Naturwissenschaften 70, 282–287 (1983) © Springer-Verlag 1983

# Relation between Neurophysiological and Mental States: Possible Limits of Decodability

Alfred Gierer
Max-Planck-Institut für Virusforschung, D-7400 Tübingen

Validity of physical laws for any aspect of brain activity and strict correlation of mental to physical states of the brain do not imply, with logical necessity, that a complete algorithmic theory of the mind-body relation is possible. A limit of decodability may be imposed by the finite number of possible analytical operations which is rooted in the finiteness of the world; it is considered as a fundamental intrinsic limitation of the scientific approach comparable to quantum indeterminacy and the theorems of logical undecidability. An analysis of these limits, applied to dispositions of future behaviour, suggests that limits of decodability of the psycho-physic relation may actually exist with respect to brain states with self-referential aspects, as they are involved in mental processes. Among possible empirical approaches to such aspects, studies on a class of "metatheoretical" jokes may be helpful which suggest that our brain is capable of immediate perception of hidden inconsistencies resulting from self-applications of concepts and logical operations. Limits for an algorithmic theory of the mind-body problem suggested by this study are formally similar to other intrinsic limits of the scientific method such as quantum indeterminacy and mathematical undecidability: they are related to self-referential operations. Hard sciences, despite their reliability, universality and objectivity, depend on metatheoretical presuppositions which allow for multiple philosophical interpretations.

## The Mind-Body Relation as a Scientific Problem

A motivation for studying the physical basis of biological phenomena in general, and of brain functions in particular is to approach a better understanding of consciousness and the "self"; and yet, the mind-body problem is somewhat repressed in the scientific community because our deep interest in it is in conflict with its conceptual evasiveness. It is difficult even to state explicitly what the main problems are. Some philosophers claim that the problem disappears upon careful epistemological analysis, whereas others consider it an unresolvable feature of human nature. However, there are aspects of the problem which can be stated with some clarity and for which possible solutions, and solubility as such, can be rationally investigated. One of these aspects is discussed in this paper: The relation between mental and neurophysiological states.

Mental states – emotions, intentions, dispositions, motivations etc. – are given to our consciousness directly, often without mediation of our senses, and generally without knowledge of concurrent brain activity. They can be expressed in gestures or words. A few "raw feelings" have inborn expression such as laughter, smiling, and weeping: on this elementary basis cultural learning leads to a highly complex repertoire, mostly verbal, for the expression of mental states. A dictionary of synonyms such as "Rogers Pocket Thesaurus", listing words belonging to various domains, shows that about half of the 15000 words belong to the mental, and half to the physical domain. Though the attribution to one or the other domain is disputable in many cases, even a very restrictive and critical assessment of what is mental would leave us with some thousand words. In combinations, they permit an almost infinitely subtle expression of our mental state, far beyond the "raw feelings" inferable by simple physiology like blushing or a trembling voice.

What is the relation of these mental states to the neurophysiological state of the brain? There is

abundant empirical evidence that a relationship exists, including correlations of brain activity with emotional expressions, as well as effects of neurosurgery, ablations by accidents, localized electric stimulation and the action of drugs on mental conditions. The idea that mental states are correlated to physical states is supported by theorems relating thinking and physical processes. It has been shown that any formalizable mental procedure can be realized by digital computers [1]. This theorem does not apply directly to brains because the latter do not operate on a digital basis. However, their element, the nerve cell, has capacities of processing information surpassing and including those of a digital element. Therefore, any formalizable process should be within the scope of suitably constructed neural networks. This consideration does not show how the brain actually works and what its limitations are, and they leave open the possibility that there are brain properties which are not formalizable; but they support the general notion that the mental state is strictly correlated to the physical state of the brain. This seems to suggest that the mental states occur as a sequence in time parallel, or even epistemologically identical, to brain states [2]. Since the brain obeys the laws of physics, mental states appear to result from a sequence of physical events determined by the laws of mechanics and to be definable by physical expressions under the control of the nervous system. This is the classical behaviourist viewpoint. It can be restated in a more subtle way by considering mental states as system properties, subsuming classes of physical states; but the implicit conjecture would remain that all mental states can be inferred in principle from the physical states of the brain which are, in turn, determined by the laws of physics.

