, in this sense.¹³ Apparently, then, artefact understanding is, to a large extent, convergent with human understanding in this respect.

Therefore, an operational definition of 'intelligence' or 'understanding,' with respect to computational matters, might be at least convergent for both humans and artefacts. Considering our stated objective, however, it seems adequate to operationally define 'artefact (computational) intelligence' in terms of human (computational) intelligence. In order to do this we shall say, where 'I(a)' is interpreted as 'a is intelligent,'

$$I(A) \equiv S(A) \quad (4.1-4)$$

Hence, we have [by (4.1-2), (4.1-3), and (4.1-4)],

$$(\exists P)(I(A) \supset (x)(C(P,x) \supset C(A,x))) \quad (4.1-5)$$

¹³See [MIN 4], [ARB 1], [NEU 1], and [WIE 1].

The Gödel Argument is specifically used to deny propositions of the form (4.1-5) by falsifying the consequent.

We want to mention, finally, the anti-machine bias that might serve to definitionally rule out the possibility of an SNM by permitting nothing [by definition] to count as an SNM. What must be guarded against is an unwarranted assumption regarding the limitations of modeling.

Apparently, one of the reasons that it is assumed that the human brain cannot be sufficiently modeled is that the existence of such a model might challenge human dignity. For somewhat similar reasons, the findings of Copernicus were denied. At any rate, some find it demeaning that a mere "machine" could model human rationality. Most often, theological and metaphysical doctrines are the bases of such feelings; we, however, choose not to consider such issues in this study. Suffice it to say that, as we mentioned earlier (in §2) a great deal of the mind-body problem, as well as the anti-machine bias, is a result of misconceptions which are based upon Seventeenth Century mechanism, as a result of which a strict dichotomy was formed between "lifeless," inert matter, and "vital" spirits. The dichotomy rests on a confusion, because later findings have indicated that matter is not inert, in the Seventeenth Century mechanist's sense; as a result of these findings, the meanings of the terms 'matter,'

'machine,' etc. have undergone radical changes. Confusion results from using Seventeenth Century meanings in Twentieth Century contexts. Since "machines" are made of matter, and, hence, are in the Seventeenth Century mechanist's view, "lifeless" and "inert," they are considered to be incapable of modeling human rationality. It should be noted, however, that, as in the case of the concept of "matter," the concept of "machine" has undergone a remarkable transformation, particularly within this century.¹⁴ The transformation is so striking, in fact, that (as Toulmin has indicated) if Descartes were to be confronted with a Twentieth Century computer, he might well remark, "That is not what I mean by a 'machine.'"¹⁵

The term 'machine' has nevertheless retained some of the emotional stigma that was placed on it by Seventeenth Century mechanism. It is primarily for this reason that we prefer the more benign and potentially less offensive term 'artefact.' It is also to our advantage that we have introduced 'artefact' as a technical term, i.e., as being synonymous with 'physical model' (as explicated in §2.0).

In summary, then, the logical possibility of Mechanism₂ depends on the answer to the question of

¹⁴For some excellent discussions of this point, see [TOU 3], [SMA 3], [TOU 2], and [TOU 5].

¹⁵See [TOU 3], p. 825.

whether or not neural artefacts can think₂ (or, alternatively, 'are intelligent₂', where the concept of "intelligence₂" corresponds in an obvious manner to the concept of "thinking₂"). We have established that artefacts do not, and cannot think₁. Hence, the Gödel Argument will be effective only if it can successfully deny the possibility of a thinking₂ neural artefact.

Since artefact and human computational intelligence seem to converge, however, it appears that the Gödel Argument can be used to deny the possibility of a thinking₂ artefact.

We allow, therefore, that the Gödel Argument may be applied to the appropriate issue (i.e., the issue of whether or not artefacts can possibly be intelligent₂) and, hence, is potentially devastating to Mechanism₂. So, while making use of our working definition of neural artefact sufficiency, we now turn to the Gödel Argument itself, together with some of its more significant variations, in order to determine its logical structure and its effectiveness.

§ 4.2

THE GÖDEL ARGUMENT

Gödel's first theorem states, in effect, that for every system that is both consistent and expressively powerful enough to produce elementary arithmetic, there are well-formed formulae which are true under the intended interpretation, but which cannot be proved in the system, hence making the system incomplete.

We presume that Gödel's proof is sufficiently well known that a recasting of it here would be superfluous.¹⁶ Suffice it to say that Gödel demonstrated how, for every system, S_0 , which is powerful enough to produce simple arithmetic, an arithmetical formula, G_0 , can be constructed which represents the metamathematical statement, ' G_0 is not provable in S_0 .', which is not provable in S_0 , but which is true.

¹⁶Gödel's own proof may be found in [GÖD 1]. Other accounts of his proof are available in [KLE 1], especially pp. 204-213 and 298-316; [MEN 1], pp. 142-149; [CAR 1], pp. 129-134; [QUI 1], Chapter 7; [QUI 2], pp. 245-248; and [CHW 1], pp. 61-68. Less rigorous accounts may be found in [ARB 1], pp. 119-138 and [BEN 1], pp. 14-17. Informal, intuitive accounts of Gödel's proof may be found in [KLE 1], p. 205; [BOH 1], passim; [ROS₄ 1], pp. 155-24; [NAG 1], passim; and [DeL 1], passim. but especially pp. 59-60.

The Gödel Argument, which has been presented by such people as Rosenbloom, K. R. Popper, E. Nagel, J. Kemeny, J. Lucas, S. Jaki, and others, makes use of Gödel's findings by suggesting that for any candidate SNM, a human brain can always produce a formula, namely a Gödelian formula, that cannot be produced by that given candidate SNM. If this is correct, then there is always at least one thing that a human brain can do that any given artefact cannot do.

There have been, since Rosenbloom's first formulation of the argument¹⁷ in 1950, several variations of the argument which differ from one another in some matters of significant detail. Accordingly, we shall present first a generalized formulation of the argument and then we shall indicate some of the most significant variations. Some of the more prominent variants of the GA, in their original form, have been placed in Appendix II. We shall refer to them, in pointing out differences. Finally, we shall demonstrate an uninteresting way in which the GA may be temporarily avoided.

In a generalized form, the GA may be structured as follows:

(GA-1) The human brain can produce elementary arithmetic.

¹⁷See [ROS₂ 1], p. 208.

- (GA-2) Therefore, in order for an artefact, A_i , to model a human brain, it must be able to produce elementary arithmetic.
- (GA-3) Assume that P is a person such that an artefact A_0 models P (in the sense that A_0 can compute everything that P can compute).
- (GA-4) Artefacts are definitionally regarded as being instantiations of formal systems, i.e., for every artefact, A_i , there is a formal system, S_i , such that A_i satisfies S_i .
- (GA-5) Assume that A_0 is a candidate SNM and that it satisfies a formal system S_0 .
- (GA-6) Assume that S_0 is consistent.
- (GA-7) Then, a Gödelian formula may be constructed in S_0 ; let G_0 be that formula.
- (GA-8) G_0 cannot be proved in S_0 . [By Gödel's First Theorem, i.e., Prop. IV.]
- (GA-9) Therefore, A_0 cannot produce G_0 as being true.
- (GA-10) But, any "rational being," for example, P , could, if he knew Gödel's proof, convince himself that G_0 is true, even though G_0 is not provable by A_0 .
- (GA-11) Therefore, for every artefact A_i there is a true proposition which it cannot prove (or "produce as being true") but which a human brain, viz., P , can prove (or "produce").

(GA-12) Therefore, it is not logically possible than an artefact can sufficiently model the human brain.

The logical structure of the GA, as we have presented it, is most similar to the second formulation given by J. R. Lucas in his 1961 paper.¹⁸ Lucas' (second) formulation of the argument is probably the most comprehensive one that has been presented (it is rivaled only by Jaki (1969)¹⁹). The primary objective of Lucas' paper was to set out the GA in some detail and thereby attempt to convince logicians (and others) who were suspending judgment regarding the impossibility of an SNM until they had seen a well-structured argument.

Lucas cites the Nagel-Newman version of the GA as his primary source.²⁰ While the Nagel-Newman formulation of the GA is not actually an argument, much of the intent of Lucas' formulation is derived from it. There are notable differences as well, however. Nagel and Newman make a much more modest claim than the one which Lucas makes. Lucas wants to rule out the possibility of any artefact

¹⁸See [LUC 2], pp. 44-47. Also, see APP. II, (D) (1.).

¹⁹See [JAK 1], pp. 215-216, and APP. II. (E) (1.).

²⁰See [NAG 1], pp. 100-101, and APP. II, (B).

whatever being an SNM, while Nagel and Newman restrict their results to "currently conceived artificial machines."²¹

At any rate, the following table indicates the correspondence between the steps of the GA and the steps of the original arguments, as listed in Appendix II. It should be noted that we have, insofar as is possible, made explicit, in the GA, some steps that are either implicit or suppressed in some of the original arguments. This should serve to make the GA at least somewhat more coherent than most of the original arguments, and thereby, the GA may constitute a stronger attack upon Mechanism₂.

Some of the more notable differences between the various formulations of the Gödel Argument should be mentioned at this point. We have already indicated an important difference between Lucas' arguments and the Nagel-Newman argument (see APP. II, (B) and (D)). Rosenbloom's argument differs from the others in that it includes a psychological element. Rosenbloom holds that Gödel's findings constitute a mathematical analogue of what is called "introspection" in psychology; he concludes from this that Gödel's theorem may be used to distinguish human brains from artefacts (see APP. II, (A), especially (R-2) and (R-3)). A somewhat similar point is made by Jaki, who argues that additional ("extraneous") components

²¹[NAG 1], p. 101. See also APP. II, (B), (N-6).

Table 4.2-1.

GA	Corresponding steps in the original arguments (see Appendix II, (A) - (E)).
(GA-1) - (GA-3)	Implicit in all; most explicit in (R-1)
(GA-4)	(N-1), (K-1), (L ₁ -2), (L ₂ -1), (J ₁ -1) - (J ₁ -3).
(GA-5)	(N-1), (K-1), (L ₁ -2), (L ₂ -1), (J ₁ -1) - (J ₁ -3).
(GA-6)	Implicit in all; explicit in (L ₁ -3), (L ₂ -2, 3), (J ₁ -2).
(GA-7)	(R-5), (N-2), (K-3), (L ₁ -3), (L ₂ -7), (J ₁ -5).
(GA-8)	(R-4, 5), (N-2), (K-3), (L ₁ -3), (L ₂ -8), (J ₁ -5).
(GA-9)	(L ₁ -3), (L ₂ -9).
(GA-10)	(R-2, 3), (N-6), (K-6), (L ₁ -3), (L ₂ -10), (J ₁ -6, 7).
(GA-11)	(R-4, 5), (N-6), (K-6), (L ₁ -3), (L ₂ -11), (J ₁ -7).
(GA-12)	(R-5), (N-7), (K-7), (L ₁ -4, 5), (L ₂ -12, 13), (J ₁ -8).

would be required for any machine to prove its own consistency, where, allegedly, a proof of its own consistency would constitute one sort of "self-reflection" (see APP. II, (E)(1.), especially (J_1 -5)). It is not clear, in Jaki's argument, whether the additional component would be a hardware or software "component." Jaki's argument also differs from the others in that he explicitly points out that a randomizing device would not avoid the difficulty which is posed by the GA.

In one respect, at least, Lucas' argument seems to be more honest than the others, since he allows that it may be a considerable task to "write down an analogue of the machine's operation" (see APP. II, (D)(2.), (L_2 -5)). This point provides the basis of one of the strongest objections to the GA.

It should be made clear that the GA is used to deny the possibility of an SNM only indirectly. As we have seen (largely in §2) an SNM ($= M_1$), considered in its perspective in the modeling framework, is an instantiation of a mathematical-theoretical model ($= T_1$) which constitutes its "program" or "rules of operation." The mathematical-theoretical model (T_1), in turn, is an instantiation of a formal system, C , and is a model of an SNT, which also satisfies C . The SNT similarly, is satisfied by its logical model, a human CNS ($= M_0$), which

is (physically) modeled by the SNM. C also constitutes a formal model of both the SNT and the MT.

The GA, in virtue of the fact that Gödel's theorem applies only to consistent formal systems, is used to attack the formal system of which a candidate SNM is a logical model. The GA, therefore, is applicable to the mathematical model (of an SNT) which "programs" the SNM directly, and the GA is applicable to the SNM itself only indirectly. Note also that C (being the structure of the relevant mathematical model) is likewise affected by the GA, since it is a formal system of the kind specified in (GA-4) - (GA-6).

An additional, and somewhat surprising, result is pointed out by Jaki.²² Since C is a focus of the GA, and since the SNT is a logical model of C (alternatively, C is a formal model of the SNT) the GA may be taken to be applicable to the SNT and hence it may be applicable to Mechanism₁.

Specifically, if Jaki is correct, it appears that the GA implies that an SNT is not logically possible. We have attempted, in §3, however, to demonstrate that an SNT is logically possible. Hence, if our findings of §3 are correct, we may conclude that an SNT is logically possible. Therefore, we may conclude that the GA is not sound.

²²See [JAK 1], p. 217. See also APP. II, (E) (2.).

We have, of course, based our argument on the premise that an SNT is logically possible. We nevertheless contend that this argument should be ranked among the strongest that have been presented against the GA, since it indirectly indicates an equivocation in the GA, which we shall discuss in more detail in §4.3. Additionally, our argument seems to indicate that the superiority of a brain's computational output to an artefact's computational output is illusory.

In any event, several sorts of arguments have been presented in opposition to the GA which do not rely on the presumption that an SNT is logically possible. We shall discuss these arguments in §4.3.

Before considering these arguments, however, we shall briefly discuss a trivial (and allegedly unsuccessful) way of avoiding the GA. This will become a matter of concern in some of the arguments which we shall discuss in §4.3; in particular, it plays a central role in Lucas' "one-up-manship game," which we shall characterize presently.

Let P be a person/CNS such that T_0 models P in the sense that T_0 can compute everything that P can compute. Furthermore, let G_0 be the Gödelian formula for T_0 , and suppose that P "produces" G_0 and thereby uses the GA against T_0 , thus arguing that T_0 does not constitute an SNM. An easy, but trivial, way that a Mechanist₂ can avoid

P's GA is to add G_0 to T_0 as an axiom. Thus, we have $T_1 = T_0 \cup G_0$. Then, according to the proponents of the GA, P can allegedly produce a Gödelian formula, G_1 , which T_1 cannot compute. The Mechanist₂ can once again supplement T_1 by adding G_1 to T_1 as an axiom, and thereby generating $T_2 = T_1 \cup G_1$, but P can then allegedly produce yet another Gödelian formula G_2 for T_2 .

It is the essence of the GA that P can always stay ahead of any artefact T_i in the sense that P will indefinitely be able to keep producing Gödelian formulae for any T_i in the series $T_0, T_1, T_2, \dots, T_n$ ($i \leq n$). This constitutes Lucas' "one-up-manship" or "matching" game, which, according to him, P will always win. If, at some point, however, P is unable to produce such a formula, then the GA fails. If, on the other hand, P can always produce a Gödelian formula for every T_i , then, purportedly, the GA is successful.

