Interactions Between Philosophy and Artificial Intelligence: The Role of Intuition and Non-Logical Reasoning in Intelligence

Aaron Sloman

School of Social Sciences, University of Sussex

Recommended by B. Meltzer

ABSTRACT

This paper echoes, from a philosophical standpoint, the claim of McCarthy and Hayes that Philosophy and Artificial Intelligence have important relations. Philosophical problems about the use of "intuition" in reasoning are related, via a concept of analogical representation, *to problems in the simulation of perception, problem-solving and the generation of useful sets of possibilities in considering how to act. The requirements for intelligent decision-making proposed by McCarthy and Hayes are criticised as too narrow, and more general requirements are suggested instead.*

1. Introduction

The aim of this paper[1] is to illustrate the way in which interaction between Philosophy and A.I. may be useful for both disciplines. It starts with a discussion of some philosophical issues which interested me long before I knew anything about A.I., and which I believe are considerably enriched and clarified by relating them to problems in A.I., which, they, in turn, help to clarify. These issues concern non-logical reasoning and the use of non-linguistic representations, especially "analogical" representations such as maps or models. This discussion is followed by some general speculations about the conceptual and perceptual equipment required by an animal or machine able to cope with our spatio-temporal environment. Finally, there are further vague, general and programmatic remarks about the relations between Philosophy and A.I.

[1] Presented to the 2nd International Joint Conference on Artificial Intelligence, at Imperial College London, September 1971. Since writing it, the author has acquired a keener appreciation of the gap between formulating such ideas and embodying them in a computing system.

Artificial Intelligence 2 (1971), 209–225

The paper was inspired mainly by discussions with Max Clowes, but also to some extent by the attempts made by McCarthy and Hayes [12], and Hayes [8] to relate philosophical issues to problems in the design of intelligent robots. My criticisms of their work should not be taken to imply unawareness of my debts.

Although I was ignorant of the remarkable papers by Minsky while developing these ideas, I now believe that many of his comments on the state of A.I., especially in his 1970 lecture [17], are intimately connected with the main themes of this paper. I do not yet know enough about computers and programming to understand all his papers listed in my bibliography, so, for all I know, he may already have taken these themes much further than I can.

2. The Limits of the Concept of Logical Validity

Within Philosophy there has long been a conflict between those who, like Immanuel Kant [9], claim that there are some modes of reasoning, or adding to our knowledge, which use "intuition", "insight", "apprehension of relations between universals", etc., and those who claim that the only valid modes of reasoning are those which use *logically valid* inference patterns. (I shall analyse this concept shortly. The problem of valid inductive reasoning, from particular instances to generalisations, is not relevant to this paper.) Although various attempts have been made to show that non-logical, intuitive, modes of reasoning and proof are important (e.g. I. Mueller [18] attempts to show that diagrams play an essential role in Euclid's *Elements*), nevertheless, the prevailing view amongst analytical philosophers appears to be that insofar as diagrams, intuitively grasped models, and the like, are used in mathematical, logical or scientific reasoning they are merely of psychological interest, e.g. in explaining how people arrive at the *real* proofs, which must use only purely logical principles. According to this viewpoint, the diagrams in Euclid's *Elements* were stricly irrelevant, and would have been unnecessary had the proofs been properly formulated.

A similar viewpoint seems to prevail in the field of A.I., despite the recent "semantic" approach, which takes non-linguistic models or interpretations into account in attempts to make the search for proofs, or for solutions to problems, more efficient. (For example, Gelernter [7], Lindsay [11], Raphael [19].) The manipulation of non-linguistic structures appears to be tolerated as "heuristics" but not accepted as a variety of valid proof. This prevailing view seems to be implicit in the following quotation from McCarthy and Hayes [12]:

> '. . . we want a computer program that decides what to do by inferring in a formal language that a certain strategy will achieve its assigned goal. This requires formalising concepts of causality, ability, and knowledge.' (p. 463)

Although McCarthy and Hayes do not discuss the question explicitly, their stress on the need for a "formal language" and "formalising concepts", and other features of their essay, suggest that they would not admit the *autonomy* of non-linguistic modes of reasoning. Their concept of a "formal language" seems to include only languages like predicate calculus and programming languages, and not, for instance, the "language" of maps. In his Turing lecure [17] Minsky inveighed at length against this sort of restriction, but failed to characterise it adequately: it is not, as he suggested, a case of concentrating on form (or syntax) while ignoring content, but a case of concentrating on too narrow a range of types of representations (or "languages"). Formalisation, for instance of syntactic and semantic rules, is indispensable: what is now needed is formalisation of the rules which make non-linguistic, non-logical reasoning possible. I shall support this remark by showing that logically valid inference is a special case of something more general.

What is meant by calling an inference, or step in a proof, from premisses $p_1, p_2, \ldots p_n$ to a conclusion c, "valid"? The fact that syntactic tests for validity can be used by machines and by men has led some to forget that what is tested for is not a syntactic relation but a semantic one, which I shall now define.