On the other hand, general arguments have been put forward suggesting that the mind-body relation is unresolvable. The most radical versions postulate that mental states intervene with physical states of the brain, say neural circuits, causing processes to occur which would not occur without intervention. Recent versions of such interactionist theories have been proposed by Popper and Eccles [3], and Sperry [4]. Intervention challenges the validity and completeness of physical laws in relation to brain activities, and thus the universality of physics. This universality cannot be established on purely logical grounds, but empirically physics has proven to be fully applicable to biological problems wherever it has been tested. The living domain has properties not found in inorganic nature; this, however, is not because the laws of physics are suspended, but because, generally, systems of components have properties that the components themselves do not have and which can be related to elementary physical processes (such as molecular reactions) only by suitable conceptualization. Physics turned out to be the very basis for an understanding of biological principles and facts including those of molecular genetics, the generation of structure and form [5], and the function of neuronal systems. In the absence of any evidence to the contrary, it seems implausible to give up the universality of physics in the biological domain. Therefore, we assume that mental states are related to, and change with, physical states of the brain. This implies that the scope and limits for a resolution of the mind-body problem are to be sought by analysis within the framework of physics and logic. Thus we may question whether mental states can be defined and communicated in an unambiguous manner, and whether they can be analysed without interference by the process of analysis; further, statements made about mental states may change the state after analysis [6]. Apart from the epistemological problems of definition of mental states and their dependence on the analysis, there is the far more general question of whether intrinsic limitations of science and the scientific procedure prevent, in principle, a complete resolution of the mind-body problem. Such intrinsic limitations are given by three aspects of science, quantum indeterminacy, logical undecidability, and the fact that the world (and life) is finite.

## Scope and Limits of the Scientific Method

Quantum indeterminacy implies that future physical states given by position and velocity of particles cannot be calculated completely on the basis of present physical states if events occurring in atomic dimensions are significantly involved. In many cases, quantum indeterminacy does not significantly affect the macroscopic world because objects consisting of many atoms allow for predictions which, though they are statistical in principle, have a very high degree of precision. However, in other cases such as nucleation in meteorology, and sexual recombination of genetic material in biology, evolving macroscopic properties are strongly dependent on individual events within atomic dimensions: The genetic constitutions of future organisms are subject to quantum indeterminacy and are thus not predictable in principle; long-term development of the weather appears to be another example for unpredictable macroscopic features.

Whether quantum indeterminacy also plays a role in brain function is an open question which will not be discussed here. In any case, however, quantum indeterminacy implies that in the long run the physical *environment* of an organism is not determined, that its future is thus open and that a large variety of physical events are consistent with the present physical state. This implies that general statements on the future make sense only if they refer to general systems properties rather than to a particular physical state given by the position and velocity of atoms.

Mathematical decision theory has led to the conclusion that within any formal system which is rich enough to encompass formal logic and arithmetic, there are statements which, though they can be formalized, are not decidable in a finite number of steps within the system; among them the self-referential statement of consistency of the system as a whole. Formalization can be possible in a richer system but the latter now allows for new statements that cannot be decided, among them the question of the consistency of the enriched system. This implies that any formal system depends on unformalized presumptions and no complete internal formalization of a system is possible [7].

Intuitively, it appears that such theorems impose limits to the self-understanding of the human brain. However, as has often been pointed out, no immediate conclusions can be drawn in this respect. Though the number of essentially different possible states of the brain is very large, it is finite in principle whereas the theorems of undecidability refer to infinite numbers of states. From a mathematical point of view a finite number of cases appears as decidable by assessing all possible states individually.