While keeping this "matching" situation in mind for reference, we shall presently consider various of the arguments which may be used to attempt to refute the GA. This enterprise should enable us to assess the effectiveness of the GA and thereby provide us with a basis for a decision regarding the possibility of an SNM (and, consequently, the truth of Mechanism²).

§4.3

ARTEFACTS, BRAINS, AND GÖDEL

Opponents of Mechanism₂ (Mentalists₂) often hold that it may be possible for an artefact to model any human rational behavior, but that it is not possible for an artefact to model every such behavior. Lucas and (especially) Jaki,²⁴ for example, feel that the point which is at issue in the GA is whether or not a single artefact can (simultaneously) simulate every aspect of "mind-like" behavior. Lucas apparently takes the one-up-manship game to provide the primary basis for this contention.

Proponents of Mechanism₂, on the other hand, view the contention that any facet of "mind-like" behavior can be modeled, as supporting their views. To show this, let $\{F_1, F_2, \dots, F_n\}$ be an exhaustive set of human rational behavior (i.e., outputs), and let $\{A_1, A_2, \dots, A_n\}$ be neural artefacts that sufficiently model the respective members of $\{F_1, F_2, \dots, F_n\}$. We make the unproblematic assumption that the various kinds of human behavior can be finitely listed. Then, as Mechanists₂ would

²⁴See [LUC 2], pp. 47-48; and [JAK 1], pp. 216 and 218-219.

hold,²⁵ a single artefact could be constructed which would consist of a conjunction of $A_1, A_2, \dots, A_n$ (viz., in terms of the mathematical systems which $A_1, A_2, \dots, A_n$ satisfy, the new artefact A' would satisfy a system S which included all of the systems which are satisfied respectively by $A_1, A_2, \dots, A_n$, i.e., the larger system S would incorporate the systems, which are satisfied respectively by $A_1, A_2, \dots, A_n$, as "subroutines"). Hence, if the Mechanists₂ are correct, we might want to say that the union of $F_1, F_2, \dots, F_n$ constitutes a "person" (or, at least, the union would constitute the set of outputs of a "person") or "brain," whereas the union of $A_1, A_2, \dots, A_n$ constitutes an SNM. Hence, we have,

$$F_1 \cup F_2 \cup \dots \cup F_n = \text{a "person"/"brain"/CNS} \quad (4.3-1)$$

and

$$A_1 \cup A_2 \cup \dots \cup A_n = \text{an SNM} \quad (4.3-2)$$

This, of course, cannot be the case if a person can always out-Gödel every candidate SNM, since, in this case, there will always be at least one F_i such that the candidate SNM cannot model F_i . A Mentalist₂ would hold

²⁵See, for example, [MIN 4], pp. 446-447; [MIN 1], passim.; [ARB 1], p. 140; and [GOO 2], p. 146.

that, considering the GA, there is always at least one F_i for every artefact A_λ such that A_λ cannot "produce" F_i , where F_i is, of course, a Gödelian formula for A_λ .

There are, roughly, three sorts of arguments that can be used in opposition to the viewpoint of Mentalism₂ and specifically in opposition to the GA. (I.) First there is the argument, which was first proposed by Turing, to the effect that human brains may have the same limitations that are attributed to neural artefacts by the use of the GA. (II.) The second sort of Mechanism₂ argument is the argument that the possibility of an SNM is ruled out definitionally (and, hence, arbitrarily) by the GA. (III.) Finally, it is argued, in the last sort of Mechanism₂ argument, that the GA is based on the presumption that the formal system/"program" of an SNM can be known; it is claimed, furthermore, that this presumption is unjustified and, for this reason, that the GA fails.

We shall presently consider each of these sorts of argumentation in order to determine the viability of the GA and to determine the consequent effect that the GA may have on Mechanism₂.

I.

One way of appraising the import of the GA is to say that it constitutes an attempt to prove that artefacts, in virtue of their defining characteristics, have some (computational) limitations which are not shared by humans. In particular, it is argued that human brains do not have the computational limitations, with respect to the production of Gödelian formulae, that are characteristic of artefacts. Computational limitations, therefore, have been used by Mentalists₂ to characterize a way in which people are, in principle, different from artefacts.

The import of the GA may be symbolically represented in the following manner: Let 'Ax' abbreviate 'x is an artefact,' let 'Px' abbreviate 'x is a person,' and, finally, let 'Gxy' abbreviate 'x can out-Gödel y,' viz., 'x can produce a Gödelian formula for y.' Then, we have:

$$(x) (y) ((Ax \cdot Py) \supset ((Gyx \cdot \sim Gxy)))^{26} \quad (4.3-3)$$

That is, any person can out-Gödel any artefact, i.e., a person can produce a Gödelian formula for any artefact, but it is not the case that an artefact can produce a Gödelian formula for any person. Mentalists take this to

²⁶This formulation was suggested by Richard Hall, in private conversation.

be the basis for the contention that a person, P, is, in principle, different from, or superior to, any artefact.

Generally speaking, propositions of the form (4.3-3) are used, by Mentalists₂, to defend the contention that humans can be "intelligent," or can "think," but that artefacts cannot. We have already indicated some of the confusions which result in the contention that artefacts cannot think.

We shall attempt to refute the GA, and maintain Mechanism₂, by denying the proposition:

$$(x) (y) (Ax \cdot Py \supset Gyx) \quad (4.3-4)$$

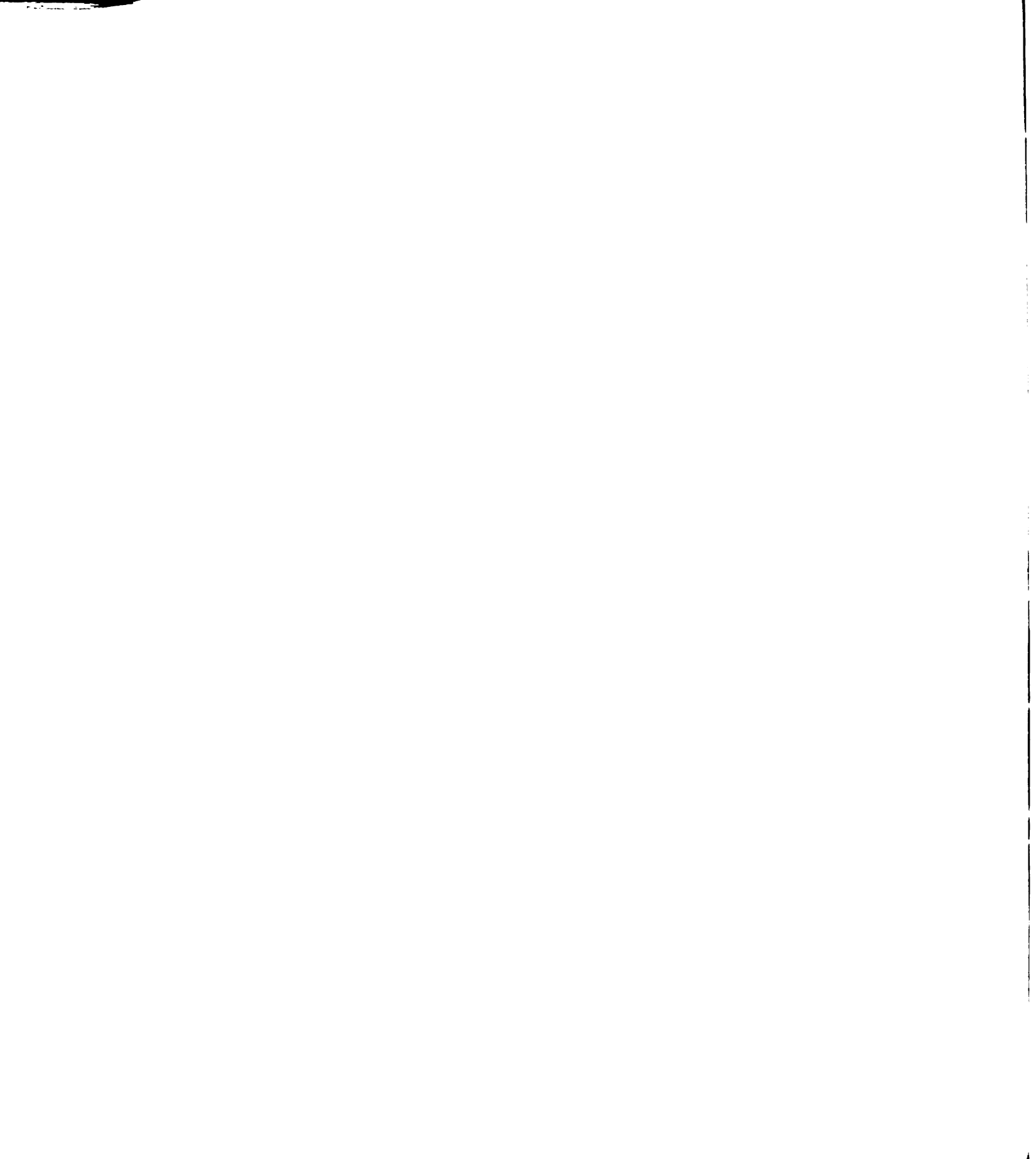
It should be noticed that (4.3-4) is a logical consequence of (4.3-3). Hence, if we can show that (4.3-4) is false, we can obtain a denial of (4.3-3).

The import of (4.3-4) may be expressed by the contention that artefacts will always have some computational limitations that people do not have. So, our denial of (4.3-4) will amount to being a denial of this contention.

It will be necessary, for the purpose of our present discussion to consider two closely related kinds of computational limitations which artefacts allegedly have but that proponents of the GA suppose that humans do not have. First, Mentalists₂ often argue that humans have

access to "informal," "intuitive," and inductive proofs, whereas artefacts do not. This amounts to saying that humans have different methods of producing their computational outputs than artefacts do; the conclusion which is drawn is that the computational processes of any person are in principle superior to any artefact's computational processes, hence, for this reason, the consequent of (4.3-3) is supposed to be true. Secondly, it is contended, by Mentalists₂, that every person, P, is computationally more complex than any artefact A. Hence, it is concluded that P is essentially more complex than any A. Here again, this alleged divergence in complexity is taken to be in support of the contention that the consequent of (4.3-3) is true.

Considering first the contention that humans have methods of computation which are different from any artefact's methods, it is often contended, by Mentalists₂, that artefacts are, by definition, consistent, deductive, etc., whereas humans are not. This is apparently taken to be one primary reason why humans are allegedly always able to produce Gödelian formulae for any artefact, and are, in this respect, computationally superior to artefacts. In order to determine the sense in which human computational processes differ from artefact processes, therefore, we must examine the possibility that humans are either not consistent or not deductive. We shall



presume, for the present, that artefacts are both consistent and deductive, while recognizing that this characterization of artefacts is based on definitional grounds. We shall discuss the difficulties which are related to an overly restricted definition of 'artefact' later.

Gödel's theorem applies only to formal systems which are consistent and deductive.²⁷ Hence, the GA is taken to apply directly to the formal system which an artefact A_i satisfies [by definition, see [GA-4)]. It is a supposition of the GA that for every neural artefact, A_i , A_i 's formal system is (deductive and) consistent, although this assumption is not stated explicitly in most formulations of the GA.²⁸ It is, consequently, assumed in the GA that the entire output of any artefact A_i is a deductive, deterministic output (since, by definition, A_i 's formal system is entirely deductive).

In order to distinguish between P and any A_i , then, granting for the moment that an A_i must be both consistent and deductive, we must consider the possibility that P is either not consistent or not deductive.

If P is inconsistent, then, of course the GA presents no computational obstacle for P since Gödel's

²⁷See [JAK 1], p. 219; see also [GÖD 1], passim.

²⁸See Benaceraff's consistency objection, [BEN 1], p. 21.

theorem applies only to consistent systems (we presume that, if P were "inconsistent," then the SNT which P satisfies would be inconsistent).

Clearly, if P , or an A_i , was basically inconsistent, then they could generate as outputs any propositions whatever, to include Gödelian formulae which state their own consistency. If P was inconsistent, then it seems apparent that one minimal requirement for an artefact, A_i , to be an SNM (i.e., for A_i to model P , with respect to P 's computational output), would be that A_i would have to be inconsistent. If A_i is inconsistent, however, A_i could, as we have just indicated, have any formula whatever as an output. If this were the case, clearly P could not out-Gödel A_i .

This situation would be unacceptable to a Mentalist₂ for two reasons. First, the assumption that P is inconsistent, seems, as we have seen, to lead to the consequence that an SNM must also be inconsistent; the result of this circumstance is that P cannot produce a formula which the SNM will be unable to produce. This result refutes the GA and constitutes a denial of Mentalism₂. Secondly, the Mentalist₂ proposes the GA only to attempt to secure a certain degree of dignity, uniqueness, and perhaps, superiority for human beings. The assumption that P is inconsistent may, initially at least, provide a convenient way to establish a distinction between

(inconsistent) human and artefacts, which are [by definition] consistent, deterministic, etc., but this assumption involves a trade-off which is apparently unacceptable to Mentalists₂. On the one hand, a distinction can be made between humans and artefacts, on the basis of the inconsistency assumption, but on the other hand, it seems that a certain amount of the dignity, etc. that Mentalists₂ would attempt to purchase for humans, by making the inconsistency assumption, would be lost. It is probably for these, and similar reasons, that Mentalists₂ typically contend that humans are consistent. Suppose, on the other hand, that P is consistent. Lucas makes the claim that P is consistent, but he, of course, adds that the GA does not apply to P.²⁹ The basis of Lucas' claim may constitute the alleged non-quantitative distinction between P and any A_i for which we have been searching. The basis of Lucas' claim is that P can informally prove his own consistency, and therefore, P may be consistent without being a victim of the GA.

Two major points are at issue here. First, as Lucas argues, both P and A_i are consistent (Lucas, of course, does not use our symbolism). But, secondly, P can allegedly (informally) prove his own consistency whereas A_i cannot, the reason being that P has access to

²⁹ See [LUC 2], pp. 55-57.

informal proofs whereas A_i does not [by the definition of 'machine' or 'artefact'; this point is made explicit in (GA-4) and in the corresponding premises in the original arguments]. Therefore, P and A_i are similar in both being consistent, but they are dissimilar in that P can prove things informally whereas A_i [by definition] cannot; specifically, P can informally "prove" his own consistency (and P can informally "prove" Gödelian formulae).

Lucas' account of how P "proves" his own consistency is extremely obscure, and in view of the severe criticism that it has received, it seems to be the weakest part of his 1961 paper. Generally, he relies upon metaphysical considerations, such as those regarding the "unity" of human consciousness,³⁰ which are totally unrelated to the mathematical (and, hence, non-metaphysical) premises of the GA, which give it its apparent rigor and appeal. Lucas' failure to adequately describe how P can "prove" his own consistency, or, for that matter, how P can informally produce successive members of the sequence $G_0, G_1, \dots$, indicates a serious defect in the GA. This finding may also be taken as an indication that the purely "non-metaphysical" nature of the GA is illusory, since, in the final analysis, appeal is made to metaphysical argumentation, viz., regarding the "unity of consciousness."

³⁰See [LUC 2], pp. 56-57.

We must conclude, at this point, that the Mentalist₂ requires that both humans and artefacts be consistent. We have seen that the Mentalist's₂ account of human consistency (at least Lucas' account) is rather obscure and relies upon ontological rather than mathematical argumentation. Artefacts, on the other hand, are supposed, by Mentalists₂, to be logically consistent [by definition].