In general, whether a statement is true or false, i.e., what its tiuth-value is, depends not merely on its structure, or meaning, but also on facts, on how things are in the world: discovering the truth-value requires the application of semantic interpretation procedures in investigating the world. However, some statements are so related that by *examining* those procedures, instead of *applying* them, we can find that certain combinations of truth-values cannot occur, no matter what the world is like. "London is larger than Liverpool" and "Liverpool is larger than London" are incapable of both being true: they are *contraries* of each other. Similarly some pairs of statements are incapable of both being false: they are *subcontraries.* More generally, when certain combinations of truth-values for statements in some set S are impossible on account of (i) syntactic relations between those sentences and (ii) the semantic rules of the language, then the statements in S are said to stand in a *logical relation.* (A more accurate definition would have to make use of the concept of "logical structure". Although intuitively clear, the precise definition of this concept is very difficult.) Inconsistency, i.e., the impossibility of all statements in the set being true, is one important logical relation. Another is validity of inference, i.e. the case where what is ruled out as impossible is the conclusion, c, being false while all the premisses $p_1, p_2 \ldots p_n$ are true. Thus, logical validity is a special case of the general concept of a logical relation, namely the case where the combination of truth-values $(T, T, \ldots T: F)$ cannot occur.

My main claim is not merely that these are semantic concepts, concerning meaning, reference, denotation (e.g., denotation of truth-values) as well as form (syntax, structure), but that they are special cases of still more general concepts, which I shall now illustrate, with some examples of valid reasoning which are not logical. Many more examples can be found in Wittgenstein [23].

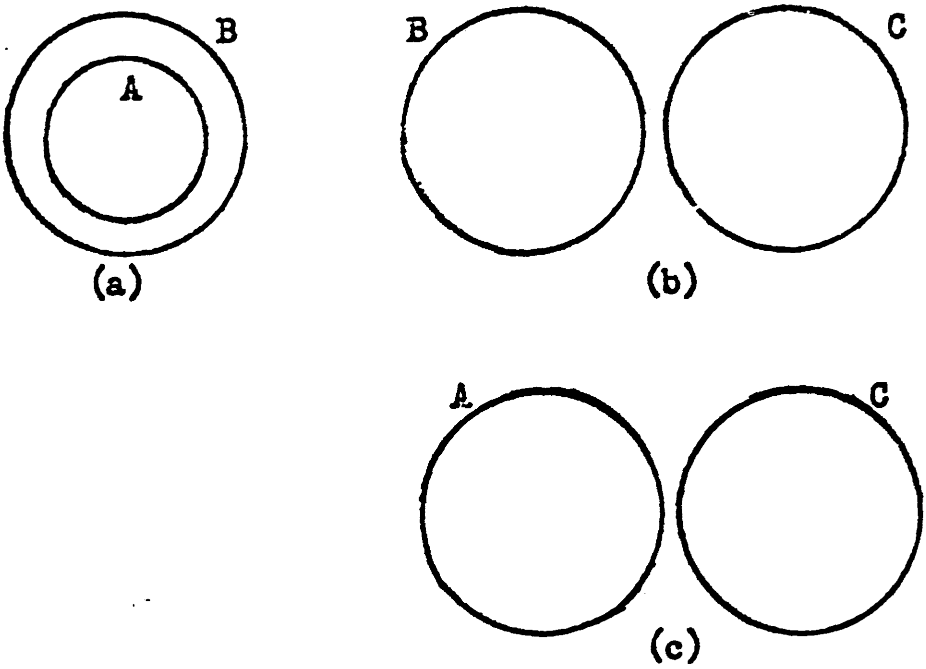

FIG. 1

Consider the familiar use of pairs of circles to represent Boolean relations between classes, as in Fig. 1, where (a) represents a state of affairs in which the class *A* is a subclass of *B*, (b) represents a state of affairs in which the classes *B* and *C* have no common members, and (c) represents *A* and *C* as having no common members. If *A* is the class of male persons in a certain room, *B* is the class of students in the room and *C* the class of redheads in the room, then clearly for each of the three figures whether it correctly represents the facts depends on how things are in the world (i.e. in the room). Nevertheless, the "inference" from (a) and (b) to (c) is valid, since no matter how things are in the room, it is impossible for the first two to represent the state of affairs while the last does not: that combination of semantic relations is ruled out, given the "standard" way of interpreting the diagrams. (How is it ruled out?)

