Journal of Philosophy, Inc.

Teleology, Error, and the Human Immune System

Author(s): Mohan Matthen and Edwin Levy

Source: *The Journal of Philosophy*, Vol. 81, No. 7 (Jul., 1984), pp. 351-372

Published by: Journal of Philosophy, Inc.

Stable URL: <http://www.jstor.org/stable/2026291>

Accessed: 28/05/2009 15:11

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=jphil>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

Journal of Philosophy, Inc. is collaborating with JSTOR to digitize, preserve and extend access to *The Journal of Philosophy*.

<http://www.jstor.org>

THE JOURNAL OF PHILOSOPHY

VOLUME LXXXI, NO. 7, JULY 1984

TELEOLOGY, ERROR, AND THE HUMAN IMMUNE SYSTEM*

THE subject of this paper is the explanation of natural processes by means of the goals toward which they are directed.¹ We wish to examine a species of such explanation—*error-ascription*. The term 'error' is ours: we use it in order to bring out the similarity between certain explanations of natural processes, used for example by immunologists, and certain explanations used in action theory. Our point is that these explanations as used by immunologists are committed to goal-ascriptions and that the need for explanations of this type signals a system more complex than those to which other goal-ascriptions are applicable. We shall also argue that there is nothing anthropomorphic about the use of error-ascription, nor of goal-ascription in general.

This is our plan. In the first three sections of the paper we shall attempt to show how the phenomena that interest us are relevant to teleology. Section I is primarily expository: in it we describe certain key features of the human immune system and what we call its "errors." Our language in this section is teleologically loaded, but this is for expository convenience only—we attach no importance to the mere availability of teleological concepts for use in describing our phenomena. In section II we discuss two representative attempts to reduce explanations by goals to efficient-causal explana-

* An earlier version of this paper was read at the Tri-university Conference of British Columbia at Simon Fraser University in February 1982, at the Canadian Philosophical Association Congress at Ottawa in June 1982, where Paul Thompson commented, and at McGill University in October, 1982. We have profited greatly from the discussion at all these occasions. John King-Farlow was very helpful in preparing the final version.

¹ There is another sort of use to which teleology is commonly put, namely, explaining the presence of enduring features or traits by means of the functions they serve—for example explaining the presence of kidneys in mammals by their excretory function. The distinction is sketched in section III, below.

tions (those of Ernest Nagel and Jonathan Bennett). Our aim here is to show that it is difficult to give a satisfactory treatment of error within the frameworks of these analyses. In section III we discuss what error is and, with particular reference to the human immune system, how error-ascription is committed to teleology.

The upshot of the first three sections is that the presence and persistence of error make difficult the reduction of teleological to causal explanation. The analogous point for human psychology is not unfamiliar: one reason for using goal-ascriptions and beliefs to explain action is that error prevents us from positing a direct link between the external occasions of action and action itself.² Why then does this crucial point seem to be overlooked or even suppressed in the accounts of teleology in insensate systems? Primarily, we think, because it is believed that although goal-ascriptions (and hence error-ascriptions) may have pedagogical or heuristic value when applied to nonsentient entities, they have no genuine explanatory worth. In the last two sections of the paper we discuss two elaborations of this attitude.

I. THE HUMAN IMMUNE SYSTEM AND ITS ERRORS

Immunologists have identified the goal of immune systems to be the maintenance of a safe internal environment for the organisms they serve (their "owners"). For example, David Koffler says:

A primary function of the immune system is to distinguish self from non-self and thereby protect the body's internal environment from infection and other challenges from the external environment.³

More picturesquely, we have the following statement by Niels Kaj Jerne:

The special capability of the immune system is pattern recognition, and its assignment is to patrol the body and guard its identity.⁴

²In *Brainstorms* (Montgomery, Vt.: Bradford Books, 1978), p. 63, D. C. Dennett quotes B. F. Skinner attempting to make the direct link: "Young people refuse to get jobs 'not because they feel alienated but because of defective social environments'. A man 'makes a distinction not through some mental act of perception but because of prior contingencies.'" If the social environments are not really defective, but merely perceived to be so, or if the prior contingencies do not really exist, the direct explanation fails, but an indirect explanation via the mental states is still available. [Dennett does not make this point explicitly; rather he protests (correctly) Skinner's belief that the direct link excludes the indirect.] See Adam Morton, "Because He Thought He Had Insulted Him," this JOURNAL, LXXII, 1 (Jan. 16, 1975): 5-15, for points related to error in human psychology.

³"Systemic Lupus Erythematosus," *Scientific American*, CCXLIII, 1 (July 1980): 52-61.

⁴"The Immune System," *Scientific American*, CCXXIX, 1 (July 1973): 52-63.

Obviously, these statements are not intended to have any theological implications, nor do their authors intend to attribute to the immune system a mental life of its own. Rather they are snatched from the pens of sober and distinguished scientists seeking only to describe the immune system and explain its working.

Our claim is that part of the value of these attributions of goals to the immune system is that some of the things that the immune system does can be described and understood as being errors made in trying to achieve these goals. It is with this application of goal-ascriptions that we shall be concerned here. But before we can discuss how error is philosophically significant we must describe in outline how the immune system works and how it errs.

Our quotations tell us that the human immune system seeks to eliminate challenges to the safety of the body's internal environment, whether from unfamiliar molecules, viruses, bacteria, or cells. One of the mechanisms in the process by which it achieves this end involves *antibodies*—these are protein molecules that attach to certain specific three-dimensional structures called *epitopes* or *antigenic determinants* on the surfaces of threatening entities, or *antigens*.⁵ A key event in the immune response is an antibody attaching to an antigen that has the appropriate sort of epitope on its surface—when this happens, the result is a fairly large entity known as an *immune complex*, and this can be removed from the body by the immune, circulatory, and excretory systems.