There remains, however, the possibility that humans can differ essentially from artefacts in that humans can be non-deductive while artefacts are necessarily deductive. There are two points to consider relative to this possibility. First, the only thing that requires artefacts to be deductive is the Mentalist's₂ definition of 'artefact.' This definitional restriction begs the question. Specifically, the Mentalist₂ asks whether or not humans are essentially different from artefacts; he then definitionally distinguishes between humans and artefacts (viz., humans are non-deductive and artefacts are deductive, both by definition); finally, he concludes that humans are essentially different from artefacts, thus begging the question. Secondly, Minsky, and others, have recently developed some inductive artefacts which have been at least moderately successful; also, as we shall indicate shortly, self-amending artefacts may be designed which inductively survey their own syntax.

We conclude that Mentalists₂ have not proved that the computational processes of humans are necessarily different from (or superior to) those of artefacts. The only way that Mentalists₂ have argued for this contention is by using arbitrary definitions which beg the question. The possibility remains, however, that while humans and artefacts may have similar computational processes, humans may still differ essentially from artefacts in that they are essentially more complex than any artefact. Hence, we shall have to consider this possibility.

If it could be shown that humans are essentially more complex than any artefact, then it would follow that artefacts have computational limitations which are not shared by humans.

In order to adequately examine the critique the Mentalists₂ view that humans are essentially more complex than any artefact, we shall discuss the following considerations: (1.) We shall discuss Turing's observation that there is no proof, in the GA, that a human will always be able to continue to produce Gödelian formulae, for any artefact. (2.) Secondly, we shall attempt to analyze the general concept of one thing being "essentially more complex" than another thing, and then we shall apply our findings to the alleged divergence in complexity between humans and artefacts. (3.) Thirdly, we shall also consider, in some detail, the possibility of an

appropriately designed, self-amending artefact being comparable to a human with respect to computational output.

(4.) Finally, since a proof of the contention that a human will always be able to produce a Gödelian formula, for any artefact, is required in order to maintain the GA, we shall consider the consequences, for Mentalism₂, of there being such a proof.

The question which we must consider, at this point, is, does the GA prove that this alleged divergence of computational limitations exist between humans and artefacts, or does it simply include the assumption that the alleged divergence exists?

Turing has argued that it is assumed, but not proved, in the GA, that the alleged divergence exists. He points out, in particular, that ". . . it has only been stated, without any sort of proof, that no such limitations apply to the human intellect."³¹ Turing notes that people often give the "wrong answers" to questions, and, hence, we should not be too hasty in drawing conclusion about the fallibility of artefacts, from the evidence which is provided in the GA. Additionally, Turing notes that the computational superiority of a person, P, can only be evidenced by a triumph over a particular artefact. As Turing expresses this point:

³¹[TUR 1], p. 16.

There would be no question of triumphing simultaneously over all machines, in short, then, there might be men cleverer than any given machine, but then again, there might be other machines cleverer again, and so on.³²

We shall presently discuss this matter in some detail.

It should be immediately noticed, however, that, as Turing indicates, the contention that humans do not have the computational limitations which artefacts have, is assumed, but not proved, in the GA.

In order to characterize the way in which people and artefacts differ with respect to computational complexity, consider the one-up-manship game³³ which we described in §4.2. In order for the GA to have the desired result, a human being, P , must be able to produce a Gödelian formula, G_i , for any artefact, A_i . If, as we have indicated, A_i is continually augmented by the addition, at each stage in the one-up-manship game, of the Gödelian formula G_i as an axiom (thus producing $A_{i+1} = A_i \cup G_i$), then P must, according to the GA, be able to continually produce successive Gödelian formulae in the series $G_i, G_{i+1}, \dots$. The submerged assumption of the GA is, therefore, that P will never be unable to continue

³²Ibid.

³³Lucas prefers the term 'matching game,' since an SNM need not surpass P (i.e., be "one-up") in order to pass the test which is posed by the GA. An SNM need only match, or equal, P in the generation of the sequence $G_0, G_1, \dots, G_j$ of Gödelian formulae.

generating the sequence $G_i, G_{i+1}, \dots$ (where n is an arbitrarily large number).

Although both Mechanists₂ and Mentalists₂ typically discuss the GA in terms of a one-up-manship or "matching" game (as we shall) it is not strictly necessary to do so. One could also speak in terms of a rank ordering of artefacts such that each successive artefact has, as an axiom, the Gödelian formula with the smallest Gödel number of the last. The first member, A_0 , of the ordering, would have, as its formal system, a system, S_0 , which is minimally capable of producing elementary arithmetic. Then, where G_0 is the Gödelian formula for S_0 , the next member of the ordering would be $S_1 = S_0 \cup G_0$. Hence the ordering would proceed as follows: $S_0, S_1, S_2, \dots$ where for each S_i , $S_i = S_{i-1} \cup G_{i-1}$. There will, of course, be a corresponding series, $A_0, A_1, A_2, \dots$, of artefacts. Therefore, an alternative way of stating the submerged assumption of the GA is to say that P can always produce a Gödelian formula for each successive member of the sequence $S_0, S_1, S_2, \dots$ (which, of course, corresponds to a sequence of artefacts $A_0, A_1, A_2, \dots$.)

It appears that storage limitations are not at issue here, i.e., this is not a question of whether a human or a machine will run out of internal storage first. If we grant both P and A_i an indefinite amount of external storage, e.g., scratch paper, then only a minimum amount

of internal storage is required by either of them. The finiteness of the internal storage of P or A_i could not prevent P or A_i from constructing some (constructible) infinite ordinals. A Turing Machine, for example, needs only one binary digit of internal storage.³⁴ Therefore, it appears that the distinction between P and A_i cannot be made on the basis of storage capacities, especially if possible temporal restrictions are not considered to be relevant (i.e., if one assumes that the computations in question could be carried out "given enough time"; it should be noted that Lucas makes this assumption with regard to the production of Gödelian formulae by P (see Appendix II, (D)(2.), (L_2 -5)). Lucas states, to be specific, that P could continue the generation of the sequence $G_i, G_{i+1}, \dots$ if at each stage in the sequence, P was "given enough time" to carry out the required computation).

The Mentalist₂, however, does not attempt to argue that humans have a larger storage capacity than artefacts; rather, the Mentalist₂ argues, simply, that humans are, in some general sense, essentially more complex than any artefact.

The general tenor of the GA is that for every P , P 's brain/CNS is of a higher level of complexity than any A_i . This is most obvious in the Nagel-Newman formulation

³⁴See [SHA 1], p. 157 and [GOO 2], p. 147.

of the GA (see Appendix II, (8), (N-6)). So, we may rephrase our present question as: "Is P essentially more complex than any A_i ?" As soon as this question is asked, it seems to be somewhat ludicrous.

It is not at all clear what would be meant by saying that one thing is essentially more complex than another thing (for example, one might want to say that the Empire State Building is more complex than the Eiffel Tower, but it is not clear what would be meant by saying that the Empire State Building is essentially more complex than the Eiffel Tower).

The qualification ("essentially") is required, by Mentalists₂, however, since omitting it would make the distinction between P and an A_i merely quantitative. That the distinction between P and currently existing A_i 's is merely quantitative is precisely what Mechanists₂ claim and what Mentalists₂ deny. It seems to be the case that no "essential" distinction of the sort that proponents of the GA require can be established.

If the primary distinction between P and A_i is based upon their divergent levels of computational-deductive complexity, then this seems to provide no distinction between P and artefacts in general, with regard, of course, to computational output. A_i can be challenged with a Gödelian formula, G_i , by P, but it can also be challenged by an artefact A_j ($j > i$). Until the level

(of complexity) υ is reached, A_υ will perform as effectively in challenging an artefact A_i as P will. A_υ , however, will be able to out-Gödel A_i only until the level υ is reached, i.e., until $A_i = A_\upsilon$ with respect to deductive complexity. The concept of "levels of computational-deductive complexity" may be effectively characterized in terms of the series $A_0, A_1, \dots$ where $A_1 = A_0 \cup G_0$, $A_2 = A_1 \cup G_1$, and so on, and where G_i is a Gödelian formula for artefact A_i . Given the sequence $A_1, A_2, \dots$, then, we may say that for any two artefacts A_j and A_i ,

A_j is (computationally-deductively) more complex than A_i if and only if $j > i$. (4.3-5)

i.e., A_j is more complex than A_i if and only if A_i occurs earlier in the sequence $A_1, A_2, \dots$ than A_j does.

It seems apparent that every normal, rational adult, human (who knows Gödel's proof, of course), is more complex than currently existing neural artefacts. We have also found, however, that no case can be made for the contention that people are essentially more complex than artefacts, firstly because no clear sense can be made of the proposition that for any two things such that one is more complex than the other, one is "essentially" more complex than the other, and secondly because for any artefact, A_i , such that a person, P , can produce a Gödelian

formula for A_i , an artefact, A_j , can be constructed such that A_j can produce a Gödelian formula for A_i , i.e., A_j can do the same thing that P can do. This seems to indicate that there is no "in principle" distinction between a person, P , and artefacts, as such. This also seems to indicate that the computational complexity distinction which obtains between a person, P , and currently existing artefacts is a quantitative (rather than qualitative) distinction.

Therefore, if the distinction between P and A_i is quantitative, as it seems to be, then it follows that P could out-Gödel A_i only if P 's level of complexity ν' was higher than that of A_i . Hence, no distinction between P_ν , and an artefact A_ν could be made unless it were proved that it is logically necessary that $\nu' > \nu$, and the GA has not done this.

An SNM, therefore, would have to be an artefact A_ν , where, for at least one P_ν , $\nu \geq \nu'$. P_ν , then, would be no more able to produce a Gödelian formula for A_ν than he is presently able to produce Gödelian formulae for other human beings (i.e., for other human brains or CNS's) or for himself.

Another consideration, which favors Mechanism₂, is the observation that artefacts may be self-amending. Where A_i is an artefact of an arbitrarily high level of complexity and μ is a formalization of a person P 's method

of generating Gödelian formulae for the series of artefacts, $A_0, A_1, \dots$, then A_i (where A_i is any member of the series $A_0, A_1, \dots$) need only be amended with μ , resulting in $A_i' = A_i \cup \mu$, in order to match P's performance in the one-up-manship game. A_i can, therefore, be expected to perform as well as P can in a one-up-manship game; if P and A_i are playing the game with a third artefact, A_h ($h \geq i$), then it is assumed, but not proved, in the GA, that A_i would fail to produce the next term of the sequence $G_0, G_1, \dots, G_i$ whereas P would not fail to produce the next term, viz., P would not fail to produce G_{i+1} .

There is an additional difficulty, however. Even if an artefact A_i were to be augmented with μ , in the manner which we have indicated, to form A_i' ($= A_i \cup \mu$), then, it may be argued that P could still produce a Gödelian formula for A_i' . This difficulty is dependent upon the assumption that A_i' is of a level of complexity that would permit P to produce a Gödelian formula for it (viz., it is assumed that where v is the complexity level of A_i' and v' is the complexity level of P_v , $v' > v$).

This assumption is defective in at least two respects. First, it is not clear that P could comprehend the formal system of A_i if this formal system were extremely complex (we shall discuss this difficulty shortly). Secondly, the assumption is logically defective.

A_i' ($= A_i \cup \mu$) could, if our earlier contentions are correct, amend itself in the appropriate manner, given the function of μ , resulting in $A_i'' = A_i' \cup \mu \cup G_i = A_i' \cup G_i$, where G_i is the Gödelian formula which P has produced for A_i'' . If P again produces a Gödelian formula, G_j , for A_i'' , then A_i'' can once again amend itself, using μ , to form $A_i''' = A_i'' \cup \mu \cup G_j = A_i'' \cup G_j$. But, it seems clear that, considering just A_i' and its series of augmentations, the series could be continued indefinitely (viz., the series $A_i', A_i'', A_i''', \dots$) thereby generating an infinite regress. Furthermore, a mechanization of the method of constructing a G_i at any stage of the series $A_0, A_1, A_2, \dots$ would result in an axiomatization of elementary number theory; the process of A_i' systematically amending itself could be continued throughout the constructible ordinals. There would, admittedly, always be a Gödelian formula if one assumes, of course, that the formal system S_i which is satisfied by any given A_i must be both deductive and consistent. It cannot be proved that A_i' would always be able to continue the series $G_0, G_1, \dots$. But, what is required of a Mentalist₂ is that, in order to preserve the GA, he must prove that P can always continue the series, i.e., that P can always produce the next member of the series, $G_0, G_1, \dots$ in a one-up-manship game.

To summarize our findings, at this point we have seen that the point of view which is taken by proponents of the GA is essentially refuted by the observation that Gödel's construction could be carried out by an artefact, A_i . Specifically, we have demonstrated that for every artefact, A_i , such that a person, P , can produce a Gödelian formula for A_i , a more complex artefact, A_i' , can be constructed which can do the same thing.

It is still possible that a Mentalist₂ could argue for his viewpoint on the basis of transfinite counting. What we have labeled by a μ (= P 's method of producing Gödelian formulae), amounts to Gödel's proof. Therefore, the amended artefact, $A_i' = A_i \cup \mu$, may be said to "know" Gödel's proof. Clearly, if A_i is consistent, $A_i \cup \mu$ will still be consistent. More correctly, the system S_i which A_i satisfies, when amended with μ to form $S_i' = S_i \cup \mu$ will remain consistent.

According to the proponents of the GA, a person, P , will be able to produce a Gödelian formula, G_i , for every artefact A_i , since G_i will not be provable in A_i 's formal system S_i . But, as we have seen, an artefact $A_i' = A_i \cup \mu$ could do the same thing as P . It is significant to observe that A_i' could also be "programmed" to produce the sequence of Gödelian formulae, $G_0, G_1, G_2, \dots$, by repeatedly applying μ . Additionally, A_i' could be "programmed" to produce each of the theorems of the series

$A_0, A_1, A_2 \dots$ which correspond to the above mentioned series of Gödelian formulae. Assuming that these theorems could be arithmetized, in the Gödelian sense, the production of the theorems of the series, $A_0, A_1, A_2, \dots$, would constitute transfinite counting.³⁵ It is our contention that a suitably designed artefact could go as far in transfinite counting as any person, P , can go. Let A_i' be such an artefact; then, A_i' would still have a formal system which "programs" it, and a Mentalist₂ could retort by saying that A_i' could still be improved, i.e., by adding its Gödelian formula to it. This can most certainly be done if the Mentalist₂ can give an adequate account of how the improvement would be accomplished.

An adequate account of how P can always improve any given artefact A_i would constitute an explanation of how P is always able to produce Gödelian formula for any given artefact, viz., an explanation of how P is always able to continue the series, $G_0, G_1, G_2, \dots$. The best concept of explanation which we have is the covering-law model.³⁶ Since, in the framework of the covering-law model, explanations are either deductive or stochastic, it follows that the proof, that P can always improve any

³⁵See [GOO 2], passim.

³⁶This point was suggested by R. Hall, in private conversation.

given artefact, could also be formalized (i.e., "programmed") and be added to the formal system of an artefact, unless, of course, arbitrary definitional restrictions prohibit any artefact from being probabilistic or inductive.