## Finitism as a Fundamental Constraint of Scientific Deduction

However, whereas neither quantum indeterminacy nor limits of decidability appear to impose immediate and direct constraints on the understanding of brain properties, limitations are suggested if, in addition, finitistic aspects are considered [8]. The universe and human life are finite. There are some $10^{80}$ stable atomic particles in the universe (within the range of the physically possible observation), and the universe is of the order of $10^{40}$ elementary times old (events shorter than the elementary time would interfere with the stability of elementary particles). No real number of operations can surpass an upper limit of $10^{80} \times 10^{40} = 10^{120}$, and a realistic estimate would be much below this number. On the other hand, such large numbers occur easily, on a combinatorial basis, as the number of *possible* states, such as the number of possible combinations of genes in the progeny of an organism, of possible combinations of words on a single page, of possible combinations of people in one room, or, closer to our context, of different possible states of a single brain. General statements which are true for all possible combinations in a given case then require a finite algorithmic theory to be proven. A mathematically finite relation may be physically meaningless if the number of steps required to reach the decision is above $10^{120}$ and thus above physically reasonable numbers. A general statement subsuming a sufficiently large number of possible cases may not be decidable even if a decision is possible for any individual case. Moreover for complex systems the number of possible true statements themselves surpasses any physically reasonable number. Therefore there may be statements which can be proven once they are proposed but which cannot be discovered in an algorithmic way; such statements can be found only by chance and intuition, perhaps with negligible probability.

Since the finite size and time scale of the universe is part of physics we consider the limitation of possible operations as a limitation of the scientific approach which is as fundamental as quantum indeterminacy and mathematical undecidability. This concept can be restated in a negative form: It does not make sense, in principle, not even as part of thought experiments, to look at the world as would an imaginary supercosmic computer.

## The Relation between Physical and Mental States May not be Fully Decodable

This finitistic aspect may have implications for the relation between complex mental states (which could be expressed by the combinations of words of the mental domain) and the physical state of the nervous system describable by the connectivity and activity of the neural network in the brain. Despite the presumption that there is stringent correspondence of mental to physical states, the psychophysic relation need not be fully resolvable by physically reasonable limited algorithms.

Examples of mental states which lend themselves to analysis in this context are dispositions towards future behaviour, which, in turn, depends on future external events. Mental dispositions refer to an open future because quantum indeterminacy implies that the future physical environment of a

person is in the long run undetermined in principle. The number of possible future environments in physical terms of positions and velocities of matter is beyond finitistic limits. It follows that meaningful dispositions for future behaviour cannot refer to individual sequences of physical events in terms of positions and velocities of matter, but only to future features, circumstances and events described in general terms, each subsuming a large variety of physically different states. For example, the present disposition of a person for emigration in the future consists of a list of conditions of future events and circumstances affecting the choice whether and where to emigrate; the disposition for a change of profession is equivalent to a description for choices depending on future circumstances. The question of whether the physical state of a human brain corresponds to a given behavioural disposition cannot be decided by testing all possible real cases of the future physical environment in terms of basic laws of physics with respect to the response of the brain because their number is beyond finitistic limits. Further the total number of conceivable dispositions which can be expressed by combinations of words certainly surpasses physically reasonable numbers. There are much more than $10^{120}$ different dispositions each of which is describable in a few sentences. Therefore, it is impossible for a given state of the brain to analyse any conceivable disposition individually even if the disposition, once found, were decidable. Aside from chance and luck, the only way to find a disposition is by an algorithm within finitistic limits, relating brain states and dispositions. There is no logical guarantee that such an algorithm exists in any case. Therefore, from a finitistic point of view, the validity of physics for the brain and the existence of a unique correlation of the mental state of the brain to its physical state does not imply that a complete algorithmic theory for the psychophysic relation is possible. There may be mental states that cannot be inferred from the state of the brain by physically reasonable, finitistic procedures.