The removal of antigens depends, then, on there being antibodies around in numbers sufficient to tag and eliminate them. But the body does not keep on hand enough of each antibody that it is capable of producing when the need arises. If an unfamiliar bacillus were suddenly to be introduced into a human body, the chances are that there would not be a sufficient quantity of the appropriate antibody ready to meet the challenge. What would be present is an antibody factory—something called a *B-cell*—that *could* produce the antibody in question. When the bacillus is first introduced, a chain of events is initiated that ultimately results (one hopes) in the production of a large quantity of the appropriate antibody and the consequent elimination of the intruder.

This is what happens. The B-cells (about 10^{12} in number) circulate, each displaying on its surface a sample of the antibody it can produce. When this sample attaches to something, it signals the

⁵ Immune responses are of two kinds: humoral, involving the production of antibodies, and cellular, involving killer T-cells. In this paper we discuss only the former.

presence of an antigen, and so the (possible) need for more production. Accordingly, the immune system as a whole might "help" the B-cell that has signaled. (It may "suppress" it too—more on this in a moment). When helped, the B-cell undergoes cell division, producing many clones of itself; and each clone produces the antibody.⁶ The net result is that the antibody in question is produced in prodigious quantity, and the invader is eliminated.

It is not the case, as one might imagine, that *every* antigen is something to be eliminated. Antibodies exist in an extraordinary variety, and will attach to all sorts of things, *self* tissue included. Now of course the point is to identify and eliminate non-self, and to identify and leave self unharmed.⁷ This is the function of the help-suppress decision. B-cells signal the presence of something they *can* attack; the system decides whether the thing whose presence is signalled *should* be attacked. This decision serves what is called *self-tolerance*.

The list of things that an immune system seeks to destroy (and the list of things it will tolerate) is a characteristic that is, as it turns out, acquired, not innate. To quote Jerne again (55):

You might think that self-tolerance derived from nature's being wise enough to construct the genes coding for your antibodies in such a way as not to give rise to combining sites that would fit epitopes occurring in your own body. It can easily be shown however that this is not so. For example your father's antibodies could recognize epitopes occurring in your mother; some antibody genes inherited from your father should therefore code for epitopes inherited from your mother.

But this would be inconsistent with self-tolerance—a system of this sort would destroy its owner. It follows that self-tolerance must be acquired—probably this process starts at the embryonic stage, and it is helped along by mother's milk.

If the recognition of what is to be attacked is not ensured by the patterns that are pre-programmed into your genes, how is it achieved? It now appears that there are good grounds for regarding the immune system as an immensely complex *network* of reactions. When a B-cell produces antibodies, the immune response does not stop there. These antibodies are themselves antigens which induce the production of further antibodies, which induce the production of yet others, and so on. This proliferating process results ulti-

⁶ It also produces immunological "memory" cells which make a swifter response possible if the antigen is encountered again.

⁷ Cancer may be, in a sense, intermediate between self and non-self, and so complicates the theory. Cancerous tissue carries "self markers"; yet it is normally eliminated by the immune system.

mately in an equilibrium state—what Jerne, the originator of the network theory, has called an “*eigenbehaviour*”, “a dynamic steady state of the system as its elements interact” (59). Whereas self antigens in normal concentrations are part of what contribute to this steady state, foreign antigens and self-antigens in abnormal concentrations “modulate” it by perturbing the equilibrium. When this occurs there is a resultant network of new activity. It is this activity that constitutes the body’s immune response; it ceases when regulation occurs, i.e., when a new equilibrium is reached.

In summary, then, the immune response goes like this. The body is patrolled by B-cells each displaying a sample of the antibody that it is capable of producing. When these B-cells attach to an epitope (through the antibody they display) they may be helped or suppressed—if the process is working as it should, it will be helped if the thing it “senses” is harmful. When helped, the B-cells proliferate and produce antibody in vast quantities. Each antibody attaches to an intruding thing and forms an immune complex which is removed.

We come now to what we shall call *errors* of the human immune system.

(i) In lupus (systemic lupus erythematosus) the immune system *misrecognizes* and, hence, seeks to eliminate its owner’s own DNA. In this disease the actual damage appears to be mediated by the huge amount of immune complex that the body has to remove—for example, the kidneys may become clogged with the stuff, resulting in fatal nephritis.

(ii) In multiple myeloma—a type of cancer—the immune system goes berserk. Instead of reacting only to certain specific circumstances, i.e., when it attaches to an appropriate antigen, it floods the body with antibodies, and these cause a number of severe problems including kidney or liver failure.

(iii) There is a kind of physician-induced encephalomyelitis (inflammation of the brain and spinal cord) in which the immune system attacks its own nervous system. This disorder does not arise from a misrecognition of self (as occurs in lupus and other diseases). Instead what happens is that, for example, an anti-rabies vaccine prepared in a rabbit’s spinal cord sets off an immune reaction, and this response is “cross-reactive” with self. Thus the attack on a correctly identified foe turns out to be an attack on parts of the self. This is because in certain respects parts of self resemble the foe. (Cross-reactivity can also be harnessed for beneficial ends: that the body’s defenses against cowpox are also effective against smallpox is the basis of vaccination against smallpox.)