Now consider Figs. 2a and 2b, each representing a configuration composed of two horizontal rigid levers, centrally pivoted and joined by an unstretchable string going round a pulley with fixed axle. (A deeper analysis of this example would require a much more elaborate and explicit statement of the semantic rules.) If the arrows represent direction of motion of ends of levers, then it is

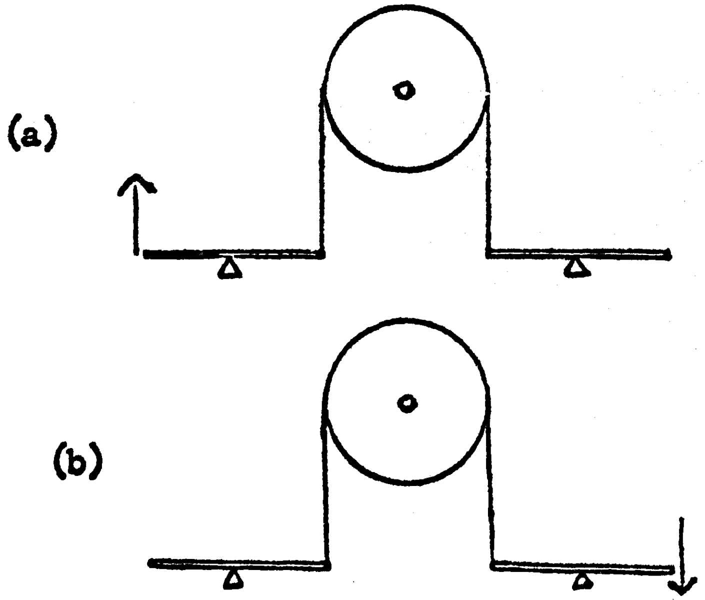

Fig. 2

impossible for any situation to be represented by (a) unless it is also represented by (b), even though whether a particular situation is or is not represented by each of the figures is a matter of fact. Thus the inference from (a) to (b) is valid. Anyone who does not find this immediately obvious may be helped by being shown figures with arrows in intermediate positions, as in Fig. 3. This is analogous to the use of a sequence of intermediate steps in a logical proof to help someone see that the conclusion does follow from the premisses: one person may require such intermediate conclusions though another does not. (It would be of some interest .o discuss the case of a person who understands each step, but cannot grasp the proof as a whole—but space limitations prevent this. Problem: how do we know where to insert the intermediate arrows?)

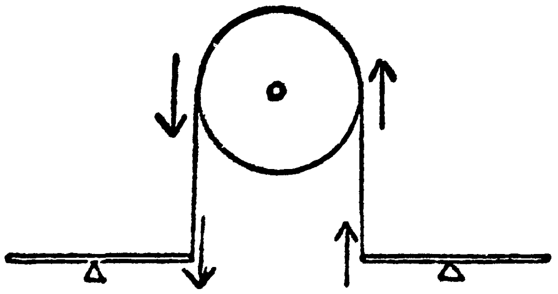

Fig. 3

3. Generalizing the Concept of Valid Inference

What these two examples illustrate, is that the concept of a valid inference, or a valid step in a proof, can be generalised in two ways beyond the definition given above. *First* the inference may involve non-verbal representations, instead of only a set of statements. *Second*, validity need not concern only truth-values, but also represented objects, configurations, processes, etc. Thus, the inference from representations $R_1, R_2, \ldots R_n$ to R_c is *valid* in the generalised sense if structural (syntactic) relations between the representations and the structures of the semantic interpretation procedures make it impossible for $R_1, R_2, \ldots R_n$ all to be interpreted as representing anything which is not also represented by R_c. We can also express this by saying that R_c is jointly *entailed* by the other representations.

In this sense (a) and (b) of Fig. 1 together entail (c). Similarly (a) in Fig. 2 entails (b). Explicitly formulating the interpretation rules relative to which these inferences are valid would be a non-trivial exercise. Once they have been made explicit, the possibility arises of indicating for any valid inference exactly which are the rules in virtue of which the step from "premisses" to "conclusion" is valid. When a proof contains such explicit indications it is not merely valid but also *rigorous*. So far, relatively few representational or linguistic systems are sufficiently well understood for us to be able to formulate proofs which are rigorous in this sense. For instance, we can do this for some of the artificial languages invented by logicians, in which various logical symbols are *defined* in terms of their contribution to the validity of certain forms of inference (e.g., the rule of "universal instantiation" is part of the definition of the universal quantifier). But the fact that we do not yet understand the semantics of other languages and representational systems well enough to formulate rigorous proofs does not prevent us from recognising and using valid proofs. Similarly, it need not prevent a robot.

I conjecture that much intelligent human and animal behaviour, including the phenomena noted by Gestalt Psychologists, involves the use of valid inferences in non-linguistic representational systems, for instance in looking at a mechanical configuration, envisaging certain changes and "working out" their consequences. The use and manipulation of rows of dots, or sticks of different lengths, to solve arithmetical problems, instead of the manipulation of equations using numerals and such symbols as "+" and "−" is another example. What philosophers and others have been getting at in talking about our ability to "intuit" or "see" connections between concepts or properties can now be interpreted as an obscure reference to this generalised concept of validity. (My own previous effort [20] was also obscure.) One of the sources of confusion in such discussions is the fact that although we sometimes use and manipulate "external" representations, on paper or blackboards for instance, we also can construct and manipulate diagrams and models "internally", i.e., in our minds. This has led to a certain amount of mystique being associated with the topic. By placing the topic in the context of A.I., we can make progress without being side-tracked into the more fruitless variety of philosophical debate about the ontological status of mental processes, for the ontological status of the internal manipulations within a computer is moderately well understood.