Apparently, the Mentalist₂ loses either way. A proof that P can always improve any artefact is required in order for him to maintain his position, since, in the final analysis, such a proof is necessary to show that a person, P, is in principle, different from any artefact. But, if the Mentalist₂ produces such a proof and the proof is an adequate one, then it could be programmed also and added to the formal system of an artefact.

It is not clear, however, that such a proof is even logically possible.

Suppose that Pr is the proof that P can always continue the series $G_0, G_1, \dots$. If a Pr were provided by a Mentalist₂, then it could be axiomatized and used as an augmentation of the formal system S' of A_i' , thus generating $S'' = S' \cup \text{Pr}$, and hence generating A_i'' , so that A_i'' could always continue the series $G_0, G_1, \dots$. Unfortunately, no such proof as Pr is possible. Where τ is the smallest non-constructible ordinal, Pr would lead to the conclusion that τ is constructible. Therefore, Pr

has as a consequence, the contradictory proposition that τ is both constructible and not constructible.³⁷

We may conclude, therefore, that the GA cannot be maintained, since it cannot be proved that a human will always be able to produce a Gödelian formula, for any artefact (proposition (4.3-4)). We have seen, in particular, that the divergence of computational processes between humans and artefacts, which would allegedly have the result which the Mentalist₂ desires, is based only upon arbitrary definitional grounds which we have seen to be not well-founded. Secondly, we have seen that no complexity distinction between humans and artefacts can be maintained, and we have seen that a proof for (4.3-4) is logically contradictory. Hence, we may conclude that (4.3-4) is false; so, we can obtain a denial of (4.3-3). Consequently, we may conclude that, to this extent, the GA has been shown to fail.

II.

Artefacts are commonly treated by contemporary Mentalists₂ in much the same way that they were treated by Seventeenth Century dualists. We have already

³⁷This argument is suggested, in part, by some remarks of W. V. O. Quine's, which are reported by J. J. C. Smart, and, in part, by some remarks made by I. J. Good. See [SMA 2], pp. 109-110; [GOO 2], passim.; and [GOO 1], passim.

indicated the conceptual difficulties which may result from carrying Seventeenth Century mechanistic conceptions of "matter" and "machine" over into Twentieth Century contexts.³⁸ It seems apparent that the Seventeenth Century concept of "machine" is used by Mentalists₂ in the GA. Jaki, for example, explicitly states that the GA can be successfully used only because of the ". . . rigidity inherent in the concept of machine."³⁹ This rigid concept of "machine," as described by Jaki, makes a machine,

. . . a combination of resistant bodies so arranged that by their means the mechanical forces of nature can be compelled to do work accomplished by certain determinate motions.⁴⁰

This quotation could easily have come from Descartes rather than coming, as it does, from a Twentieth Century methodologist. What is important, for our purposes, is that passages of this sort, by claiming an ontological distinction between mental processes and mechanical processes, definitionally restricts the sorts of things that (physical) artefacts can sufficiently model; when certain aspects of "determinism" are also considered, these may rule out, allegedly, the possibility of an SNT and, correspondingly, a sufficient MT (= T_1).

³⁸For a more classically philosophical treatment of these confusions, see [PAS 1], Chapter 3.

³⁹[JAK 1], p. 217.

⁴⁰Ibid.

These points are substantiated by some remarks made by Lucas, in his 1961 paper. After presenting the GA and considering some objections to it, he discusses Turing's "Critical Mass Thesis."⁴¹ The critical mass thesis is the view that currently existing artefacts are "subcritical," i.e., they are not yet complex enough to have the higher intellectual attributes (e.g., "consciousness," "thought," etc.) predicated of them. Turing feels, however, that eventually an extremely complex, hence "supercritical" artefact will be constructed; when an artefact's degree of sophistication reaches a certain "critical mass," then it will be appropriate to attribute the higher intellectual characteristics to it. It will then constitute an SNM. Suppose that such a "supercritical" device were to be constructed, then, in Lucas' words,

. . . it might turn out that above a certain level of complexity, a machine ceased to be predictable, even in principle; and started doing things on its own account, or, to use a very revealing phrase, it might begin to have a mind of its own.⁴²

We have certain reservations about saying that such a device would have a "mind" (if what is meant is "human mind"), since it would be merely a model of a "mind." In any event, such a device would constitute an SNM. As

⁴¹See [TUR 1], p. 25.

⁴²[LUC 2], p. 58.

Lucas holds, however, ". . . it would cease to be a machine . . . " if it (i.e., an artefact) reached a "supercritical" level of complexity. Specifically, Lucas holds that,

If the mechanist produces a machine which is so complicated that this ceases to hold good of it [i.e., it is no longer true of the "machine" in question that it is predictable, strictly determinant, etc.], then it is no longer a machine for the purposes of our discussion, no matter how it was constructed. We should say, rather, that he had created a mind, in the same sort of way as we procreate people at present.⁴³

This suggests that, for Lucas, if an artefact is not defeated by the GA, viz., if there is an A_1 that can win over or match a person P in the one-up-manship game, then it is not a "machine." It suggests that no artefact is permitted to win a one-up-manship game with a person P , and that this is in virtue of definitional criteria. The fallacy that this finding introduces in the GA is obvious.

Specifically, the GA includes the assumption that an artefact A_0 models P . But, by definition, A_0 is a "machine." Furthermore, a "machine" cannot model P because it is definitionally prohibited from doing so, since anything that (sufficiently) models P is not a "machine." Since A_0 is a "machine," therefore, it follows that A_0 does not model P and A_0 cannot model P [by definition].

Considering our account of physical modeling (in §2), it is clear that an "artefact" or "machine" is a

⁴³Ibid.

physical model, or for present purposes, simply a "model." Hence, we may substitute 'model' for every occurrence of 'machine' in the above argument; we have a result which clearly indicates the fallacious reasoning as well as indicating the significance of using a Seventeenth Century meaning for 'machine' in the argument. When we actually carry out the substitution, we have: (1) the GA includes the assumption that a model A_0 models P . (2) But, [by definition] A_0 is a model. Furthermore, (3) a model cannot model P , because it is definitionally prohibited from doing so, since anything that (sufficiently) models P is not a model. Since A_0 is a model, therefore, it follows that (4) the model A_0 does not model P , and that (5) the model A_0 cannot model P .

By (1) and (4), we have A_0 models P and A_0 does not model P , which is a contradiction.

In summary, the GA includes the assumption that an artefact A_0 both does and does not model P . As we mentioned earlier, our main reason for choosing to use the term 'artefact' was to avoid confusions such as the one which we have just discovered. In the final analysis, the confusion is based upon ontological and conceptual matters. It should be pointed out, furthermore, that, in this instance, Lucas and Jaki, at least, have abandoned once again the mathematical argument of the GA and have resorted to metaphysical argumentation (viz., to attempt

to demonstrate that "machines" are essentially different from humans). It, therefore, seems safe to conclude that the GA is not a purely mathematical (and, hence, non-metaphysical) argument as it initially appeared to be.

III.

It is most clear, in the accounts of the GA which are given by Lucas and Jaki, that there is an assumption implicit in the GA regarding P's knowledge of an artefact A_i 's formal system. The GA, in effect, is dependent upon P's ability to demonstrate that for any formal system that is adequate to generate arithmetic there is a Gödelian formula for the system, and that this formula is true. What is overlooked by proponents of the GA is that the argument depends upon the manner in which P is given the formal system of a candidate SNM, A_i . It is further presumed that P can be given, i.e., can know, the formal system of A_i . It is clear that P must be able to determine the formal system/program of A_i in order to carry out the GA with respect to A_i .⁴⁴

It is not clear, as we have indicated, that a model-maker who is constructing a very complex system such as an SNM will, practically speaking, be able to know what its detailed formal structure is, although it is certainly

⁴⁴This point is made by P. Benacerraf. See [BEN 1], p. 28.

logically possible that he know this. As M. Minsky notes, if a computer, for example, were given a very complex program, it does not follow that the programmer will have complete knowledge of what will be the output of the program.⁴⁵ This is especially true of evolutionary (i.e., "learning" or "growing") systems, such as the self-amending system which we mentioned, A_i' (or, to be absolutely precise, the System S_i' which A_i' satisfies). Similarly, one does not expect a logician or a mathematician to know all of the consequences (theorems) of a set of axioms which he has constructed.

Hence, we have the following result: The formal structure of an SNM (or, for that matter, of a brain) may be too complex to, practically speaking, be known simultaneously except in idealized and carefully controlled circumstances. But, the formal structure of an SNM must be known in order for the GA to be applied to it. Therefore, we may conclude that even if the GA were a sound argument, it would be practically impossible (or at least extremely difficult) to apply. We also draw the somewhat weaker conclusion that, in view of the practical difficulties of determining the formal structure of an SNM (especially a self-amending SNM such as A_i') from moment to moment, and considering that (in view of the Turing Test

⁴⁵See [MIN 4], p. 447.

situation, which could, of course, be reconstructed in terms of the one-up-manship game) an SNM would operate on real time, it is not at all evident that a person P would be able to keep pace with an SNM in a one-up-manship/matching game.

As evidence for these conclusions, we should mention that an SNM would be expected to be as (structurally/formally) complex as a human brain. The reason that P cannot apply the GA to other persons (or to himself; recall that, by the reflexivity of the isomorphism relation, P is isomorphic to P , i.e., $P \simeq P$, hence, P is a model of P) is that the formal systems ($SNT \vee C$) which other human brains satisfy, are too complex for P to be practically able to apply the GA to them (or to himself). This is an aspect of the Brain-Model Problem, which we shall discuss in §5.

* * *

To summarize our findings, we have seen that, firstly, human brains are likely to have the same computational limitations that artefacts have, and that no essential distinction can be made between humans and artefacts on the basis of their respective computational outputs or on the basis of computational limitations. Secondly, we have found that a definitional fallacy is in the GA which results in the consequence that an artefact, A_i , both does and does not model a person, P . Thirdly,

and finally, we have found that the proponents of the GA have overlooked practical difficulties which may be involved in the determination of the principles and conditions of operation of an SNM, and this determination is required for the production of Gödelian formulae for an SNM, and consequently, is required for the application of the GA.

We conclude that the proponents of the GA have failed to show that an SNM is not logically possible. To this extent, then, we may still take Mechanism₂ to be viable.

§4.4

DISCUSSION

Lucas (1961) noted that most logicians were reserving judgment regarding the Gödel Argument until they had seen the entire GA set out in detail. Lucas attempted to do this. We, on the other hand, have attempted to present our opposing arguments in detail.

Lucas, of course, hoped to convince logicians (and others) that the GA was a sound argument and that it could be defended against some strong objections. We have, however, found that the GA is defective in several crucial respects and that it does not refute Mechanism₂ as Lucas, Jaki and other Mentalists₂ had hoped that it would.

The extent to which the GA is effective, and hence, applicable, now becomes clear. To be specific, it seems evident that the GA can be applied to all or most existing neural artefacts and that, therefore, a person P will win a matching game with all or most existing neural artefacts. In effect, we are in agreement with the Nagel-Newman formulation of the GA which is explicitly restricted, in application, to "currently conceived

artificial machines."⁴⁶ It is openly admitted, by Mechanists₂ that currently conceived artefacts have negligible intelligence with respect to humans; it is further admitted that new concepts in artefact design will be required before an "intelligent" artefact can be constructed. The implication is that while the GA applies to currently conceived artefacts, it may not be applicable to future artefacts.

Additionally, as we have indicated, the GA is to a large extent based upon a confused conception of "machine," "thinking," "intelligence," etc. The confusions regarding the concepts of "machine," "thinking," "intelligence," etc. may, however, be largely dissolved when these concepts are placed in the perspective of the physical modeling framework which we presented in §2.

It is our contention that Mechanists₂ have been in error if they have held that artefacts can attain human intelligence. Since they are not, in fact, human artefacts cannot have human intelligence. We have argued that it is logically possible, however, that they can attain artefact intelligence, which converges with human intelligence, especially, as we have indicated, in the case of computational "intelligence."

As we noted, Turing replaced the question "Can artefacts think?" with the question "Can an artefact pass

⁴⁶[NAG 1], p. 101, and APP. II, (B), (N-6).

Turing's Test?" We have replaced this question, in turn, with the question, "Can an artefact sufficiently model a human brain/CNS?" This latter question provides a basis for making a distinction between intelligence and modeled intelligence.

Additionally, Turing felt that within fifty years an SNM would be developed and that, at that time, a person could say 'Artefact A_0 thinks.' without expecting to be contradicted. Considering our recent findings, however, this supposition is not altogether correct. As we pointed out earlier, the determination of whether or not artefacts may be said to think will be based upon a decision, not a discovery. If one takes thinking to include both human and artefact thought, then (and only then) would it be correct to say that artefacts think. The decision, in this regard, will be based upon a decision regarding the use of the word 'thinking,' together with the establishment of criteria for its use. We contend that it is more correct and more graphic to say that humans think whereas sufficient neural artefacts model thought.

Generally speaking, the Gödel Argument may be taken to lead to the conclusion that a sufficient neural artefact would, perhaps, have to be far different from the neural models which have presently been conceived and constructed. Almost certainly, an SNM would have to be more complex than currently existing neural models.

This outlook, however, presents a difficulty for investigators in the field of neural modeling which, while being tremendously overwhelming, is nevertheless one of "mere" practicality.

We conclude, therefore, that a sufficient neural artefact is logically possible. Accordingly, we conclude that, at least insofar as the Gödel Argument has been shown to fail, Mechanism₂ is viable. We maintain these conclusions, however, while holding some reservations regarding the contention that a sufficient neural artefact will ever actually be constructed.

§5.0

MECHANISM₂ AND THE BRAIN-MODEL PROBLEM

The Brain-Model Problem, as we indicated in §2.4, is really a cluster of problems, some of which are scientific or practical problems, and some of which are logical or philosophical problems. The most important aspect of the Brain-Model Problem, for the purposes of this study, is expressible in terms of the question, "Is it logically possible to construct a sufficient neural artefact (SNM) that is simple enough to be understood?" This question is clearly based on the presupposition that an affirmative answer can be given to two questions which are logically prior to this one, namely, "Is an SNT logically possible [Mechanism₁]?" and "Is an SNM logically possible [Mechanism₂]?"

We concluded, in §3, that, to the extent to which the self-refutation arguments which we discussed were unsuccessful, an SNT may be considered to be logically possible. Similarly, we concluded, in §4, that inasmuch as the GA was unsuccessful, an SNM may be regarded as being logically possible. The Brain-Model Problem leaves us with one remaining question: "Can an SNM be understood?"

Some clarification is required at this point. Clearly, one theorizes in order to gain a viable understanding of a given subject matter, and, to an extent, one constructs models (theoretical or physical) for the same reason. However, in the instrumental view of modeling, which we have taken, models are considered as being constructed in order to (heuristically) facilitate theory formulation, and therefore, models may be regarded as only indirectly contributing to understanding.

To be specific, one theorizes about the human brain (and its functions) in order to understand it. So, a person who "understands" the human brain may be said to do so only in virtue of a theory which he knows, that has the human brain (or CNS) as its logical model. If he completely understands the human brain, then the theory in question would constitute an SNT. Similarly, one could only be said to understand an SNM in virtue of knowing the theory, MT (= a mathematical-theoretical model of the SNT) which it satisfies.