(iv) In the case of *immune-complex diseases* the recognitional aspect of the immune system might work perfectly well and respond correctly to an invader. But it might then turn out that the immune complex formed by the attachment of antibody to antigen is actually harmful to self tissue. [Note that the antigen is not self-tissue, in contrast with diseases such as those described under (i) and (iii).] Thus in serum sickness the correct immune response to a vaccine—for example, a vaccine to counteract the effects of tetanus toxin—will sometimes result in immune complexes that cause a rash, joint pains, etc.

II. TWO REDUCTIVE THEORIES OF GOAL-AScription

Before we go on to discuss the content of calling the forms of behavior just described *errors*, we should like to show that they constitute puzzling cases for a certain approach to the philosophical analysis of goal-ascription. According to some philosophers, there is nothing wrong with goal-ascriptions as such—but there is nothing distinctive about them either. It is just that talk of goals is natural in a particular sort of causal context. The first of the theories we shall summarize claims that it is the organization of the causal links within a system which is characteristic of entities properly described as goal-directed; according to the second it is the nature of the entities linked which is characteristic.

(a) Let us consider first the system-property view propounded originally by Gerd Sommerhoff and held in various forms by Ernest Nagel and Morton Beckner.⁸

Nagel takes *plasticity* and *persistence* as characteristic of goal-directed processes. (These are the terms used by Nagel. A process is *plastic* if its characteristic outcome can be achieved through a number of different pathways. It is *persistent* if the outcome is achieved in spite of perturbations in the initial conditions of the process.) Nagel traces these characteristics of goal-directed processes to the fact that their outcomes are achieved through the interaction of several causally independent—"orthogonal", as he calls them—mechanisms. Therefore Nagel identifies as goal-directed any process that employs orthogonal mechanisms, and identifies as a goal the characteristic outcome of such a process. Because each *orthogonal mechanism can be described without reference to goals*, Nagel

⁸ Sommerhoff, *Analytic Biology* (New York: Oxford, 1950); Beckner, "Furction and Teleology," *Journal of the History of Biology*, II (1969), reprinted in Marjorie Grene and Everett Mendelsohn, ed., *Topics in the Philosophy of Biology* (Boston: Reidel, 1976); and Nagel, "Teleology Revisited," this JOURNAL, LXXIV, 5 (May 1977): 261–301, reprinted in *Teleology Revisited and Other Essays* (New York: Columbia, 1979). Page references to the articles will be to the books cited.

declares that "there is nothing teleological" in goal-ascriptions (292).

The difficulty that Nagel's view faces is this. In a disease like lupus, the immune system "misidentifies" its owner's DNA as a foreign substance that is to be attacked. Given this misidentification, it goes on to act normally—that is, it attacks what it has identified as non-self. But this ill-grounded attack is both plastic and persistent if any behavior of the immune system is—it is just as complex and just as "ingenious" as normal behavior. So Nagel must say that the destruction of self is the goal of this behavior. Moreover, he must say that the destruction of *non*-self is not a goal of the behavior, since that is not even a possible outcome of what the system is doing. It strikes us that there is something counterintuitive about this position. (We shall elaborate later.)

(b) Jonathan Bennett's theory of teleology⁹ is very different from Nagel's. It is helpful in understanding Bennett first to put his guiding intuitions into mentalistic terms. We can describe a conscious agent's behavior *B* at a certain time *t* as directed toward goal *G* if and only if at *t* that agent believed that *B* would attain *G* and was prepared to do whatever was necessary for attaining *G*. (There is a certain simplification here that does not affect the points at issue.) To extend this conception to nonrational or inanimate entities, Bennett thinks that we should (nonteleologically, nonmentalistically, nonanthropomorphically) identify in these agents a state of "registering" (cf. believing) that *B* will attain *G*, and a state of being "ready" for *G* (cf. being prepared to do whatever is necessary in order to attain *G*). Then, if we can connect these two states to the behavior *B* by a causal law, Bennett will identify the teleological explanation in terms of *G* with the explanation in terms of that causal law. For example, Bennett thinks that there is a state *H* that a polar bear enters when it is hungry (its blood-sugar level is low), a state *R* that it enters when it "registers" that it must kill a nearby seal in order to satisfy its hunger (*R* will include seeing the seal, identifying it as killable, etc.), and further that there is a causal law that connects states *H* and *R* to attacking the seal. To say that the polar bear killed a seal in order to satisfy its hunger is simply to invoke this law to explain the bear's behavior.

How will Bennett deal with phenomena such as lupus behavior? It looks as if he can maintain that it is directed toward self-preservation, because Bennett relies on a law that links the system's *register-*

⁹ Chapter 2 of Jonathan Bennett, *Linguistic Behaviour* (New York: Cambridge, 1976).

ing that a certain behavior will preserve self to the operation of the immune response. And clearly this law does operate in the case we described. But can he accommodate the direction of lupus behavior to self-destruction? To do so he would have to maintain that in certain circumstances the system is ready (cf. wants) to destroy itself, and this seems strange, or at least in need of explanation. At any rate it is, or appears to be, incongruent with the functional characterization of the immune system as concerned with self-protection.

It is possible to summarize the differences between Nagel and Bennett in this way: both are concerned with the goals that explain behavior, but at different levels. Nagel is concerned with what we might call "transient" goals—the goals manifested in a piece of behavior, regardless of whether one would wish to attribute the same goals to the system over longer stretches of time. This is the result of his looking to the causal processes themselves to reveal their goals. Bennett, on the other hand, is concerned with more permanent goals. He is concerned with how these goals explain pieces of behavior, of course, but in discussing the criteria for goal-ascription (54–59), he is concerned with diachronic considerations—not only the actual behavior he wants to explain, but longer-established patterns of behavior. Bennett assumes, in other words, that the teleological explanation of behavior will involve *dispositions* to seek certain goals.