There are, of course, many problems left unsolved by these remarks. For instance, there are problems about the *scope* of particular inference patterns: how far can they be generalised, and how does one discover their limits? (Compare I. Lakatos [10], and S. Toulmin's discussion in [22] of the use of diagrams in optics.) More importantly, does the ability to generate, recognize and use valid inferences require the use of a "metalanguage" in which the semantic and syntactic relations can be expressed and which can be used to characterise inferences explicitly as valid? Many persons can recognize and use valid inferences even though they have learnt no logic and become incoherent when asked to explain why one thing follows from another: does this imply that we unwittingly use sophisticated metalinguistic apparatus long before we learn any logic? Is *social* interaction required to explain how we can learn the necessary consequences of semantic and syntactic rules? These deep and difficult problems arise as much in connection with the use of language as in connection with the use of non-linguistic representations, so I have no special responsibility for answering them merely because of my defence of non-linguistic systems as having an autonomous status not reducible to the status of heuristic *adjuncts* to linguistic ones.

4. Analogical vs. Fregean Modes of Representation

How should one decide which sort of representational system to use in connection with a given problem? It may be impossible to give a useful general answer to this question, but I shall try to show that for certain sorts of problems "analogical" systems have advantages over general languages like predicate calculus. If this is so, then the hunt for *general* problem-solving strategies and search-reducing heuristics may prove less fruitful than the study of ways in which highly *specific* topic-dependent modes of representation incorporate rich problem-solving powers in their very structures. Contrast Hayes [8]:

> ... for the robot, generality is all-important, and powerful—problem dependent—heuristics just will not be available.' (p. 536)

Clearly it will depend on the robot: and why should we aim to design only robots whose *general* intelligence surpasses that of humans and other known animals?

In order to make all this more precise we need an analysis of the linguistic/non-linguistic distinction which I have hitherto used without explanation. Detailed investigation shows that there is a whole family of distinctions to be explored. For the moment, I shall merely explain the contrast between "analogical" and "Fregean" modes of representation. Pictures, maps and scale models are largely analogical, while predicate calculus (invented by Frege), programming languages and natural languages are largely, though not entirely Fregean. The contrast concerns the manner in which the parts of a complex representing or denoting configuration, and relations between parts, contribute to the interpretation of the whole configuration, i.e., the manner in which they determine what is represented, expressed, or denoted.

In an *analogical* system properties of and relations between parts of the representing configuration represent properties and relations of parts in a complex represented configuration, so that the structure of the representation gives information about the structure of what is represented. As two-dimensional pictures of three-dimensional scenes illustrate, the correspondence need not be simple. For instance, in Fig. 4 distances in the picture represent distances in the scene in a complex context-sensitive way. Similarly, the relation "above" in the picture may represent either "above", or "further", or "nearer" or "further and higher", etc., in a scene, depending on context, such as whether a floor, wall or ceiling is involved. Consequently the interpretation of an analogical representation may involve very complex procedures, including the generation of large numbers of *locally* possible interpretations of parts of the representation and then searching for a *globally* possible combination. For an example see Clowes [2]. (The use of search-procedures structurally related to the picture is another example of the use of an analogical representation.)

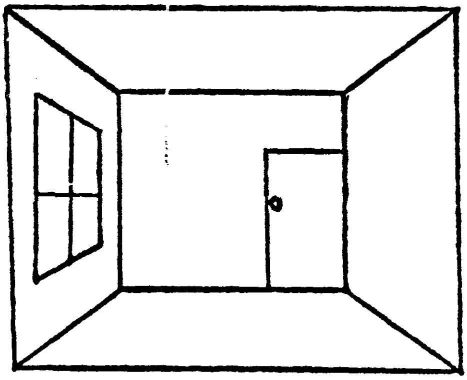

FIG. 4

By contrast, in a Fregean system there is basically only *one* type of "expressive" relation between parts of a configuration, namely the relation between "function-signs" and "argument-signs", (Frege's syntactic and semantic theories are expounded in [5].) For example, the denoting phrase "the brother of the wife of Tom" would be analysed by Frege as containing two function-signs "the brother of ( )" and "the wife of ( )", and two argument-signs "Tom" and "the wife of Tom", as indicated in "the brother of (the wife of (Tom))". Clearly the structure of such a configuration need not correspond to the structure of what it represents or denotes. At most, it corresponds to the structure of the *procedures* by which the object is identified, such as the structure of a route through a complex "data structure". Moreover, the interpretation procedures need not involve the search for a globally consistent interpretation in order to remove local ambiguities, since objects, relations, properties and functions can be unambiguously named by arbitrarily chosen symbols. For instance, the use of the *word* "above" in English need not be subject to the same kind of local ambiguity as the *relation* "above" in Fig. 4. Frege showed how predicates, sentential connectives ("not", "and", etc.) and quantifiers could all be used as function-signs. Consequently, predicate calculus is purely Fregean, as is much mathematical notation. Natural languages and programming languages, however. are at least partly analogical: for instance, linear order of parts of a programme corresponds, to a large extent, to temporal order of execution. (Devices such as "*go to*" which upset this correspondence are neither Fregean nor analogical. These two categories are by no means exhaustive.)