We may safely assume that an SNM would be on a level of complexity that is comparable to that of the human brain, considering that the behavioral outputs of an SNM would have to be as varied and as sophisticated as human brain outputs. It may be expected that an SNM would be as difficult to understand as the human brain. Hence, the aspect of the Brain-Model Problem which we are

presently considering may be phrased in either of two ways: First, "Can the human brain (and its operations) be understood (viz., by a human being)?," and secondly, "Can an SNM (and its operations) be understood (e.g., by a human being, or by another SNM)?"

In the most important and interesting respect, these questions deal with self-understanding (and are, in that sense, self-referring). If we may be permitted a rather cavalier treatment of the concept of "degrees of complexity,"¹ we may generalize these questions in the following manner: Let S_i^n be a physical system (e.g., a human brain or an SNM) having a degree of complexity = n , where the relative degree of complexity is, generally speaking, determined by the sophistication of the system's behavioral output. Then, the question which is posed by the Brain-Model Problem becomes,

Can S_i^n understand any S_u^n (and its operations)?" (5.0-1)

where S_u^n is an arbitrary system having a degree of complexity = n . This question may also be asked with regard to systems which are not of the same level of complexity,

¹Turing employs this concept, for example in his 1950 paper; at this point, we want to use 'degree of complexity' in a somewhat more general sense than we did in §4. See [TUR 1], passim.

Can S_i^n understand any $S_u^{n+\sigma}$? (5.0-2)

where $n + \sigma > n$, i.e., can a system understand another system that is more complex than it is? It is generally assumed that any system can understand another system that is less complex than it is; the problem is to determine whether or not a system can understand another system that is either as complex as it is or is more complex than it is. More precisely, the central question posed by the Brain-Model Problem, which we shall consider here is,

Can S_i^n understand any S_u^σ ? (5.0-3)

where $\sigma \geq n$.

This generalized question has several interesting applications. We may ask, for example, whether or not a neurophysiologist will be able to develop an SNT that he can understand. Alternatively, and more intuitively, we may ask, is the human brain understandable by other human brains?

On the other hand, we may ask, would an SNM be understandable by a human being? Alternatively, could an SNM be understood by another SNM, or by itself?

An intriguing speculation which has been made by M. Minsky is that an SNM would very likely have the same mind-matter convictions that people do, and therefore would be unwilling to admit that it was a "mere"

machine.² Accordingly, an SNM may have the same limited understanding of its own "mental" operations that people presently do of theirs. As we have already indicated, internal storage limitations may not be a crucial factor,³ in self-understanding. As Minsky argues, the Brain-Model Problem is based upon the dimorphism of self-models which people have and which an SNM is likely to have.

In the present section, we shall discuss Minsky's argument and attempt to assess its effect on the possibility of understanding an SNM, if one were to be actually constructed. We shall then make some brief suggestions for future research.

²See [MIN 2], passim. and [MIN 4], pp. 448-449.

³See [MOO 1], [PEN 2], [NEU 2], [NEU 3], and [ARB 2].

§5.1

THE BRAIN-MODEL PROBLEM

It was in his 1965 paper that Marvin Minsky discussed, in some detail, the aspect of the Brain-Model Problem that is our present concern.⁴ In an attempt to avoid terminological confusions, we must note that Minsky uses a rather imprecise concept of "model." He apparently feels that the concept of "model" which he uses is virtually incapable of being rigorously characterized. Generally speaking, he regards the modeling relation as being an essentially ternary one, which holds between an "observer" or modeler, the model itself, and, finally, the thing which is modeled. In Minsky's words,

To an observer B, an object A* is a model of an object A to the extent that B can use A* to answer questions that interest him about A.⁵

⁴See [MIN 2], passim. The Brain-Model Problem was introduced by Minsky, in his doctoral dissertation; see [MIN 5]. His 1965 paper, i.e., [MIN 2], is reprinted in a very slightly revised form in Minsky, M. (ed.) Semantic Information Processing. MIT Press. Cambridge: 1968, pp. 449-450. We shall refer only to the original article as it appeared in the Proc. I.F.I.P. Congress, 1965.

⁵[MIN 2], p. 45.

At an initial level, at least, this concept of "model" closely approximates our concept of "theory." One theorizes about an object A in order to be able to answer questions about A. Roughly speaking, then, we might characterize Minsky's "model"-concept in terms of a verbal/conceptual description or explanation of a given object.

As we mentioned in §2, central nervous systems, in a sense, "model" their environments (viz., by "mapping" their environments onto themselves). In terms of Minsky's characterization of the modeling relation as being ternary, B would be a (total) human CNS, A would be the environment of B, and A^* would be the portion of B, $B'(\subset B)$, such that the elements of A are mapped onto the elements of B' . Insofar as this sort of "modeling" provides a basis for answering questions about the environment of a CNS, it constitutes "modeling" in the sense which Minsky intends. It should be noted that neural modeling and neural coding converge at this point.

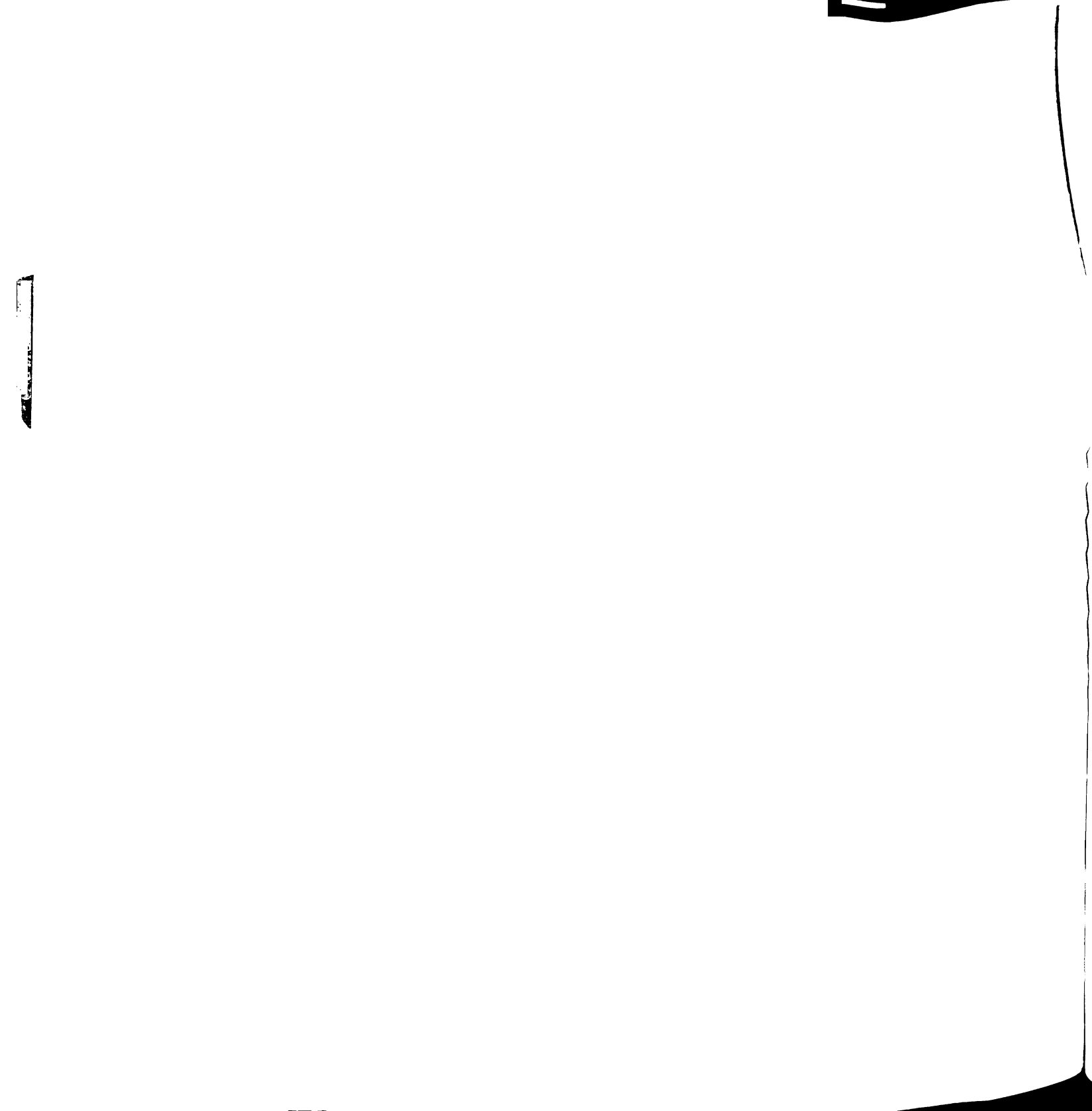
It is essential to Minsky's characterization of the Brain-Model Problem that both human beings and artefacts answer questions on the basis of the environmental models which we have described. It would be convenient to speak of some internal mechanism in either humans or neural artefacts which conducts the reflective enterprise in question; unfortunately, however, there are no simple,

discrete operations or mechanisms in either humans or in (existing) neural artefacts which can be pointed out as performing this function.⁶

Consider a person, P , together with his total environment W (which would include his own body). Let W^1 be the first-level environmental model of P 's W . Then, according to Minsky, P answers questions about W in terms of statements made about the structures in W^1 which correspond to structures in W . Ideally, since W^1 is embodied in physical structures within P , one might want to say that W^1 is a physical model of W . Certainly, however, one would not want to say that W^1 is isomorphic to W , in the old Gestalt sense of 'isomorphism.' Since W^1 is verbalized, or at least is capable of being verbalized, it is probably more correct and it is certainly more graphic to characterize W^1 as being a theory about W , i.e., W satisfies W^1 .

The Brain-Model Problem arises when we consider the status of P 's model of himself. P is a part of the world, and, in a sense, P is a portion of his own environment. Hence, $P \subset W$. It follows that P 's model of himself, P^1 , must be contained in P 's model of his total environment, i.e., $P^1 \subset W^1$. Therefore, P may answer questions

⁶See [MIN 2], p. 45.



about P in terms of statements about the (formal) structure of P^1 , which corresponds to the structure of P .

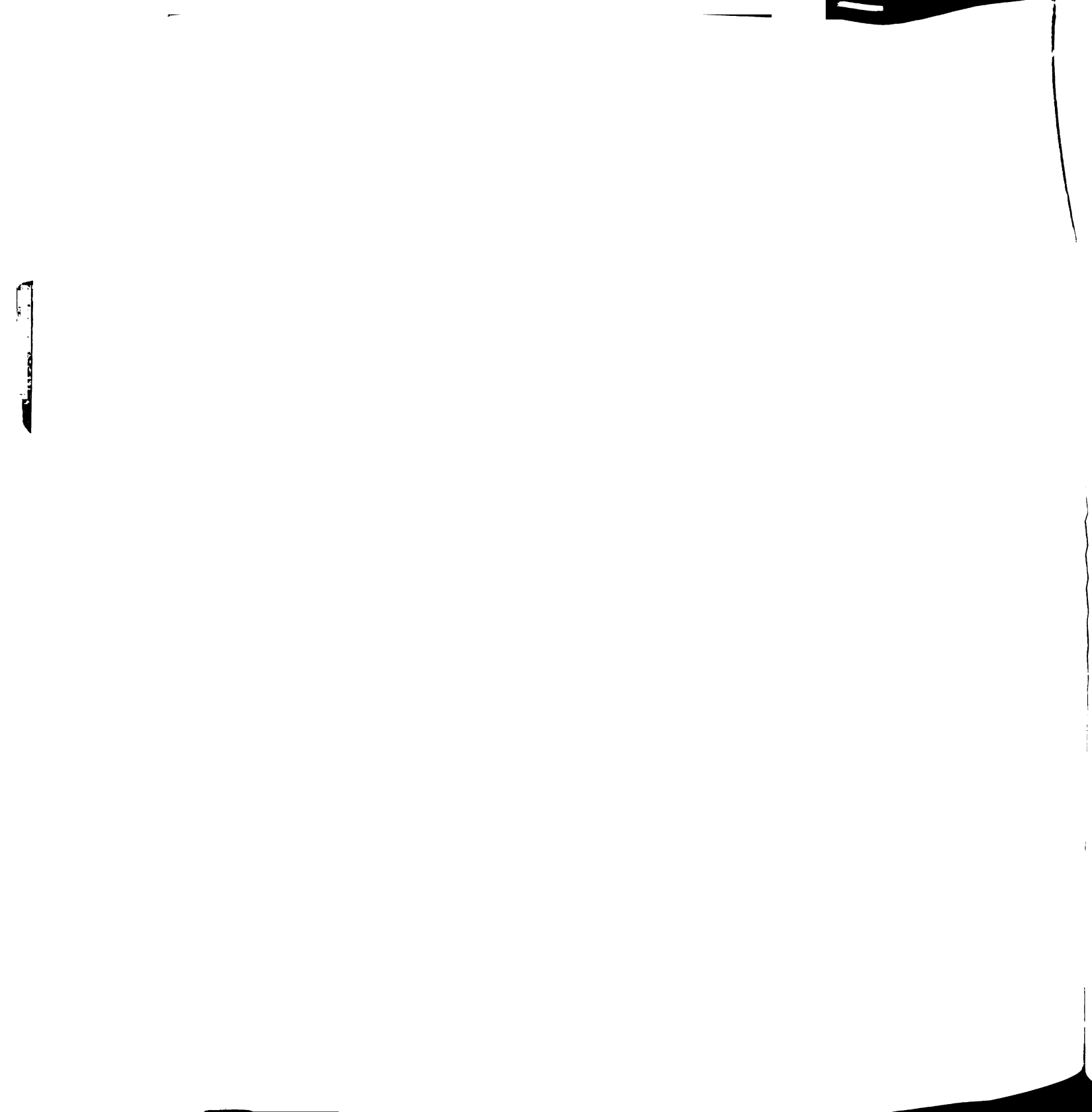
It is implicit, in Minsky's 1965 paper, that the situation would be much the same for an SNM. Specifically, let A = an SNM, and let W = the environment of A . Then, we have $A \subset W$. Furthermore, A answers questions about W in terms of W^1 , and A answers questions about A in terms of A^1 . Since $A \subset W$, we have $A^1 \subset W^1$.

Minsky makes a further distinction between various degrees of generality of P 's (or A 's) questions about W .⁷

This distinction calls upon some concepts that are similar to those employed in the object language/meta-language distinction. Generally speaking, if P asks specific questions about an object, say a tree, then he consults his environmental model W^1 . For example, if P asks how tall a given tree is, he would use W^1 to answer this question. A very general question, however, such as a question regarding what "kind" of thing a tree is, would require a higher-level, more abstract environmental model W^2 , since, in Minsky's view, the question would not be applicable directly to W , but it is, instead, applicable to W^1 .

At any rate, we may generalize Minsky's views to include P 's questions about himself. If P asks, for

⁷See [MIN 2], p. 46.



example, "How tall is P?," then he may obtain an answer to this question by consulting $P^1 (\subset W^1)$. On the other hand, if P asks, "What kind of thing is P?," then P obtains an answer in terms of the more generalized model $P^2 (\subset W^2)$.

Similarly, for A, we have $A (\subset W)$, $A^1 (\subset W^1)$ and $A^2 (\subset W^2)$.