As long as we are concerned with nonerroneous behavior it will be harmless to focus on just one level of goal-ascription, because the transient goal is congruent with the more permanent one. The problem arises when we look at the cases of error recounted above, such as that of lupus, because here two otherwise reasonable intuitions about goal-directed behavior seem to conflict. On the one hand, we feel, with Nagel, that goal-directedness is a characteristic of action, and so we attach importance to the undoubted similarity that exists between the relationship of lupus activity to self-destruction and that between other activities and their goals. On the other hand, Bennett seems right to think that goals are the sorts of things that are entertained dispositionally and so it is imperative that we be able to explain why a thing that is normally oriented toward self-tolerance behaves here to achieve self-destruction.

How does error-ascription help us here? It accepts both the "higher" (and more permanent) goal of rejecting non-self and the "subsidiary" (transient) goal of attacking the owner's DNA, but takes the incongruity between the two as something that needs spe-

cial explanation. This explanation is that the system has a "false belief" (or false registration, to use Bennett's term). The higher goal and the false registration together "imply" the lower goal, and this is why the lower goal is adopted without abandoning the higher. The implicational structure of this explanation has the consequence that we must accord explanatory force to statements of the following sort:

Whenever the immune system is ready to *destroy non-self* and it registers that it must *destroy its owner's DNA in order to destroy non-self* then it is ready to *destroy its owner's DNA*.

This is, of course, a causal law; that is, it causally links states of the immune system. But this statement of the law has an intentional infrastructure—the states it links are presented as having intentional objects (represented by the italicized parts), and the explanatory work of reconciling the system's functions with its transient goals is tied to the logical relationship that holds between these intentional objects and the *falsity* of one of them.

These observations lead us to our main point. Suppose that Nagel and Bennett were right to suppose that it is *just* the causal structure of the above law that contributes to its explanatory force. Then it should make no difference how the law is stated; in particular it should make no difference whether the intentional-state descriptions above were replaced by chemical descriptions of the same states. But such a restatement would reveal no objects that stand in logical relations and none that bear truth values—as a result it would not resolve the conflict in goals that arises in such cases. Intentionality occurs essentially in error-ascriptions. It follows that the explanatory value of error-ascription is not exhausted by the causal law that supports it.

It should be noted that our point does not rely solely on the non-translatability of an intentional statement into a nonintentional idiom. Nagel and Bennett may well concede this.¹⁰ But their reduction may not be intended as a translation; for they may believe that the teleological/intentional element in the causal statement above is irrelevant to explanation as such (though it may be pedagogically or heuristically useful). Nagel and Bennett do not say this explicitly, but the point has been made by some others, such as D. C.

¹⁰ One proponent of the system-property view, Morton Beckner, makes the point explicitly: "My conviction that teleological language cannot be translated into non-teleological language is based on the following hunch: if *A* is a teleological sentence, and if *B* translates *A*, *B* would also be teleological" (*op. cit.*, fn 8, p. 209).

Dennett, who are to some degree sympathetic to the use of the intentional idiom.¹¹ The general reason for this attitude has been succinctly stated by Carl Hempel: Teleology may give us a feeling of being more at home with biological phenomena—the analogy with action does this—but “understanding in this sense is not what is wanted in science, and a conceptual scheme that conveys insight into the phenomena in this intuitive sense does not for that reason alone qualify as a scientific theory.”¹² Accordingly, the reductionist may wish to claim that his causal analysis captures everything scientifically relevant in goal-ascriptions, but not the intentional element. He is satisfied if the goals yielded by his analysis agree with those which our intuitions identify as goals. Our point is that in cases of nonerroneous behavior Nagel and Bennett may well have achieved this modest goal, but that in cases of error they fall short of it because of their failure to take into account the *levels* of goal-ascription.¹³ We have argued that without these levels it is not possible to account for all our intuitions about goal-ascription in cases of error.

Our conclusions thus far are simple. There is a problem in biology which is to be solved by the use of the intentional idiom, including goal-ascriptions. This problem is not an artifact of the in-

¹¹ See D. C. Dennett, “Intentional Systems”, this JOURNAL, LXVII, 4 (Feb. 25, 1971): 87–106, esp. pp. 91–96. Dennett claims that intentional states are attributed to things like computers “as if” they really belonged to these things. He calls this “conceptually innocent anthropomorphizing” and likens it to “taking out a loan of intelligence” which must be repaid by “finding and analyzing away” the various intelligent entities we posit. The implication, we take it, is that teleology or, more generally, intentionality is useful only insofar as it helps us discover causal connections.

¹² *Philosophy of Natural Science* (Englewood Cliffs, N.J.: Prentice-Hall, 1966), p. 71.

¹³ It is worth making note of the fact that, according to Bennett, a “teleological law” is applicable *only when* error is possible. For he discusses the case of a certain lake which has the (coincidental) property that, when its level rises to the point where it threatens to inundate some flora, its surface area increases, and so evaporation quickens and the flora are not inundated. On the other hand, when the level falls to the point where it threatens to beach some shellfish, surface area decreases, evaporation slows down, and the level falls no further (75–78).

Bennett admits that one *could* say that the lake registers that it is about to kill wildlife and takes action to prevent this—i.e. that one could produce a “teleological law” true of the lake. But he asserts that this would be unscientific even *faute de mieux*, because “the concept of registration is idle”: to say that the lake’s *registering* that an increase of surface area would help it not to inundate the flora adds nothing to saying that the lake’s increase of surface area would help it not to inundate the flora.