A Fregean system has the advantage that the structure (syntax) of the expressive medium need not constrain the variety of structures of configurations which can be represented or described, so that very general rules of formation, denotation and inference can apply to Fregean languages concerned with very different subject-matters. Contrast the difficulty (or impossibility) of devising a single two-dimensional analogical system adequate for representing political, mechanical, musical and chemical structures and processes. The *generality* of Fregean systems may account for the extraordinary richness of human thought (e.g. it seems that there is no analogical system capable of being used to represent the facts of quantum physics). It may also account for our ability to think and reason about complex states of affairs involving many different kinds of objects and relations at once. The price of this generality is the need to invent complex heuristic procedures for dealing *efficiently* with specific problem-domains. It seems, therefore, that for a frequently encountered problem-domain it may be advantageous to use a more specialised mode of representation richer in problem-solving power. For example, an animal or robot constantly having to negotiate our spatio-temporal environment might be able to do so more efficiently using some kind of analogical representation of spatial structures and processes. A great deal of sensory input is in the form of spatial patterns, and a great deal of output involves spatial movements and changes, at least for the sorts of animals we know about, so the internal decision-making processes involve translation from and into external spatio-temporal configurations. It seems likely, therefore, that the translation will involve less complex procedures, and be more efficient, if the internal representations of actual and envisaged states of affairs, changes, actions, etc., use a medium analogous in form to space-time, rather than a Fregean or other linguistic form.

A great deal more needs to be said about Fregean, analogical and other types of representation or symbol, but I haven't space for an extended survey. Instead I shall now try to describe and illustrate in more detail some ways in which analogical representations may be superior to Fregean or linguistic types.

5. Advantages of Analogical Representations for an Intelligent Robot

An intelligent agent needs to be able to discover the detailed structure of its environment, to envisage various possible changes, especially changes which it can bring about, and to distinguish those sequences of changes which lead to desired or undesired states of affairs. Rumour has it that not all species can do these things in the same contexts: a dog, unlike a chicken, can think of going *round* a barrier to reach food visible on the other side. Similarly, first-generation robots may only have very limited capacities. A minimal requirement for coping with our environment, illustrated by the chicken/dog example, is the ability to consider changes involving relatively smoothly ordered sequences of states, such as going round a fence, turning a knob, moving one end of a stick into the hollow end of another, moving a plank until it bridges a ditch, etc. By contrast, the contexts for intelligent action which appear to have attracted most attention in A.I., such as searching for logical proofs, playing chess, finding a route through a space composed of a network of points and arcs, acting in a world composed of interacting discrete finite automata (compare McCarthy and Hayes [12, pp. 470ff], involve search spaces which have no obvious usable order to organisation, so that in order to make problems tractable new organising patterns have to be discovered or invented and new means of representing them created. Of course, these contexts are very important, and are to be found in our environment also (e.g. assembling a mechanism from general-purpose components). But they are also much more difficult, and attempting to tackle them without first understanding how to satisfy the above minimal requirement may be unwise.

For example, here are some problems which we (who? chimps? two-year old children?) seem able to solve effortlessly when the problem is represented analogically, but which sometimes become much more difficult in a different format (e.g., an arithmetical format, using equations and co-ordinates, etc.). In Fig. 2a which way is the right-hand end of the right-hand lever moving? In Fig. 5, where *A* represents the dog, *B* the food, and the dashed line a fence, find a representation of a route from dog to food which does not go through the fence. (Noticc that this requires a grasp of how the latter relation is represented analogically.) In Fig. 6, where *AB* represents a ditch, *CD* a movable plank, find a way of moving the plank until it lies across the ditch. In Fig. 7, representing rail connections between towns, find a route between the two asterisked towns passing through the smallest number of other towns. In Fig. 8, where the lines represent rigid rods lying in a plane, loosely jointed at *A, B, C, D* and *E*, what will happen to the angle *CDE* if *A* and *D* move together? Our ability to solve such problems "easily" (and many more examples illustrating this could be given) seems to depend on the availability of a battery of "subroutines" which we can bring to bear on parts of spatial configurations, transforming them in specific ways representing changes of certain sorts in *other* configurations.