So, we have a modeling heirarchy that may be intuitively characterized in terms of degrees of generality and mathematically characterized in terms of recursive depth. Hence, we have a modeling heirarchy that may proceed as,

$$W^0, W^1, W^2, \dots W^n,$$

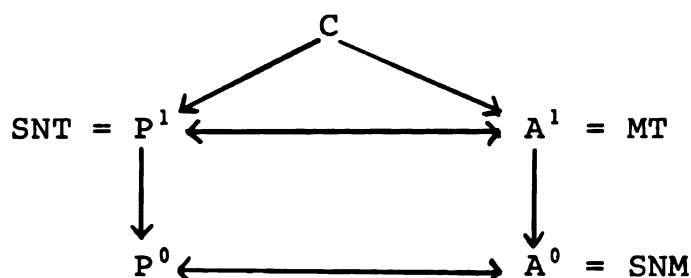
or,

$$P^0, P^1, P^2, \dots P^n,$$

Or, finally, as

$$A^0, A^1, A^2, \dots A^n.$$

For our purposes, consider that $P^1 = \text{SNT}$, and that $A^1 = \text{MT} =$ a mathematical theory which is satisfied by A. Therefore, in terms of the modeling framework, we have



(Figure 5.1-1)

In terms of the aspect of the Brain-Model Problem that is of immediate interest to us, we would like to determine whether or not P^0 can "understand" P^1 , or alternatively, we would like to determine whether or not an artefact A^0 is capable of "understanding" A^1 .

Presumably, P^0 would understand P^1 if P^1 constituted a unified explanation/description of P^0 which was simple enough to be understood.

The first difficulty, as is evidenced by the division between mental events and physical events in thought and language, is that P^0 is not unified. P^0 's environmental model W^1 is distinctly bipartite.⁸ One part of W^1 deals with mechanical and physical matters, etc. and the opposite part deals with things such as goals, intentions, social matters, etc. The division in W^1 is also reflected in P^1 , thus forming a basis for the popular mind-matter convictions.

⁸This point is suggested by Minsky in [MIN 2], p. 46, and [MIN 4], p. 449.

It should be noted that environmental models should be considered as being the end-products of a long development which includes the learning of a language which represents the bipartite structure of a W^1 . This development process would also include the process of socialization which, in part at least, contributes to the conceptual dimorphism of thought which is represented in language. As Minsky expresses these points,

An infant is not a monist: it simply hasn't enough structure in $[P^1]$ to be a dualist yet; it can hardly be said to have a position on the mind-body problem yet.⁹

Minsky's central argument for the proposition that statements of the form (5.0-3) should be answered in the negative is based upon the commonly held belief in a mind-body dualism which is a consequence of having a developed W^1 (and, hence, having a developed P^1). As we have indicated, when P^0 is asked a very general question about himself, he answers the question by giving a general description of his model of himself, where P^1 is his model of himself and P^2 is P^0 's description of P^1 . The conventional way in which P^0 will express the bipartite structure of P^1 is for him to claim that he has a mind as well as a body.

⁹ [MIN 2], p. 46.

Unfortunately, as Minsky argues, the separation of the two parts of P^1 is so vague and indistinct, and furthermore, their interconnections are so complicated (and accordingly, difficult to describe and specify) that P^0 's account of his mind-body distinction is inevitably confused and unsatisfactory.¹⁰ It seems to follow from this that P^0 's understanding of P^1 will, consequently, also be confused and unsatisfactory.

The confusion regarding P^1 is based upon more, however, than just the bipartite structure of P^1 . The concept of a "part" of an environmental model is, itself, quite obscure. As Minsky notes, the concept of a part of a computer program is much more complicated and vague than the concept of a part of a physical object. Apparently, the concept of a "part" of an environmental model is comparable to the concept of a "part" of a computer program with respect to conceptual vagueness.¹¹

We may conclude, then, that since P^0 's model P^1 of himself is not unified, P^0 's understanding of P^1 is adversely affected.

It has been argued, by Turing and others,¹² that the development of an SNM will have, as one of its

¹⁰See [MIN 2], p. 47 and [MIN 4], p. 449.

¹¹See [MIN 2], p. 47.

¹²See [TUR 1], passim. and [ARB 1], passim.

consequences, a unification of language (and possibly of science). In view of this, we should note that Minsky's "bipartite model" thesis is based, to some extent, upon the "two language" view regarding the philosophy of mind. According to this latter view, the mind-body problem is a pseudo-problem which results from a clash between two divergent languages, viz., two divergent ways of talking about the same thing. If this view is correct, then one way to resolve the problem would be to develop a unitary language, as Turing hoped would happen. Turing claimed, in effect, that once such a language is accepted, one could say that artefacts "think" without expecting to be contradicted. He, of course, felt that a unitary language would develop as a matter of course after an SNM had been successfully constructed and the fact that an SNM was in existence became widely known.

The establishment of a unitary language would, then, have the likely effect that P^0 's model P^1 of himself would likewise be unified. This effect would be realized, however, only in an ideal situation. As Minsky indicates, the heuristic value of the bipartite model is too great to warrant its unification, even if the means for unifying it became available, viz., as the result of scientific developments.¹³

¹³See [MIN 2], p. 47.

The primary value of the bipartite model is, apparently, its simplicity. It seems likely that a unified language of the sort that we have mentioned would be much too complicated to use for "every day" purposes.

We may conclude, then, that P^0 's model P^1 of himself is likely, for purely practical reasons, to remain bipartite even if an SNM (or SNT) is developed which could potentially result in the unification of language. Since P^0 's understanding P^1 is dependent upon the lack of vagueness in P^1 's structure, we seem to be forced to conclude that P^0 may not be able to adequately understand P^1 .

We, however, do not view the situation as being as hopeless as Minsky does. As we noted earlier, the mind-body problem is largely based on misconceptions which result from Seventeenth Century mechanism. Revisions which have taken place in the concept of "matter" within the last century have made the distinction between mind and matter not so much incorrect as irrelevant.¹⁴ There is, however, no assurance that a complete reduction of psychology to physical science could be accomplished (recall that in §3 we stipulated that there may be two adequate explanations for the same event). Nor is there any assurance that such a reduction would be desirable or that it would resolve the Brain-Model Problem, considering the

¹⁴For an argument for this point, see [SMA 3].

convenience and the heuristic value of the mind-matter separation in language. We conclude, therefore, that partly by choice and partly by necessity, the bipartite structure of P^0 's model P^1 of himself is likely to remain.

Considering, then, that P^0 's understanding of P^1 is likely to be adversely affected by the lack of unity in P^1 , we turn, now, to the question of whether or not P^1 will be simple enough to be understood, if it is sufficient (i.e., if it constitutes an SNT).

Recent findings regarding self-reproducing automata seem to indicate that an automaton can reproduce itself completely in all detail.¹⁵ The logical consequence of this is that a system can produce another system which is as complex as it is. This consequence reflects favorably on our conclusions regarding the logical possibility of constructing an SNM.

It is, however, impossible for an artefact (or a person) to indicate, by examining and analyzing its internal states from one moment to the next, exactly what it is doing at each step.¹⁶ An artefact (or person) that attempted to do this could never get beyond the first step.

¹⁵A recent paradox regarding self-reproducing systems, however, remains to be resolved. See [ROS₁ 2] and [ROS₁ 3].

¹⁶This point is made by Minsky in various places. See [MIN 1], passim. and [MIN 2], p. 48.

There is, on the other hand, no logical reason why an artefact (or person) cannot comprehend or "understand" its principles of operation. In addition, given enough storage (or memory), it seems possible for an artefact to understand every detail of one of its earlier states. Correspondingly, given enough information regarding specific conditions and given its principles of operation, an artefact could predict a future state of its own system.

So, we have the conclusion that the principles of operation of an artefact or human brain are likely to be simple enough to be understood from moment to moment. Considering the complexity of the systems in question, however, a given state of a system would have to be singled out, recorded in some fashion, and analyzed part by part in order to be understood.

Even at present, many of the principles of the operations of individual components of the human brain are known. In this respect, we may expect a person, P, to understand the principles of operation of the basic components of his brain. On the other hand, the principles of operation of the basic components of, for example, a digital computer are known; so, if an SNM were to be physically similar to a digital computer, we may expect that it would be able to understand the principles of operation of its basic components. Considering the complexity of a brain, or an SNM, however, it seems unlikely

that the total systemic functions of either a brain or an SNM could be understood except, as we have indicated, when a single state is selected, recorded in some manner, and studied in detail. As in Harmon's example, the deduction of ensemble functions from single-unit functions, in a brain or in an SNM, is as difficult as the problem of deducing the properties of water from one's knowledge of hydrogen and oxygen molecules. We must emphasize, however, (as Harmon does) that this is a practical difficulty at our present level of understanding; the deduction of brain or SNM ensemble functions from principles of single unit functions may not be an in-principle impossibility.

We have three consequences of this conclusion to consider which are related to our earlier findings. It should be noted, firstly, that the Brain-Model Problem indicates that Russell's Test may be very difficult to perform, but it does not make the test impossible to perform. Secondly, regarding the Description Paradox, it should be noticed that the laws of an SNT would apply to any brain state. We have seen that a person may not be able to "know" his brain states from moment to moment, therefore, he could not "know" from moment to moment successive instances of an SNT (which is satisfied by his brain states); the Description Paradox is fallaciously based on the assumption that a person could know these instances of an SNT. As we have indicated, a person would

not get beyond the first instance. Finally, as Benacerraf indicated, one must know the "program" of a neural artefact in order to apply the Gödel Argument to it. Considering our recent findings, it seems impossible that a person could keep up with a self-amending artefact in a matching game; similarly, one artefact could not keep up with another one. These observations also provide an indication of why one person cannot apply the Gödel Argument to another person.

Generally speaking, the Brain-Model Problem has the result that, in virtue of deeply entrenched mind-matter convictions, and in virtue of the complexity of the human brain and, consequently, of an SNM, there are considerable practical difficulties involved in having a situation wherein one human brain understands another (or itself), one human brain understands an SNM, and surprisingly, in having a situation wherein one SNM understands another SNM (or itself). As we have indicated, the practical difficulties are of two sorts; first, there are the difficulties which are posed simply by the complexities of the systems in question; secondly, there are the difficulties that result from the practical and heuristic value of bipartite environmental models (as we noted, the value of the bipartite structure of these models makes it unlikely that they would become unified in any circumstances).

As Minsky appraises this situation,

When intelligent machines are constructed, we should not be surprised to find them as confused and as stubborn as men on their convictions about mind-matter, consciousness, free will and the like A man's or a machine's strength of conviction about such things tells us nothing about the world, or about the man, except for what it tells us about his model of himself.¹⁷

If this appraisal is correct, then an SNM may be expected to have the same mind-matter convictions that a person does, although this would follow only if the SNM had a bipartite model of itself.

The source of the bipartite self-model of a human being seems to be clear enough, but the source of a similar division in an SNM seems to be less clear. Perhaps what Minsky has in mind is the existence of an SNM self-model that has the same simplicity and, hence, heuristic value that the self-models of human beings have. In this case, one would have to argue that the model-dimorphism in self-models is a function of model-simplicity. We cannot see how this argument would proceed. A more viable approach might be to argue that an SNM, A^0 , could not be said to sufficiently (physically) model a human, P^0 , unless A^0 had a bipartite model of itself, A^1 , that was comparable to P^0 's bipartite self-model P^1 . In any event, the provision of A^0 with a bipartite self-model

¹⁷[MIN 2], p. 48. See also [MIN 4], p. 449.

A^1 may simply be a function of model construction and "programming," viz., A^0 would be constructed in such a manner that it would be able to communicate in a human language, and human languages presumably embody a representation of the bipartite structure of human environmental models; hence, it may be the case, that as soon as an SNM were provided with a language, it would also be given the basis for a bipartite self-model and, more generally, a bipartite environmental model. Additionally, a bipartite self-model may be required in order for an SNM to pass a generalized Turing Test. It is not clear that such a self-model would be required in a Turing Test which was restricted to computational outputs, since it is not evident that computational/mathematical language embodies a bipartite structure.

It is a further consequence of the Brain-Model Problem that an SNM should not be expected to have any better an understanding of its own operations than a human being presently has of his own operations. Also, it may be expected that, in the sense which we have indicated, a person who constructed an SNM would have no better an understanding of the SNM's operations than he does of his own operations.

§5.2

SUGGESTIONS FOR FUTURE RESEARCH

In the present study, we have found that a great deal of additional work is required in some areas. Firstly, more research is required to resolve the conceptual and linguistic confusions which result from the mind-matter dimorphism which is present in natural language. Secondly, we have found the need for viable intelligence criteria for both humans and neural artefacts.

Regarding Mechanism₁, much additional research of both a conceptual and linguistic nature remains to be done in order to resolve confusions which result from the existence of two allegedly incompatible modes of language, namely the thing-language and the mind-language. To a large extent, the alleged incompatibility of, for example, actions and "mere" motions, or reasons and causes, is a consequence of this confusion. As we have indicated, a systematic up-dating of the natural language may resolve the confusion if the connotations which have been carried over from Seventeenth Century mechanism can be removed. A successful completion of this project would presumably remove many mind-matter confusions while



leaving the basic linguistic (and conceptual) mind-matter dimorphism intact; as we have shown, this latter dimorphism has a practical and heuristic value which precludes its removal.

Regarding Mechanism₂, it seems clear that, in view of the considerable practical difficulties which must be considered in the construction of an SNM, a reevaluation of the long-range goals of cybernetics may be required. In any event, as we have noted, the final appraisal of an SNM will be based upon a decision and not a discovery. This decision will require adequate criteria for both human and artefact intelligence. It appears that a single criterion would not suffice. In view of the findings of recent behavioral studies (e.g., the works of Ryle and Malcolm), it seems clear that criteria for human intelligence would be applicable only to human beings; it is our contention that this is an uninteresting truism which in no way affects the possibility of an SNM. Correspondingly, we may expect the criteria for artefact intelligence to be applicable only to artefacts. Since artefacts and humans are at least nominally non-identical (of course, any physical system is a model of itself), we conclude that their respective intelligence criteria should also be non-identical.

An adequate criterion for human intelligence has not yet been developed. Perhaps an operational definition

is the best that can be expected. At any rate, it would seemingly be very difficult to determine whether an artefact was an SNM until intelligence criteria for the system which an SNM models were made clear. We may expect artefact intelligence criteria to parallel human intelligence criteria and also be closely associated with criteria for modeling adequacy, in a general sense. In any event, we stress that a firm distinction should be made between human and artefact outputs in order to avoid confusions; to be specific, we stress the point that, in view of the relationship between a prototype and its physical model, humans are said to think, while artefacts may only be said to model thinking.

We wish to make one final comment regarding the heuristic value which one might expect to obtain from an SNM.

There is a parable about ancient map-makers who were commissioned to produce a map of a portion of a kingdom. The map was supposed to be complete in every detail. When the map-makers had completed their task, they had a perfect one-one model of a portion of the countryside, but they were dismayed to find that when the map was spread out, it covered the countryside which it was supposed to model. The moral of this story is that the map which the map-makers produced was, practically speaking, useless; the original countryside which the

map-makers were commissioned to model was as useful for navigational purposes as the map which was ultimately produced. They had done their job too well.

It may be the case that a similar situation confronts cyberneticians. An SNM may well turn out to be no more heuristically valuable than the brain which it models.