But when would registration *not* be idle? Presumably if the system could register something false or fail to register some true but relevant condition. This is error! It is noteworthy that Bennett is skeptical about the application of teleology to inanimate things like hearts (78). Did he fail to notice (or would he refuse to allow) that some of these things can make mistakes?

tentional/teleological idiom itself, and can be stated without its aid. Just ask: given that the immune system has evolved because eliminating non-self is conducive to survival, why does it ever eliminate self? (One obvious answer—"Because it is diseased and no longer has the properties evolution gave it"—is inadequate. See section IV below.) Our response to Bennett and Nagel is that error provides an intriguing case where reductionist claims must be withdrawn because a nonevasive explanation of this phenomenon makes essential use of the intentional idiom. This conclusion has a corollary for action theory: if biology can accommodate the intentional, surely psychology can as well. Why then should the scientific treatment of human action involve the elimination of intentional description?

III. WHAT IS AN ERROR?

The considerations of the last section prepare us for the following characterization of error. We shall use unabashedly anthropomorphic terminology at first, and discuss later how we can purge this terminology from science.

Error ascription occurs in a context where behavior is explained by reference to practical reasoning, that is, by finding goals and beliefs of an agent the propositional objects of which stand in a certain logical relation to the explanandum behavior (under a certain description). In such a context we often find that goals themselves stand in parallel logical relations—that some of these, called *subsidiary*, can be derived from others, called *higher*, via beliefs. Without going into the controversial details of imperative inference here, we can say that subsidiary goals are those rendered by the belief in question *necessary* for the achievement of the higher goals.

Error occurs when, because of false beliefs (or the lack of relevant true beliefs), an agent adopts a goal as subsidiary when in actual fact it is not necessary for, or actually defeats, the achievement of the higher goal.

The forms of behavior of the immune system we described above fall under this definition. In lupus the system attacks the self in the mistaken "belief" that it is non-self. In cross-reactive encephalomyelitis and immune complex diseases it attacks or harms self because it "does not know" that attacking the foreign vaccine will have these unfortunate side effects. The vaccinated person's response to smallpox is mistaken (though beneficial); the system takes smallpox to be cowpox.

Some salient points about error-ascription may here be noted.

First, it occurs in the context of teleological explanation. This is obvious: only when some behavior is explained in terms of a goal

can it be a problem to understand why that goal was adopted. Error-ascription responds to this difficulty by tracing the adoption of that goal to the falsity of a belief.

Second, error ascriptions explain *actions*, *behavior*, because they invoke the mistaken subsidiary goals of a system to explain unusual behavior. They are therefore to be distinguished from another kind of teleological explanation—functional explanation—which seeks to explain an enduring trait of some organism (e.g., that it has kidneys or an immune system) by means of the goals toward which that trait is (normally) oriented.¹⁴ Reductive theories of functional explanation have aims that are tangential to ours: they seek to identify and give a causal (e.g., evolutionary) account of those goals which explain the presence of enduring traits.¹⁵ Such analyses may be useful to us in identifying higher goals (e.g., that the immune system seeks to eliminate non-self), but they have nothing to say about the issue of how goals, mistaken or otherwise, are explanatory of the activity of organs in an organism.

Third, the explanation of behavior as erroneous, in common with the explanation of more normal behavior in terms of goals, invokes properties of the immune system as a whole, as opposed to the chemical properties of its constituent parts.

This may be illustrated by reference to the cross-reactive encephalomyelitis described in section 1. This disease occurs when the system manufactures antibodies to eliminate a foreign vaccine, but, because of an epitopal similarity between the molecules of this vaccine and molecules in the owner's central nervous system, these antibodies cause damage to that system.

It should be understood that chemically there is little (if any) difference between the laws that govern the antibody's attaching to the foreign stuff and its attaching to indigenous stuff. These laws depend on a certain topological match between antigen epitopes and antibody epitopes, and *ex hypothesi* this match is to be found in both the self stuff and the foreign stuff in question. Thus if the two events were reproduced side by side in a test tube (instead of side by side in an organism) there would be nothing to distinguish their nature. The same could be said for the subsequent chain of events.

Within an immune system, however, things are different. The

¹⁴ See part B of Nagel, *op. cit.*; and Hempel "The Logic of Functional Analysis", in his *Aspects of Scientific Explanation* (New York: Free Press, 1965), for important discussions of this sort of teleological explanation.

¹⁵ For example see Larry Wright, "Functions", *Philosophical Review*, LXXXII, 2 (April 1973): 139-168.

antibody attaching to indigenous cells is abnormal not because it is chemically abnormal, but because it is abnormal for the immune response to result in self-destruction. The self/non-self distinction is a difference that is relevant to understanding a system as a whole, but not relevant to chemistry.

Fourth, there may well be no type-type reduction from statements about an immune system as a whole to statements about chemistry. This is so because antibodies are manufactured in each body under that body's genetic control. Thus there may be differences in the chemical identity of, for example, the antibodies that attack smallpox in two different humans. Moreover, the nature of the equilibrium states of the immune network are a product of the system's history. (This is why immune responses can be "learned".) These individual differences between immune systems have the consequence that a given property of a whole system can have different chemical realizations in different systems and, since self-tolerance is acquired, it can even have different realizations in the same organism at different times. Of course any given property of a system will have a chemical realization, but it does not follow that every *explanation* in terms of systemic properties will have a chemical reduction. To put it differently: A systemic explanation assimilates an event to those in systems with the same systemic properties. But systems with the same systemic properties may not have the same chemical properties. So systemic explanations achieve an end different from chemical explanations.

IV. ERROR AND MALFUNCTION

We have argued that the role of error-ascription is to explain the subsidiary goals of an organ in terms of its higher goals. We now wish to show how this can be important in understanding the functioning of a system. We shall do this by discussing an objection to the application of this sort of analysis to nonsentient systems.