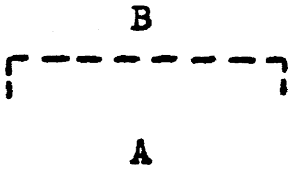

Fig. 5

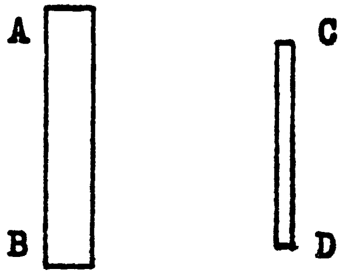

Fig. 6

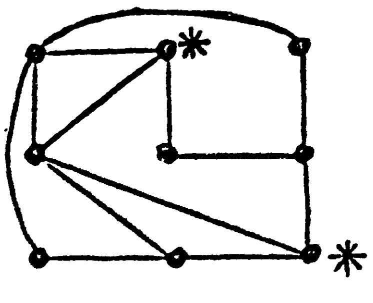

Fig. 7

For instance, while looking at or thinking about some configuration, we can imagine or envisage rotations, stretches and translations of parts, we can imagine any two indicated parts joined up by a straight line, we can imagine X moving in the direction in which Y is from Z, we can imagine various deformations of a line while its ends are fixed, such as bending the line sideways until it passes through some third specified point or until it no longer crosses some "barrier". For example, something like this last procedure could be used to find the route round the fence (see Fig. 5), or even a route round a number of barriers—though more than just bending of a straight line may be required if some of the barriers are bent or curved. A similar routine might be used to find the best route between asterisked points in Fig. 7, though more complex procedures are required for more complex route-maps.

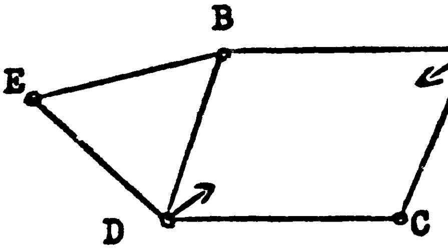

FIG. 8

Of course we cannot always do the manipulations in our heads: we may have to draw a diagram on paper, or re-arrange parts of a scale model, in order to see the effects. (Try imagining the motion of a worm and pinion without moving your hands). The difference between performing such manipulations internally and performing them externally is irrelevant to our present concerns. The main point is that the ability to apply such subroutines to parts of analogical configurations enables us to generate, and systematically inspect, ranges of related possibilities, and then, in the generalised sense of "valid" defined previously, to make valid inferences, for instance about the consequences of such possibilities. Thus, in the situation represented by Figure 2a without the arrow, one can find the movement of one lever required for producing desired movement of the other. (Of course, this leaves many unsolved problems, such as how the appropriate manipulations of the reoresenting configuration are selected and how the solution to the problem can be translated into action.)

What these examples seem to illustrate is that when a representation is analogical, small changes in the representation (syntactic changes) are likely to correspond to small changes in what is represented (semantic changes) changes all in a certain *direction* or *dimension* in the representation correspond to similarly related changes in the configuration represented, and constraints in the problem situation (the route cannot go through the fence, the rods cannot stretch or bend, the centres of the pulleys are fixed, etc.) are easily represented by constraints in the types of transformations applied to the representation, so that large numbers of impossible strategies don't have to be explicitly considered, and rejected. Hence "search spaces" may be efficiently organised. By contrast, the sorts of changes which can be made to a Fregean, or other linguistic, description, such as replacing one name with another, conjoining a new description, applying a new function-sign (such as "not-") to the description adding a qualifying phrase, etc., are not so usefully related to changes in the structure of the configuration described. (One can, of course, impose analogical structures on a Fregean system through the use of certain procedures: for instance if names of a class of individuals, are stored in an order corresponding, say, to the order of size of these individuals, then substituting those names in that order in some description, as part of a search, would be similar to the above manipulations of analogical representations. Contrast the non-analogical case where instead of the ordered list, there is a randomly ordered list of *statements* asserting for each pair of individuals which of the two is smaller in size.) For example, "failure of reference" is a commonplace in Fregean and linguistic systems. That is, very many well-formed expressions turn out not to denote anything even though they adequately express procedures for identifying some individual; for example "the largest prime number", "the shape bounded by three straight sides meeting in four corners", etc. (This topic is discussed in my [21].) It seems that in an analogical system a smaller proportion of well-formed representations can be uninterpretable (inconsistent): pictures of impossible objects are harder to come by than descriptions of impossible objects, so searches are less likely to be wasteful.

A most important economy in analogical systems concerns the representation of identity, or coincidence in complex configurations. Each part of a map is related to many other parts, and this represents a similar plethora of relationships in the region represented by the map. Using a map we can "get at" all the relationships involving a certain place through a single access point, e.g. the dot representing a town whose relations are in question. By contrast, each part of the region would have to be referred to many times, in a large number of statements, if the same variety of information were expressed in linguistic descriptions. (Thus additional semantic rules for identifying different signs as names of the same place are required.) Moreover, a change in the configuration represented, may, in an analogical representation, be indicated simply by moving a dot or other symbol to a new position, whereas very many changes in *linguistic* descriptions of relationships would be required.