BIBLIOGRAPHY

1

BIBLIOGRAPHY

- [ACH 1] Achinstein, Peter. "Models, Analogies, and Theories." Phil. of Sci. 31. No. 4. (Oct. 1964), 328-350.
- [ACH 2] _____. "Theoretical Models." Brit. J. Phil. Sci. 16. (May 1965-Feb. 1966), 102-120.
- [ACH 3] _____. "Variety and Analogy in Confirmation Theory." Phil. of Sci. 30. (1963), 207-221.
- [ADL 1] Adler, Irving. Thinking Machines. New York: The John Day Company, 1961.
- [AND 1] Anderson, Alan R. (ed.) Minds and Machines. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1964.
- [ANI 1] Anisimov, S. F. Man and Machine.
- [ANS 1] Anscombe, G. E. M. Intention. Ithaca, New York: Cornell University Press, 1957.
- [APO 1] Apostel, L. "Towards the Formal Study of Models in the Non-Formal Sciences." Syntheses. 12. (1960), 125-161.
- [ARB 1] Arbib, Michael A. Brains, Machines, and Mathematics. New York: McGraw-Hill Book Company, 1964.
- [ARB 2] _____. "Simple Self-Reproducing Universal Automata." Inf. and Cont. 9. (1966), 177-189.
- [BEN 1] Benecerraf, Paul. "God, the Devil, and Gödel." The Monist. 51. (1967).
- [BLA 1] Black, Max. Models and Metaphors. Ithaca, New York: Cornell University Press, 1962, 219-243.
- [BOH 1] Bohnert, Herbert. "The 'Intuitive' (Semantic) Proof of 'Gödel's Theorem.'" (Mimeographed, Michigan State University).

- [BRA 1] Braithwaite, R. B. "Models in the Empirical Sciences." In Logic, Methodology, and Philosophy of Science. (Eds.) Nagel, E., Suppes, P., and Tarski, A. Stanford University Press: 1962.
- [BRA 2] _____. Scientific Explanation. New York: Harper and Row, 1953, 88-114.
- [BRE 1] Bremerman, H. J. "The Evaluation of Intelligence, The Nervous System as a Model of its Environment." Tech. Rept. No. 1, Dept. of Mathematics. Seattle: The University of Washington, 1958.
- [BRO 1] Brodbeck, May. "Models, Meanings, and Theories." In Symposium on Sociological Theory. (Ed.) Gross, L. Evanston: 1959.
- [BUR 1] Burks, Arthur W. "Computation, Behavior, and Structure in Fixed and Growing Automata." Self-Organizing Systems. New York: Peragamon, 1960, 283-311.
- [BUR 2] _____, and Wang, Hao. "The Logic of Automata." JACM. 4. (1957), 193-297.
- [BUR 3] _____. "The Logic of Fixed and Growing Automata." Ann. Comp. Lab. Harvard University. 29. Cambridge: Harvard University Press, 1957, 147-188.
- [BUR 4] _____, and Wright, J. B. "The Theory of Logical Nets." In [MOO 2], 193-212.
- [BYE 1] Byerly, H. "Model Structures and Model Objects." Brit. J. Phil. Sci. 20. (1969), 135-144.
- [CAI 1] Caianiello, E. R. (Ed.) Automata Theory. New York: Academic Press, 1966.
- [CAI 2] _____. "Mathematical and Physical Problems in the Study of Brain Models." In [CAI 1].
- [CAM 1] Campbell, N. R. "What is a Theory?" In Physics: The Elements. Cambridge: University Press.
- [CAR 1] Carnap, R. "Discussion: Variety, Analogy, and Periodicity in Inductive Logic." Phil. of Sci. 30. No. 3. (1963). 222-227.
- [CAR 2] _____. The Logical Syntax of Language. 129-134.

- [CAR 3] _____. "The Methodological Character of Theoretical Concepts." In Minnesota Studies in the Philosophy of Science. I.
- [CHU 1] Church, Alonzo. Introduction to Mathematical Logic. Princeton, New Jersey: Princeton University Press, 1956.
- [CHW 1] Chwistek, L. "A Formal Proof of Gödel's Theorem." J.S.L. 4. No. 2. (June, 1939), 61-68.
- [COH₁ 1] Cohen, Jonathan. "Can There be Artificial Minds?" Analysis. 16. (1955), 36-41.
- [COH₂ 1] Cohen, L. D. "Descartes and Henry More on the Beast Machine: A Translation of Their Correspondence Pertaining to Animal Automation." Am. Sci. I. (1936), 48-61.
- [COH₃ 1] Cohen, Morris R. and Nagel, E. An Introduction to Logic and Scientific Method. New York: Harcourt, Brace, and Company, 1934, 137-141; 221-223; 286-289; 367-376.
- [COP 1] Copi, Irving M., Elgot, Calvin C., and Wright, Jesse B. "Realization of Events by Logical Nets." JACM. 5. No. 2. (1958), 181-196.
- [CRO 1] Crosson, Fredrick J. and Sayre, Kenneth M. (Eds.) Philosophy and Cybernetics. New York: Simon and Schuster, 1967.
- [CUL 1] Culbertson, J. The Minds of Robots. Urbana: The University of Illinois Press, 1963.
- [DAV 1] Davidson, Donald. "Actions, Reasons, and Causes." J. Phil. LX. No. 23. (Nov., 1963), 685-700.
- [DeL 1] DeLong, Howard. "Unsolved Problems in Arithmetic." Sci. Amer. 224. No. 3. (March, 1971), 50-60.
- [DUH 1] Duhem, Pierre. The Aim and Structure of Physical Theory. New York: Athenuem, 1962. Part I, Chapter 4.
- [FAR 1] Farré, G. L. "Remarks on Swanson's Theory of Models." Brit. J. Phil. Sci. 18. No. 2. (1967), 140-147.
- [FEI 1] Feigenbaum, E. A. and Feldman, J. Computers and Thought. New York: McGraw-Hill Book Company, 1963.

- [FLE 1] Flew, A. Mind, Body, and Death. MacMillan.
- [FOG 1] Fogel, L. J. Biotechnology: Concepts and Applications. Englewood Cliffs: Prentice Hall, 1963.
- [FRE 1] Freudenthal, Hans. (Ed.) The Concept and the Role of the Model in Mathematics and Natural and Social Sciences. D. Reidel Publishing Company. Dordrech, Holland, 1961.
- [FRE 2] _____. "Models and Probability." Syntheses. 12. (1960).
- [GEO 1] George, F. H. Automation, Cybernetics, and Society. London: Leonard Hill Ltd., 1960.
- [GÖD 1] Gödel, Kurt. On Formally Undecidable Propositions. Meltzer, B. (trans.) New York: Basic Books, Inc., 1962.
- [GOO 1] Good, I. J. "Gödel's Theorem is a Red Herring." Brit. J. Phil. Sci. 19. (1969), 357-358.
- [GOO 2] _____. "Human and Machine Logic." Brit. J. Phil. Sci. 18. (1967), 144-147.
- [GOO 3] _____. "Logic of Man and Machine." The New Scientist. 20. (1965). 182-183.
- [GRO 1] Groenewald, H. J. "The Model in Physics." Synthese. 12. (1960), 222-227.
- [GRÜ 1] Grünbaum, Adolf. "Are Infinity Machines Paradoxical?" Science. 159, 396-406.
- [GUN 1] Gunderson, Keith. "The Imitation Game." In [AND 1], 60-71.
- [HAL 1] Hall, Richard J. "What Do We Immediately Perceive?" (Unpublished manuscript, Michigan State University).
- [HAN 1] Hanna, J. "Explanation, Prediction, and Description." Synthese. 20. (1969).
- [HAR₁ 1] Harmon, L. "Problems in Neural Modeling." In Neural Theory and Modeling. (Ed.) Reiss, R. F. Stanford: Stanford University Press, 1964.

- [HAR₁ 2] _____, and Lewis, E. "Neural Modeling." Bell Telephone System Tech. Pub. Mono. 5228.
- [HAR₂ 1] Harré, R. Theories and Things. London: Sheed and Ward, 1961.
- [HAR₃ 1] Harth, Erich M. "Brain Models and Thought Processes." In [CAI 1].
- [HEM 1] Hempel, C. Aspects of Scientific Explanation. New York: The Free Press, 1965, 433-441, 445-447.
- [HES 1] Hesse, Mary B. "Analogy and Confirmation Theory." Phil. of Sci. 30. (1963), 207-221.
- [HES 2] _____. Models and Analogies in Science. New York: Sheed and Ward, 1963.
- [HES 3] _____. "Models and Analogy in Science." Encyc. of Philosophy, Vol. 5, 354-358.
- [HES 4] _____. "Models in Physics." Brit. J. Phil. Sci. 4. No. 15. (Nov., 1953).
- [HOL 1] Holland, J. H. "Outline for a Logical Theory of Adaptive Systems." J. Assoc. Comput. Mach. 9. (1962), 297-314.
- [HOO 1] Hook, S. (Ed.) Dimensions of Mind. New York: New York University Press, 1960.
- [HUT 1] Hutten, E. The Language of Modern Physics. New York: The MacMillan Company, 1956, 81-89.
- [JAC 1] Jacobson, Homer. "On Models of Reproduction." Amer. Sci. 46. (1958), 2-5-284.
- [JAK 1] Jaki, S. Brain, Mind, and Computers. New York: Herder and Herder, 1969.
- [JOR 1] Jordan, James N. "Determinism's Dilemma." Rev. Meta. 23. No. 1. (Sept., 1969), 48-66.
- [KEM 1] Kemeny, J. G. A Philosopher Looks at Science. Princeton: VanNostrand, 1959.
- [KEM 2] _____. "Man Viewed as a Machine." In [WAL 1].
- [KLE 1] Kleene, Stephen C. Introduction to Metamathematics. Princeton: VanNostrand, 1952.

- [KLE 1] _____. "Representation of Events in Nerve Nets and Finite Automata." In [SHA 1].
- [KRI 1] Krimerman, Leonard I. (Ed.) The Nature and Scope of Social Science. New York: Appleton-Century-Crofts, 1969, 334-349.
- [KUI 1] Kuipers, A. "Model and Insight." Synthese. 12. (1960), 249-256.
- [LAS 1] Lashley, K. S. Brain Mechanisms and Intelligence. New York: Dover Publications, Inc., 1963.
- [LET 1] Lettvin, Jerome Y., Chung, S., and Raymond, S. "Multiple Meanings in Single Visual Units." (Unpublished manuscript.)
- [LUC 1] Lucas, J. R. "Human and Machine Logic: A Rejoinder." Brit. J. Phil. Sci. 19. (1968), 155-156.
- [LUC 2] _____. "Minds, Machines, and Gödel." Phil. 36. (1961). 112-127. Reprinted in [AND 1].
- [MAC 1] MacKay, D. M. "Brain and Will." In [VES 1].
- [MAC 2] _____. "The Epistemological Problem for Automata." In [SHA 1].
- [MAC 3] _____. "From Mechanism to Mind." In Brain and Mind. (Ed.) Smythies, J. R.
- [MAC 4] _____. "Mindlike Behavior in Artefacts." Brit. J. Phil. Sci. 2. (May 1951-Feb., 1952), 105-121.
- [MAL 1] Malcolm, Norman. "Dreaming and Skepticism." In Metameditations. (Eds.) Sesonke and Fleming.
- [MAL 2] _____. "The Conceivability of Mechanism." Phil. Rev. LXXVII. No. 1. (Jan., 1968), 45-72. Reprinted in [KRI 1]. We shall use the pagination in [KRI 1] throughout.
- [MAL 3] _____. "Knowledge of Other Minds." J. Phil. 56. (1959).
- [MAL 4] _____. "Scientific Materialism and the Identity Theory." In The Mind-Brain Identity Theory. (Ed.) Borst, C. V. London: MacMillan, 1970.

- [MAR 1] Martin, Michael. "Discussion: On the Conceivability of Mechanism." Phil. of Sci. 38. No. 1. (March 1971), 79-86.
- [MAX 1] Maxfield, Myles; Callahan, A. and Fogel, L. Biophysics and Cybernetic Systems. Washington, D.C.: Spartan Books, Inc., 1965.
- [MAY 1] Mays, W. "The Hypothesis of Cybernetics." Brit. J. Phil. Sci. 2. (May 1951-Feb., 1952), 249-250.
- [McC 1] McCulloch, Warren S. Embodiments of Mind. Cambridge: The M.I.T. Press, 1965.
- [McC 2] _____, and Pitts, Walter. "A Logical Calculus of the Ideas Immanent in Nervous Activity." Bull. Math. Biophysics. 5. (1943), 115-133.
- [MEI 1] Meiland, Jack W. The Nature of Intention. London: Methuen and Company, Ltd., 1970.
- [MEN 1] Mendelson, E. Mathematical Logic. Princeton: VanNostrand Co., Inc., 1964.
- [MIN 1] Minsky, Marvin L. Computation: Finite and Infinite Machines. Englewood Cliffs, New Jersey: Prentice Hall, Inc., 1967.
- [MIN 2] _____. "Matter, Mind, and Models." Proc. IFIP Congress. Vol. I. Spartan Books, 45-50.
- [MIN 3] _____. "Some Universal Elements for Finite Automata." In [SHA 1].
- [MIN 4] _____. "Steps Toward Artificial Intelligence." In [FEI 1].
- [MIN 5] _____. Theory of Neural Analog Reinforcement Systems and its Application to the Brain-Model Problem. Ph.D. Dissertation, Princeton University.
- [MOO 1] Moore, E. F. "Machine Models of Self-Reproduction." In "Mathematical Problems in the Biological Sciences." Proc. Symp. Appl. Math. 4, 17-33.
- [MOO 2] _____. Sequential Machines. Addison-Wesley Publishing Company, Inc.
- [MOR 1] Morowitz, Harold J. "A Model of Reproduction." Amer. Sci. 47. (1959), 261-263.

- [NAG 1] Nagel, E. and Newman, J. Gödel's Proof. New York: New York University Press, 1964.
- [NAG 2] Nagel, E. The Structure of Science. New York: Harcourt, Brace, and World, Inc., 1961.
- [NEU 1] von Neuman, John. The Computer and the Brain. New Haven: Yale University Press, 1958.
- [NEU 2] _____. "The General and Logical Theory of Automata." In "Cerebral Mechanisms in Behavior." Proc. Hixon Symp. (Ed.) Jeffries, L. A. New York: Wiley, 1951, 1-31.
- [NEU 3] _____. The Theory of Automata: Construction, Reproduction, Homogeneity. (Ed.) Burks, A. W. Urbana: University of Illinois Press, 1966.
- [NEW 1] Newell, Allen and Simon, H. "That 'Computer Behaviorism' can Account for Human Thinking." In The Nature and Scope of Social Science. New York: Appleton-Century-Crofts, 1969, 289-305.
- [PAS 1] Passmore, John. Philosophical Reasoning. New York: Charles Scribner's Sons, 1961.
- [PEN 1] Penrose, L. S. and Penrose, R. "A Self-Reproducing Analogue." Nature. 179. No. 4571. (1957).
- [PEN 2] Penrose, L. S. "Mechanics of Self-Reproduction." Ann. Human Gen. 23. (1958), 59-72.
- [PEN 3] _____. "Automatic Mechanical Self-Reproduction." New Biology. 28. (1959), 92-117.
- [PEN 4] _____. "Self-Reproducing Machines." Sci. Amer. 200. No. 6. (1959), 105-114.
- [PEN 5] _____. Penrose, L. S. "Developments in the Theory of Self-Replication." New Biology. 31. (1960), 57-66.
- [PER 1] Perkel, D. H., and Moore, G. P. "A Defense of Neural Modeling." In Biophysics and Cybernetic Systems. [MAX 1].
- [PER 2] Perkel, D. H. and Bullock, T. H. "Neural Coding." Neurosci. Res. Prog. Bull. 6. No. 3. (Dec. 1968).