Consider an artificially constructed system such as an electronic fuel-injection system for an internal combustion engine. Such a system might have an exhaust-gas sensor, which measures the oxygen content of the exhaust gases and feeds this information to a central microprocessor, which uses it as one component in a decision-making process about how much fuel is to be injected into the cylinders of the engine.

Now suppose that the exhaust-gas sensor gets damaged, and that as a result it starts providing false information to the control unit. Will it be illuminating to regard the subsequent behavior of the fuel-injection system as erroneous? A critic of our claims might not contest that it is, but he might insist that there is no objective dif-

ference between this way of looking at the matter and considering the fuel-injection system to have been reorganized as to its goals.

To illustrate this let us look at a schema for functional explanation offered by Robert Cummins.¹⁶ Cummins sees the explanandum of functional analysis to be a certain disposition of an organ. The disposition we would be concerned with here is of the fuel-injection system to provide specified amounts of fuel under varying circumstances—i.e., to embody a certain mathematical function from atmospheric pressure, temperature, throttle setting, and exhaust oxygen content, to fuel supplied. According to Cummins this disposition is functionally explained by analyzing it in terms of subsidiary dispositions in the manner of a flow diagram of the sort made familiar by elementary computing texts. At the very simplest level these dispositions will be subsumed to causal laws (757-761).

In the case we described we are concerned with a particular simple disposition, namely that of the exhaust gas sensor to provide certain signals in response to certain levels of oxygen in the exhaust gases. In terms of the Cummins scheme we can represent our claims on behalf of error-ascription like this. We take the flow diagram of a properly working fuel-injection system and explain its faulty action by the *falsity* of the information provided by the exhaust gas sensor. That it is providing false information we in turn seek to explain by the damage that the sensor has suffered. Call this a *top-down* explanation, because it takes the flow diagram (which incorporates the higher goals) as a given.

There is also a *bottom-up* view of the matter. At the simplest level Cummins "subsumes" dispositions to causal regularities. The causal regularities being manifested by the damaged sensor have changed; so its dispositions have changed; consequently so have the dispositions it serves, and so on right to the top—to the flow chart. Looked at in this way, the damaged sensor has caused a change in the flow diagram, and the fuel-injection system is now serving a different function—the mathematical function it once embodied is no longer embodied by it. But, on this view of matters, there is no *false* information; only information differently utilized.

The question is this. Why should we insist on the top-down view? In a conscious agent there *may* be an objective answer to the question of which flow chart was actually used and, consequently, as to the truth or falsity of the information. But what reason is there to take this position with inanimate systems such as the fuel-injection system? The only reason for doing so rests, it might be

¹⁶ "Functional Analysis", this JOURNAL, LXXII, 20 (Nov. 20, 1975): 741-765.

argued, on a misleading analogy between conscious agents and inanimate systems.¹⁷

The importance of this objection cannot be overemphasized. At the end of section II we asserted that error-ascriptions reveal a certain intentional structure in explanations of behavior that appeal to goal-directedness. Cummins can explain this intentional structure—for though he thinks that *ultimately* (i.e., at the level of the simplest dispositions) functional analysis is reduced to causal generalities, he nevertheless sees functional analysis as organizing these generalities in a particular way (i.e., by flow diagrams), and he can account for intentional structure as generated by this organization.

The critic armed with Cummins's analysis wishes to emphasize that in nonsentient systems the flow diagram is determined by the causal generalities, and not *vice versa*. A flow diagram, he says, is just a way of viewing a group of interlocking causal regularities. That there can be many ways of thus viewing these regularities, with nothing much to choose between them, shows, he might say, that *ontologically* we start at the bottom, with the only fixed reality. Teleology is useful, but (in nonconscious systems at least) it is of heuristic value only. Error only illustrates this point.

If all errors were due to malfunctions such a strategy might be plausible. But if we can show that normally functioning systems err, then it is insufficient, because then error has to be accommodated without alteration of the flow diagram determined by the normal causal generalities. In the case of the immune system there are at least two ways in which a *normally* functioning system can provide false information. First, the system is so designed as to act against an epitope, not against an antigen. Thus if it *helps* a B-cell that has made contact with an antigen, the B-cell then in effect attacks anything with a similar epitope. This can lead to difficulties if a self-antigen has an epitope similar to that of a correctly identified foreign antigen. (This is "cross-reactivity" and is involved in the physician-induced encephalomyelitis described above.) Second, the immune system uses the nature of perturbations to its "network" to mediate the help-suppress decision. It is rare that this mechanism fails to provide the right information, but in rare cases even a properly (i.e., normally) functioning network can do this. To use a crude anthropomorphic analogy, the immune system is capable of arriving at justified false beliefs. The difference between the immune system and the electronic fuel-injection system is that

¹⁷ Presumably somebody who insisted on a special status for conscious systems would do so on Cartesian grounds. (See sec. V below.)

in the former case (but not the latter) the causal "sensory" mechanism that leads from the environment to the internal representation of that environment permits the internal representation to be false, *even when the mechanism is undamaged*.