Finally, when we use a Fregean or similar language, it seems that our ability to apply names and descriptions to objects in the world has to be mediated by analogical representations. For instance, one can define a word such as "plank" in terms of other words, such as "straight", "parallel", "wooden", etc., but eventually one has to say of some words, to a person who claims not to understand them, "You'll just have to learn how things of that sort *look*". Similarly, any robot using such a language will have to relate it to the world via analogical representations of some sort. So even when deliberation about what to do, reasoning about problems, etc., uses Fregean languages, analogical representations are likely to be lurking in the background, giving content to the cogitations. If so, it may be foolish not to employ whatever relevant problem-solving power is available in the analogical systems.

What I am getting at is that insofar as a robot has to have at least those types of intelligence, common to humans and other mammals, involved in coping with out spatio-temporal environment, it may need to use analogical representations if it is to cope efficiently. Moreover, it should be remembered that although not as general as Fregean representations, spatial analogical representations are useful for a very wide variety of non-spatial systems of relationships, including all those where we find it useful to talk of "trees", "networks", "hierarchies", "spaces" (e.g. search-spaces!), and so on. So the efficient simulation of our sensorimotor abilities may provide a basis for the efficient simulation of a wide variety of more abstract cognitive abilities. (Compare Piaget's speculations about the role of innate motor schemata in cognitive development, reported in Flavell [3].)

What is now needed is a much more systematic and exhaustive survey of different types of representational systems and manipulative procedures, in order to assess their relative advantages and disadvantages for various sorts of purposes. Some of the ideas in N. Goodman's (6) may prove useful.

6. Summary of Disagreements with McCarthy and Hayes

It should be clear by now that although my thinking on these issues has been considerably influenced by McCarthy and Hayes [12], there are several areas of disagreement, mainly, I think arising out of their neglect of types of representational systems which have not yet been studied by logicians and mathematicians. Where they represent the world as a system of discreet finite automata, I claim that other sorts of representations are more suitable for an environment composed of configurations whose parts and relationships are capable of changing along partially or totally ordered, often continuous, dimensions of different sorts, such as sizes, positions, orientations, speeds, temperatures, colours, etc. Where they analyse the concept of *what can happen or be done* in terms of what is consistent with the interconnections and programs of the automata, I regard this as simply a special case of a more general concept which I call *configurational possibility*, namely the concept of the variety of configurations composed of elements, propࢭrties and relationships of the sorts we find in the world. (A fuller discussion would refer to other categories.) Thinking of all the things in one's present environment which might have been bigger, smaller, a different colour or shape, differently located or oriented, moving at different speeds, etc., illustrates the inadequacy of the discrete automaton representation. (Compare Chomsky's proofs of the inadequacy of certain sorts of grammatical theories, in [1] and elsewhere.) Our ability to notice, and use, such possibilities, apparently shared with other animals, must surely be shared by an intelligent active robot.

Where McCarthy and Hayes require their robot to be capable of "inferring in a formal language that a certain strategy will achieve its goal" (p. 463), I require only that it be capable of recognising a representation of an action or sequence of moves terminating in the goal, and not necessarily a representation in a "formal logical (sic) language" (p. 468). If proof is required that this strategy applied to the assumed existing state of affairs *will* lead to the goal, then a proof within an analogical medium, valid in the generalised sense defined above, will do. Whereas they claim that all this requires "formalising concepts of ability, causality, and knowledge" (p. 463), I claim that it is enough to be able to represent the existing states of affairs, generate (e.g., in an analogical system) representations of possible changes (or sequences of changes), recognize representations of changes which terminate in the goal state, and then attempt to put such changes into effect. There is no need for explicit use of such concepts as "can" or "able" so long as the procedures for generating deformations of representations are geared to what the robot can do. Do dogs and other animals *know* that they cannot do such things as fly over obstacles, push houses out of the way, etc., or do they simply never consider such possibilities in deciding what to do? There is a difference between being able to think or act intelligently and being able explicitly to characterise one's thinking or acting as intelligent. McCarthy and Hayes seem to make the latter a necessary condition for the former, whereas my suggestion is that some of their requirements can be ignored until we are ready to start designing a robot with reflective intelligence.

Of course, a great many problems have been left completely unsolved by these remarks. I have said nothing about how the ability to construct, interpret and modify analogical representations might be programmed. Are new types of computer hardware required if the sorts of subroutines mentioned above for modifying parts of analogical representations are to be readily available? How will the robot *interpret* such routines? How much and what type of hardware and programming would have to be built into a robot from the start in order to give it a chance of learning from experience what its environment is like: e.g. will some knowledge of the form of three-dimensional configurations have to be there from the beginning? Would the ability to cope with some types of *possible changes* in perceived configurations (e.g. motion in smooth curves, rotation of smooth surfaces etc.) have to be programmed from the start in order that others may be learnt?