- [PIR 1] Pirenne, M. H. "Mind-Like Behavior in Artefacts and the Concept of Mind." Brit. J. Phil. Sci. 2. (May 1951-Feb., 1952), 315-317.
- [PLA 1] Place, U. T. "Consciousness is Just a Brain Process." In [FLE 1].
- [POL 1] Polanyi, Michael. "The Hypothesis of Cybernetics." Brit. J. Phil. Sci. 2. (May 1951-Feb., 1952), 312-315.
- [POP 1] Popper, K. R. "Indeterminism in Quantum Physics and Classical Physics. II." Brit. J. Phil. Sci. 1. (1951), 179-188.
- [POP 2] _____. The Logic of Scientific Discovery. New York: Harper & Row, 1959.
- [PUT 1] Putnam, H. "Minds and Machines." In [AND 1].
- [PUT 2] _____. "Review of [NAG 1]." Phil. Sci. 27. (1960), 205-207.
- [PUT 3] _____. "Robots: Machines or Artificially Created Life." J. Phil. 61. No. 21. (Nov., 1964), 668-691.
- [QUI 1] Quine, W. V. O. Mathematical Logic. New York: Harper, 1965. Chapter 7.
- [QUI 2] _____. Methods of Logic. 245-248.
- [QUI 3] _____. "Two Dogmas of Empiricism." In Philosophy of Mathematics. (Eds.) Benacerraf, P. and Putnam, H.
- [QUI 4] _____. Word and Object.
- [RAB 1] Rabin, M. O. and Scott, D. "Finite Automata and their decision Problems." IBM J. Res. Develop. 3. (1959, 114-125.
- [RAB 2] Rabin, M. O. "Probabilistic Automata." In [MOO 2], 63-91.
- [RAS 1] Rashevsky, Nicholas. Mathematical Biophysics. Chicago: 1938.
- [RAS 2] _____. Advances and Applications of Mathematical Biology. Chicago: University of Chicago Press, 1940.

- [ROG 1] Rogers, Hartley. Theory of Recursive Functions and Effective Computability. (Mimeographed, Massachusetts Institute of Technology), 152ff.
- [ROO 1] Rood, Harold. "Malcolm and "Other Minds."" (Mimeographed, Michigan State University.)
- [ROS₁ 1] Rosen, Robert. "A Relational Theory of Biological Systems." Bull. Math. Biophysics. 20. (1958), 245-260.
- [ROS₁ 2] _____. "On a Logical Paradox Implicit in the Notion of a Self-Reproduction Automation." Bull. Math. Biophysics. 21. (1959), 387-395.
- [ROS₁ 3] _____. "The Representation of Biological Systems from the Standpoint of Categories." 20. (1958), 317-341.
- [ROS₂ 1] Rosenbloom, Paul C. The Elements of Mathematical Logic. New York: Dover Publications, Inc., 1950.
- [ROS₃ 1] Rosenblueth, Arturo. Mind and Brain. Cambridge: The MIT Press, 1970.
- [ROS₃ 2] _____, and Wiener, N. "The Role of Models in Science." Phil. of Sci. 12. (1945), 316-321.
- [ROS₄ 1] Rosser, J. B. "An Informal Exposition of Proofs of Gödel's Theorems and Church's Theorem." J. S. L. 4. (1939), 15-24.
- [RUS 1] Russell, Bertrand. "Human Knowledge." In [VES 1], 274.
- [RYL 1] Ryle, Gilbert. The Concept of Mind. New York: Barnes and Noble, Inc., 1949.
- [SCH 1] Scheffler, I. Anatomy of Inquiry.
- [SCR 1] Scriven, M. "The Compleat Robot." In [HOO 1].
- [SCR 2] _____. "The Mechanical Concept of Mind." In [AND 1].
- [SHA 1] Shannon, C. E. and McCarthy, J. Automata Studies. Princeton: Princeton University Press, 1956.

- [SHA 2] _____. "vonNeumann's Contributions to Automata Theory." Bull. Amer. Math. Soc. 64. (1958), 123-129.
- [SHA 3] _____. "A Universal Turing Machine With Two Internal States." In [SHA 1].
- [SMA 1] Smart, J. J. C. Between Science and Philosophy. New York: Random House, 1968.
- [SMA 2] _____. "Gödel's Theorem, Church's Theorem, and Mechanism." Synthese. 13. No. 2. (1961), 105-110.
- [SMA 3] _____. "Sensation and Brain Processes." In [VES 1], 424-436.
- [SMY 1] Smythies, J. R. Brain and Mind. London: Routledge and Kegan Paul, 1965.
- [SPE 1] Spector, M. "Models and Theories." Brit. J. Phil. Sci. 16. (1965).
- [STA 1] Stark, L. and Dickson, J. F. "Mathematical Concepts of Central Nervous System Function." Neurosci. Res. Prog. Bull. 3, No. 2. (1965), 24-39.
- [SUP 1] Suppes, Patrick. "A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences." Synthese. 12. (1960), 287-301.
- [SUP 2] _____. Introduction to Logic. Princeton: VanNostrand Company, Inc., 1957, Chapter 12.
- [SWA 1] Swanson, J. W. "On Models." Brit. J. Phil. Sci. 17. No. 4. (1966), 297-311.
- [TAR 1] Tarski, Alfred. "Contribution to the Theory of Models." Indagationes Mathematicae. 16. (1954), 527-588; and 17. (1955), 56-64.
- [TAR 2] _____. Logic, Semantics, and Metamathematics. Oxford: The Clarendon Press, 1956, 415-420.
- [THA 1] Thatcher, J. W. "The Construction of a Self-Describing Turing Machine." Proc. Symp. Math. Theo. Automata. (1963).
- [TOU 1] Toulmin, Stephen. "Brain and Language: A Commentary." (Mimeographed, Michigan State University.)

APPENDIX I

APP. I:(A) [JOR 1]; DETERMINISM'S DILEMMA (DD)

- (DD-1) If determinism is true, recognizing a proposition's logical truth is always also a matter of what, and only what, "heredity and environment" happen to suggest.
- (DD-2) It is possible that "heredity and environment" are so beneficent as always to suggest the right things.
- (DD-3) And it is possible that, even if they do not, they should sometimes or always contrive that we fail to affirm their false suggestions.
- (DD-4) But, neither of these is in the least likely, and even if either were likely, we would not have any reason to think so beyond what "heredity and environment" themselves suggest.
- (DD-5) Therefore, if all events have sufficient causal connections, such as "heredity and environment," any warrantable acceptance of logical truths and of arguments as valid is impossible and there can be no justifiable argument for any thesis, including determinism.

APP. I(B): [MAL 2]; MALCOLM'S RATIONAL PARADOX (MRP)

- (MRP-1) Saying or doing something for a reason (in the sense of grounds as well as purpose) implies that the saying or doing is intentional.
- (MRP-2) . . . mechanism is incompatible with the intentionality of behavior.
- (MRP-3) [Therefore], . . . my acceptance of mechanism as true for myself would imply that I am incapable of saying or doing anything for a reason.
- (MRP-4) [Therefore, if mechanism is true for me], . . . it would imply that I am incapable of having rational grounds for asserting anything, including mechanism.

APP. I(C): [TOU 5]; TOWNES' PARADOX (TP)

Let 'S' =_{df} 'Strictly causal brain mechanisms underlie all rational thought processes.' Then we have,

(TP-1) If S is true, then belief in S is non-rational.

(TP-2) If belief in S is rational, then S is false.

(TP-3) Therefore, S cannot both be true and be rationally believed.



APPENDIX II

APP. II: COMMENT

These arguments against Mechanism₂ are included here in order to make clear some of the differences in various formulations of the Gödel Argument and to make it convenient to refer to these differences in the text above.

Regarding APP. II(F), the chronology of the debate regarding the Gödel Argument, which we have presented here, is not intended to be exhaustive, but is, rather, intended to be merely representative.

APP. II(A): [ROS₂ 1], (1950)

- (R-1) ". . . one can probably prove rigorously that a machine can solve any problem which a brain, according to Wiener's model, can solve." (See [WIE 1], passim.)
- (R-2) "We may consider intelligence as the capacity for introspection, the faculty of thinking about one's own methods of reasoning and what they can accomplish."
- (R-3) "In mathematical terms, this means the capacity of using a syntax language for reasoning about an object language."
- (R-4) "It seems altogether feasible to incorporate this idea into a mathematical definition of a brain and to prove that a brain can solve some problems which a machine cannot."
- (R-5) "Theorem 4.3.4 shows that a certain problem cannot be solved by machines, i.e., that brains are necessary." (Theorem 4.3.4 is equivalent to Gödel's first theorem.)

APP. II(B): [NAG 1], (1958)

- (N-1) "Today's calculating machines have a fixed set of directives built into them; these directives correspond to the fixed rules of inference of formalized axiomatic procedure."
- (N-2) "Gödel showed in his incompleteness theorem, there are innumerable problems in elementary number theory that fall outside the scope of a fixed axiomatic method."
- (N-3) "Given a definite problem, a machine of this type might be built for solving it, . . ."
- (N-4) ". . . but no such machine can be built for solving every problem."
- (N-5) "The human may, to be sure, have built-in limitations of its own, and there may be mathematical problems it is capable of solving."
- (N-6) "But, even so, the brain appears to embody a structure of rules of operation which is far more powerful than the structure of currently conceived artificial machines."
- (N-7) Therefore, "There is no immediate prospect of replacing the human mind by robots."

APP. II(C): [KEM 1], (1959)

(K-1) "We can give a general definition of what a machine is."

(K-2) "This definition is so general that it certainly includes anything we have thought of so far, and it is difficult to see how we could build a machine violating this description."

(K-3) "Given such a machine, we can always find a problem of the kind it can understand, which it cannot solve."

(K-4) "Of course, there will be a better machine which can solve this particular problem, but we can ask an even tougher question which it in turn cannot answer."

(K-5) "These results prove rigorously that every conceivable thinking machine has its clearly defined limitations."

(K-6) "And yet it is not clear that a human being, given enough time, could not solve all these problems."

(K-7) The "philosophical conclusion" is, ". . . we are essentially different from machines."

APP. II(D)(1.): [LUC 2], (1961)

(L-0) "Gödel's Theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines."

(L-0') "Gödel's Theorem states that in any consistent system which is strong enough to produce simple arithmetic there are formulae which cannot be proved-in-the-system, but which we can see to be true."

L_1

- (L_1 -1) "Gödel's theorem must apply to cybernetical machines . . ."
- (L_1 -2) ". . . it is the essence of being a machine, that it should be a concrete instantiation of a formal system."
- (L_1 -3) "It follows that given any machine which is consistent and capable of doing simple arithmetic, there is a formula which it is incapable of producing as being true--i.e., the formula is unprovable-in-the-system--but which we can see to be true."
- (L_1 -4) "It follows that no machine can be a complete or adequate model of the mind . . ."
- (L_1 -5) Therefore, ". . . minds are essentially different from machines."

APP. II(D)(2.), [LUC 2], (1961)

L_2

- (L_2 -1) "We understand by a cybernetical machine an apparatus which performs a set of operations according to a definite set of rules."
- (L_2 -2) The outputs of such a machine are determined by the rules and axioms, and, hence, could, given enough time, be calculated.
- (L_2 -3) "Machines are definite: anything which was indefinite or infinite we should not count as a machine."
- (L_2 -4) "If there are only a definite number of types of operation and initial assumptions built into the systems, we can represent them all by suitable symbols written down on paper."
- (L_2 -5) "However long the machine went on operating, we could, given enough time, paper, and patience, write down an analogue of the machine's operation."
- (L_2 -6) "This analogue would in fact be a formal proof . . ."
- (L_2 -7) "We now construct a Gödelian formula in this formal system."

- (L₂-8) "This formula cannot be proved-in-the-system."
- (L₂-9) "Therefore the machine cannot produce the corresponding formula as being true."
- (L₂-10) "But we can see that the Gödelian formula is true: any rational being could follow Gödel's argument, and convince himself that the Gödelian formula, although unprovable in-the-system, was nonetheless--in fact, for that very reason--true."
- (L₂-11) Therefore, ". . . for every machine there is a truth which it cannot produce as being true, but which a mind can."
- (L₂-12) "This shows that a machine cannot be a complete and adequate model of the mind."
- (L₂-13) Therefore, ". . . we can never, not even in principle, have a mechanical model of the mind."

APP. II(E) (1.): [JAK 1], (1969)

J_1

- (J_1 -1) ". . . machines are formal systems and need no metamathematical proofs to work efficiently."
- (J_1 -2) "'Formal' in this context means specified by a finite number of components to perform well-defined functions."
- (J_1 -3) "For a machine to be a machine, it can have only a finite number of components and it can operate on a finite number of initial assumptions."
- (J_1 -4) The introduction of a randomizing device ". . . will allow the machine to choose only among a definite number of possible alternatives."
- (J_1 -5) "And as Gödel's theorem suggests, it is a basic shortcoming of all such systems, whether existing only on paper as theorems of arithmetic or implemented mechanically that they rely on a system extraneous to them for their proof of consistency."
- (J_1 -6) "The applicability of Gödel's theorem which states that no strong enough arithmetic theory can have in itself its proof of consistency to the mind-machine relationship can easily be seen if one

postulates that since minds can do arithmetic, any artificial mind should be able to do the same."

(J₁-7) "Yet a machine, being a formal system, can never produce at least one truth, which the mind can without relying on other minds."

(J₁-8) "Yet, no matter how perfect the machine, it can never do everything that the human mind can, or more concretely, no machine can be built that could simulate every aspect of mindlike behavior."

APP. II(E)(2.): [JAK 1], (1969)

J₂

(J₂-1) "In order for . . . a construction to qualify as a machine, it must be in some sense finite and definite . . ."

(J₂-2) ". . . as such, it would not have its proof of consistency within itself."

(J₂-3) "It follows, therefore, that the mechanist cannot even in principle devise an individual machine that might serve as an adequate model for the mind."

(J₂-4) "And since machines are of necessity built of physical or chemical components, it also follows that the human mind cannot be fully explained in terms of physics and chemistry."

APP. II (F): A REPRESENTATIVE CHRONOLOGY OF THE GÖDEL
PROBLEM DEBATE

YEARS	PAPERS	
1931	[GÖD 1]	
	MENTALISTS ₂	MECHANISTS ₂
1950	[ROS ₂ 1]	[TUR 1]
1951	[POP 1]	
1958	[NAG 1]	
1959	[KEM 1]	
1960		[GEO 1] [PUT 2] [PUT 1] [SCR 1]
1961	[LUC 2]	[SMA 2]
1964		[ARB 1]
1965		[GOO 3]
1967		[BEN 1] [GOO 2] [MIN 1]
1968	[LUC 1]	
1969	[JAK 1]	[GOO 1]

(Table APP. II - 1)

MICHIGAN STATE UNIVERSITY LIBRARIES



3 1293 03047 0557