Why does the immune system rely on these imperfect epistemic tools? Is this a sign of imperfect adaptation? If so, should we not say that the function of the immune system is not *literally* the detection of foreign antigens, only approximately so, and that it is only by reference to the goals of an ideal immune system that we identify some behaviors as erroneous? No; because the supposed imperfections in the tools used by the immune system are in fact of immense value. We are not genetically preprogrammed to be protected against some specific predetermined list of invaders. Rather the human immune system is capable of assessing each new antigen it encounters, and reacting to it. Moreover, it is capable of "learning" to respond to harmful antigens. These are immensely valuable features, for they enable us to function in unfamiliar environments. Moreover, the genes that control antibody manufacture are, like other genes, inherited from parents. If the system of a newborn infant were unable to learn to tolerate the new self and reject the new non-self, its mother's antibodies would attack its father's antigens, and the infant would not survive. This is why the characteristics of the new self must be "learned." But this flexibility is bought at the cost of a loss of "certainty" in the system's judgments. Thus a well-adapted and properly functioning immune system is capable of error.

The capacity of well-adapted nonmalfunctioning systems to err is found elsewhere. The nervous system and brain have the task of extracting information about a three-dimensional external world from a two-dimensional pattern of irradiation on our retinas. The "program" that governs visual perception is able to do this, but at the cost of making mistakes when confronted with illusions such as the Müller-Lyer diagram. Such errors seem to be associated with systems that not only "sense", but also process or "interpret" data from the environments in which they operate, because such processing often operates on input that underdetermines the system's output "judgments".

Clearly the ability to make reliable but underdetermined judgments is a sign of a well-adapted system, since without this ability the system might lack the ability to respond to its environment. Yet error is almost by definition a risk that attaches to making such judgments. It would therefore be a gross mistake to treat error *eo ipso* as a sign of malfunction or maladaptation.

We noted earlier that it is essential to teleological analyses that they represent a system intentionally. This can now be clarified in terms of Cummins's suggestion that a teleological analysis views a system as embodying a flow diagram, or *program*. A program is a sequence of instructions; some of these are conditional upon the fulfillment of some boolean-valued antecedent. To view a system as executing a program, we must take some of its internal states as embodying these instructions. Furthermore, to account for the fact that the system executes some of its instructions conditionally, we must also view some of its states as embodying an assignment of truth values to the antecedents of conditionals.

The utility of such a representation is obvious. In the case of an entity like the fuel-injection system discussed above, the program enables us to design and to predict the behavior of the system without knowledge of the actual causal laws that govern its behavior—since the laws of logic that govern instructions will suffice for this task. And with natural systems, we can make, confirm, and disconfirm hypotheses about the working of a system—hypotheses about the program it embodies—long before we are in any position to describe the workings of the system itself. Even if we have a good idea of the causal underpinnings of a system, hypotheses about its “program” help us organize those ideas in a systematic way.

So much is uncontroversial. The question is: can it be *objectively* right to characterize a system in this way? (It is clear that such a characterization can be objectively *wrong*.) In particular, is it really correct to suppose that there is a state embodying informational content which mediates the causal relation between the environment and the system? Though the postulation of such a state might have heuristic value, is it not more consonant with reality to short-circuit this intermediary state and to have direct causal links between environmental conditions and the behavior of the system?

The possibility that a properly functioning immune system can embody false information answers this question. Here an intermediary state is indispensable because the behavior of the system might correspond to a nonactual state of its environment. This is explained by a breakdown in the normal *correspondence* (notice that this is not *eo ipso* a breakdown in the normal causal link) between the internal state that represents the environment and the environment itself. Obviously such an explanation is not available without the existence of the intermediary information-representing state (cf. notes 2 and 13).

What this suggests is that there are degrees of goal-directedness

in systems. Already this was obvious to Cummins, who noted that functional analysis can be applied in a trivial and uninformative way to all kinds of *non-goal-directed* systems and observed:

The explanatory interest of an analytical account is roughly proportional to (i) the extent to which the analyzing capacities are less sophisticated than the analyzed capacities, (ii) the extent to which the analyzing capacities are different in type from the analyzed capacities, and (iii) the relative sophistication of the program appealed to (764).

To this we add (iv) the extent to which capacities in the analysis have inputs and outputs that are information-theoretically characterizable in an informative and nontrivial way, and (v) the extent to which these outputs are underdetermined by the "sensory" input from the environment they represent, provided that any increase in underdetermination is explainable by an increase in utility/adaptation.

On these grounds we can distinguish between the fuel-injection system, the outputs of which depart from what they represent only at the cost of the reduction of the utility of the whole, and the immune and nervous systems. Teleology is more usefully and more objectively applied to the latter.

V. IS TELEOLOGY ANTHROPOMORPHIC?

We have made a case for believing that some forms of behavior of the immune system are best understood as errors and that this involves an ascription of goals to the immune system. As we have presented it, it also involves ascribing "beliefs" to the immune system. These ascriptions strain the credibility particularly of those with Cartesian leanings. How can an inanimate and nonconscious entity have goals and beliefs? Even if representing the immune system in this way is useful, must we not insist on not taking the representation literally?

What exactly is teleology committed to? In the cases we have described, it is committed to at least this: We view certain states of the immune system and certain events in it as *informational*, that is, as correlated (perhaps many-one) to propositions (or, more generally, to information,¹⁸ but we shall ignore this below). Moreover, we regard certain states of the immune system as *goal-directed*, that is as correlated to the propositions that describe the achievements of these goals. Let us therefore postulate an information function and a directedness function: these are partial mathematical functions that take events and states, respectively, into the propositional in-

¹⁸ See Paul M. Churchland, "Eliminative Materialism and Propositional Attitudes", this JOURNAL, LXXVIII, 2 (February 1981): 67-90, esp. pp. 87/8.

formation they represent and the goals toward which they are directed, expressed propositionally.

Now what teleological explanation does in the sort of cases we have been dealing with is to represent the behavior of the system as arising out of the *logical* connections between the values of these functions—that is, it manipulates the propositions that are seen to be “present in” the states of the system in accordance with the laws of logic. This logical manipulation provides a useful alternative to manipulation by causal laws, which may in any case be unknown to us at a certain stage of inquiry; moreover it plays an essential role in understanding error.