This limitation would not be of much significance if the undecidable features were restricted to complex details of little general interest. However, the finitistic aspect may restrict our possible insight with respect to central features of mental states involving consciousness. These are often self-referential. The self is represented in the brain in an abstract manner and it is this representation which is involved in assessing behavioural strategies and acquiring behavioural dispositions. A simple example is the representation of the position of the "self" in the brain while the body moves in a dark room. Since we assume that mental states correspond to physical states, the development of strategies aimed at self-referential mental states (such as happiness, self-respect etc.) require the representation of actual and potential brain states in the brain. Brain states and corresponding mental states with self-referential aspects are of particular interest because they provide an analogy to self-referential statements within formal systems such as the statement of internal consistency. Mathematical decision theory has shown that for formal systems subsuming an infinite number of possible cases there are statements with self-referential features (of the type describable as "I am consistent") which can be expressed within the system but cannot be proven by a mathematically finite number of steps. Statements about a number of cases which, though mathematically finite, surpasses, irreducibly, physically finitistic numbers such as $10^{120}$ cannot be decided by testing each case individually; therefore, there may be undecidable statements analogous to those referring to an infinite number of cases as treated by the mathematical theorems of undecidability. We therefore assume that for the complex system of the human brain, with a number of possible states exceeding, irreducibly, finitistic limits, there are statements with self-referential aspects which are not decidable by procedures within physically finitistic limits. Although many aspects of mental states can be inferred from neurophysiology, behaviour or both there may be basic mental properties that cannot be deduced from brain states by a finitistic algorithmic procedure, or cannot be proven even if found by chance; this is expected, in particular, if self-reference and self-representation are crucially involved. Though the mental state appears to be uniquely correlated to the physical state of the brain, the mental state may not be *decodable* in all aspects on the basis of the physical state by finitistic procedures; knowledge of the mental state by conscious experience and direct verbal expression then enriches our possible knowledge in comparison to knowledge of physical parameters alone.

## The Metatheoretical Joke: An Indicator for Inconsistent Brain Functions

Possible limits of decodability might be analysed by future theoretical and empirical research. Progress in neurophysiology and psychology may obviously be helpful in this direction. Beyond this,

two possible approaches towards a better understanding of self-referential aspects of the brain will be mentioned. One might give us theoretical insights into the manner in which the brain handles inconsistencies occurring in self-referential operations. Human thought is not intrinsically consistent. In particular, there is no stringent provision against ambiguities and contradictions occurring in self-referential operations that are forbidden in formal logic. It appears rather that in the brain there is a rapid detection mechanism for some types of inconsistencies: Contradictions and inconsistencies have been postulated in the process of making a situation conscious. If the contradiction is evaluated as dangerous, the result may be increased attention, fear or even panic. If it is classified as not dangerous, the reaction is relief and often a smile or laughter. The latter type is exemplified by a class of jokes which cover a significant part of what is considered funny: They contain statements which are metatheoretically inconsistent, in that some conceptual self-application leads to a hidden contradiction. A few examples: "Nothing is inexcusable except a poor excuse". A version of the liar paradox is a man aged 31 claiming "Never trust anybody above 30". Nasredin Hodscha, appointed to be a judge, is told the case by one party in a suit, and comments "You are right". Listening to the other party, he concludes "You are also right". Both parties now argue, unisono, "It is not possible that we are right, and they are right". Nasredin Hodscha concludes "You are right". Metatheoretical jokes can also have a pictorial basis. An example is a cartoon of a manifestation of workers carrying empty posters; title: Posterpainters on strike. On the other hand, there seem to be jokes which are experienced as inconsistent though formally they are not. An example is small talk of secret service agents: "Do they know that we know that they know that we have broken their code?".

The perception of such jokes is immediate, occurring in about a second without conscious thought. Children can respond to them with surprising spontaneity. This reaction is somewhat analogous to what Julesz [9] has called "immediate perception" in figure-ground discrimination, in contrast to the detection of features of a pattern requiring formal thought and extensive time. The immediate response to metatheoretical jokes indicates that there are rapid detection mechanisms for inconsistencies resulting from self-referential activities of the brain. The formal structure of this mechanism is of interest in relation to our proposition that possible limits of decodability of the mind-body relation may be related to self-referential aspects incorporated in the brain. One of the empirical approaches to this aspect may be a formal study of metatheoretical jokes in terms of modern linguistic analysis and the corresponding logical structures. For instance, the existence (but lack of reliability) of mechanisms for immediate detection of metatheoretical inconsistencies in the absence of both prevention and (immediate) resolution poses the question as to whether there could be a complete and finitistic algorithm for such detection in terms of neurophysiological states of the brain.