I cannot answer such questions. What I am trying to do is illustrate the possibility of replacing or supplementing an excessively general and linguistic approach to problems in A.I. with a way of thinking, familiar to some philosophers, involving systematic reflection on facts, about human cognitive abilities, which are readily available to common sense (not to be confused with introspection). By asking, as some philosophers have done "How is it possible for these abilities to exist?", one is already moving in the direction of A.I. The danger is that some people in A.I. pre-occupied with the current technology of the subject and imminently solvable problems may forget or ignore some fruitful new starting points. As for the fear, expressed by Hayes, quoted above, that generality is all-important because powerful problem dependent heuristics will not be available, I hope I have at least given reasons for thinking that they can be made available.

7. Philosophy and Artificial Intelligence

Many philosophical problems are concerned with the rationality or *justifiability* of particular conceptual schemes, sets of beliefs, modes of reasoning, types of language. To reformulate these problems i. terms of the advantages and disadvantages for an intelligent robot of this or that type of conceptual scheme, type of language, etc., will clarify them and, I hope, stimulate the production of theories precise enough to be tested by using them to design mechanisms whose failure to perform as expected will be a sign of weakness in the theories. Attention paid by philosophers to the problems of designing a robot able to use, or simulate the use of, much of our conceptual apparatus may introduce much greater system and direction into philosophical enquiries (reducing the influence of fashion and historical accidents such as the discovery of paradoxes). I have tried to show, for example, how thinking about the problem of designing a robot able to perceive and take intelligent decisions helps to put logic into a broader context and brings out the importance of storing information in and reasoning with non-linguistic representations: this has important implications also for philosophy of mathematics and philosophy of science. (The sketchiness of some of my arguments is connected with the fact that this paper is part of a much larger enquiry.) Other philosophical problems (the problem of universals, problems about ostensive definition, problems about sense and reference, problems about the relation between mind and body, for example) seem to me to be quiie transformed by fairly detailed acquaintance with progress and problems in A.I. This interaction between philosophy and A.I. may also help to remedy some of the deficiencies (such as inept description and explanatory poverty) in contemporary psychology.

REFERENCES

1. Chomsky, N. *Syntactic Structures*. Mouton, The Hague, 1957.
2. Clowes, M. B. On Seeing Things. *Artificial Intelligence*. 2 (1971), 79.
3. Flavell, J. H. *The Developmental Psychology of Jean Piaget*. van Nostrand, 1963.
4. Feigenbaum, E. A. and Feldman, J. (eds.). *Computers and Thought*. McGraw-Hill, 1963.
5. Frege, G. *Translations from the Philosophical Writings of Gottlob Frege*, by Geach, P. and Black, M. Blackwell, 1960.
   The following are also relevant: *Foundations of Arithm tic*, transl. by Austin, J. L. Blackwell, 1953.
   The Basic Laws of Arithmetic transl. (with a useful introauction) by Furth, M. University of California Press, 1964.
6. Goodman, N. *Languages of Art*. Oxford University Press, 1969.
7. Gelernter, H. Realisation of a geometry-theorem proving machine, in [4].
8. Hayes, P. Robotologic. *Machine Intelligence* 5 Meltzer, B., Michie, D. (Eds.). Edinburgh, 1969.
9. Kant, I. *Critique of Pure Reason*, transl. by Smith, N. K. Macmillan, 1958.
10. Lakatos, I. Proofs and refutations, in four parts, *Brit. J. Phil. Sci.* 14 (1963–4); 1–25, 120–139, 221–245, 296–342.
11. Lindsay, R. K. Inferential memory as the basis of machines which understand natural language, in [4].
12. McCarthy, J. and Hayes, P. Some philosophical problems from the standpoint of Artificial Intelligence. *Machine Intelligence* 4. Meltzer, B., Michie, D. (Eds.). Edinburgh, 1969.
13. Minsky, M. L. Steps toward artificial intelligence, in [4].
14. Minsky, M. L. (ed.). *Semantic Information Processing*. M.I.T. Press, 1968.
15. Minsky, M. L. Introduction to [14].
16. Minsky, M. L. Descrptive languages and problem solving, in [14].
17. Minsky, M. L. Form and content in computer science. ACM Turing Lecture. *J.A.C.M.* 17 (April, 1970), 197–215.
18. Mueller, I. 'Euclid's *Elements* and the axiomatic method'. *Brit. J. Phil. Sci.* 20 (December, 1969), 289–309.
19. Raphael, B. A computer program which "understands", in [14].
20. Sloman, A. Explaining logical necessity. *Proc. Aristotelian Soc.* LXIX (1968–9), 33–50.
21. Sloman, A. Tarski, Frege and the Liar Paradox, *Philosophy* XLVI (April, 1971), 133–147.
22. Toulmin, S. *The Philosophy of Science*. Hutchinson, 1953; Grey Arrow paperbacks, 1962.
23. Wittgenstein, L. *Remarks on the Foundations of Mathematics*. Blackwell, 1956.