While the prospects for such linguistic approaches are open, one may expect general insights into the mind-body problem if decision theory is extended to systems with a finite number of essentially different states in such a way that physiological data on the complexity of the brain can be introduced and evaluated.

## Is the State of the Mind Determined by the State of the Brain?

The analysis given in this paper does not provide an explanation of, or formal criteria for, the existence of consciousness and the mind, taking them as a most elementary human experience which need not be reducible to still more elementary principles or facts. We maintain, however, that the mental state is strictly related to, and dependent on, the physical state of the brain. Our notion is thus in disagreement with the postulate of Popper and Eccles [3] that mental variables intervene with the physical state of the brain. Nevertheless, our concept is not entirely unrelated to theirs. Popper and Eccles have argued that the mental states postulated to intervene with neural circuits are, in turn, strongly interacting with what they call the "world 3" of abstract ideas created by human minds; "world 3" is thus indirectly involved in the intervention with physical states of the brain. Clearly metatheoretical concepts and procedures with self-referential aspects belong to this domain. In our theory, they intervene neither indirectly nor directly beyond the laws of physics with neurophysiological states, and yet their relation to physical states of the brain may not be fully resolvable by finitistic procedures. Therefore, mental states are given in a more comprehensive way by the inclusion of knowledge on conscious experience, and verbal communication using mental terms, as compared to physical analysis alone.

## Stringent Physicalism Is Consistent with Metatheoretical Pluralism

The limits of decodability of the mind-body problem are proposed to be closely related to self-referential operations, and self-representations, of brain states (which are involved, for instance, in complex dispositions for future behaviour). Limitations of this type are formally related to other intrinsic limitations of the scientific method: quantum indeterminacy is rooted in the limit of measuring the instrument of measurement at atomic resolution, and mathematical undecidability is due to the limitation of a formalized logical analysis of formalized logic. Despite the universality and objectivity of physical and logical analysis, the metatheoretical presuppositions of the scientific procedure cannot be fully resolved, and allow for different interpretations. Quantum indeterminacy is consistent with different philosophies on the relation between reality and knowledge, mathematical undecidability with different notions on the reality of ideas produced by human minds including the concept of the infinite. Along similar lines, we propose that limits for the resolution of the mind-body relation allow for different interpretations with regard to the determination and irreducibility of conscious experience and mental states. If the finitistic approach on which this paper is based is accepted as reflecting fundamental limitations of human knowledge, and determination means deducibility, then mental states may not be fully determined by physical states.

The spectrum of possible interpretation indicates that the metatheoretical presuppositions on which hard sciences are based are more open to different philosophical interpretations than has been assumed in times past when thinking about science was dominated by ideas and ideals of classical mechanics, self-sustained logic, analytical philosophy, and behaviorism. Rather it appears that science is consistent with different metatheoretical interpretations depending on cultural and philosophical presumptions.

I am much indebted to my colleague Dr. P. Whitington for the critical reading of the manuscript.

1. McCulloch, W.S., Pitts, W.H.: Bull. Math. Biophys. *5*, 115 (1943)
2. Feigl, H.: The Mental and the Physical, Minnesota Studies II, p. 370. Minneapolis, Minn. 1958
3. Popper, K.R., Eccles, J.C.: The Self and the Brain. Berlin-Heidelberg-New York: Springer 1977
4. Sperry, R.: Neuroscience *5*, 195 (1981)
5. Gierer, A.: Naturwissenschaften *68*, 245 (1981)
6. MacKay, D.M., in: Man and his Future, p. 153. London: Churchill 1963
7. Stegmüller, W.: Hauptströmungen der Gegenwartsphilosophie. Stuttgart: Kröner 1965ff; Unvollständigkeit und Unterscheidbarkeit. Wien-New York: Springer 1970
8. Gierer, A.: Ratio *12*, 47 (1970)
9. Julesz, B.: IRE Trans. Info. Theor. *8*, 84 (1962)

Received October 11, 1982