Donald Davidson has articulated certain conditions for the applicability of this sort of explanation.¹⁹ He demands (in our terminology) that when this kind of explanation is used with respect to the behavior of a system: (a) there should really be states of the system that have the propositions in question as informational and directedness values, and (b) these states of the system must actually have caused the behavior in question.

That these are conditions for the explanation of the behavior of the immune system can be illustrated by an example of cross-reactivity. Vaccination against smallpox involves introducing cowpox into the human body. The immune system “learns” to reject cowpox. When smallpox later enters the body its epitopal resemblance to cowpox causes the system to attack it.

Now, when the vaccinated body actually rejects smallpox, how are we to explain its behavior? It “senses” the presence of the smallpox virus, and it has the general capacity to reject it—that is, it would have attempted even before the vaccination to reject the smallpox, even though its chances of succeeding would have been much smaller. However, it is not this pre-vaccination ability that actually comes into play in the post-vaccination rejection of smallpox. Rather, another ability does come into play, namely the ability to reject cowpox, via the inability to distinguish between the two invaders.

Thus there are two sets of states that could have functioned in a teleological explanation of this behavior as far as Davidson’s condition (a) is concerned: that one is correctly used rather than the other is due to its being causally involved, that is, due to condition (b). This fact has been taken as evidence that teleological explanation is really causal. But we argued, in section II, above, that this construal overlooks the essential explanatory role played by the in-

¹⁹ See “Actions, Reasons, and Causes,” this JOURNAL, LX, 23 (Nov. 7, 1963): 685–700.

tentional infrastructure of the teleological explanation. In this particular case what needs to be explained is the nonspecificity of the immune system's response—why does one entity, cowpox, create a learned response to another entity, smallpox? The answer refers to the informational content of the states that “remember” cowpox—this content is too nonspecific to distinguish between the two events. Our claim is that this answer could not be “read off” a purely nonintentional description of the causal link (except, of course, by using background information about the informational and directedness values of the states involved). Thus we are inclined to say that the teleological explanation, though irreducible, is not autonomous—it presupposes the availability of a causal story.

We can summarize our conclusions thus far as follows. In an explanation of the following form:

I is ready to reject cowpox.

I registers this smallpox as cowpox.

Therefore, *I* is ready to reject this smallpox.

we attempt to account for the presence of a certain state of a system by means of two other states. Part of the value of this explanation lies, however, in the logical connection between three *other* entities, namely, the propositions intentionally contained in the states. But a condition that governs the applicability of this form of explanation is that the states be causally connected.

So teleological explanation is committed to propositions' being “intentionally contained” in physical states, and it is committed to the causal interaction of these states' paralleling every teleological explanation involving their propositional counterparts. Now is this as such a commitment to the immune system's having beliefs? We think not. For we are inclined to think that, in a Cartesian psychology, beliefs and desires—mental entities—provide the Davidsonian grounding for the teleological explanation of human action, but that Cartesian psychology is inapplicable to immune systems.

Think of it in this way. We use teleological explanation for human action—we explain human actions by the logical manipulation of reasons. If we are right, this means that we are committed to information and directedness functions for some human states. But on what domain are these functions to be defined? Donald Davidson thinks that they are to be defined on a domain of neurophysiological states.²⁰ But a Cartesian psychologist postulates a

²⁰ “Mental Events,” in L. Foster and J. Swanson, eds., *Experience and Theory* (London: Duckworth, 1970).

special domain for these functions, consisting of states and events characterized by consciousness and awareness—call these *thoughts*. Such a psychologist (if he accepts Davidson's argument) would presumably want to say that the teleological explanation of human action presupposes a causal connection between the thoughts that contain the relevant propositions and the action. In particular he might say that whenever we correctly explain a human action in terms of a goal and a proposition concerning how that goal is to be achieved, there is a thought known as a *desire* which intentionally contains the goal and another, known as a *belief*, which intentionally contains the proposition, and that the desire and belief cause the action. Moreover, the Cartesian has a simple account of the information and directedness functions: a thought intentionally contains a certain proposition just in case it is a state of awareness of that proposition.

The teleological explanation of behavior in the immune system is committed to anthropomorphism only if all information and directedness functions *must* be defined on such thoughts. Some philosophers have thought that this condition is satisfied, including notably Brentano, who took intentionality to be "the mark of the mental." But there is, we believe, no good reason to accept Brentano's thesis. If Davidson is right the human versions of these functions are defined on neurophysiological states, not on Cartesian mental events. (Davidson believes, of course, that the correlation between particular neurophysiological states or events and the propositions they "contain" will not be lawlike, but this does not controvert the existence of the functions we have in mind.) Even with humans, then, the explanation of actions by reasons is uncommitted to Cartesianism. Why then should teleology in physiology be so committed? The view that attributing intentional structures to nonsentient entities is anthropomorphic rests, we think, on the misconception that such structures are somehow connected to consciousness—that postulating intentionality in nonsentient entities rests on a metaphorical assimilation of these entities to sentient entities.

Our arguments are thus likely to be most appealing to a certain sort of materialist, namely, one who accepts explanations by reasons but maintains that the having of these reasons is a purely physical property of whatever has them. Such a materialist may or may not allow that humans and other animals are conscious of the reasons they have, or that consciousness is essential to *beliefs* and *desires*. However this may be, the materialist we describe has no reason to insist that every physical state with propositional "content" must be a state of awareness of that proposition. This is what

allows him to view the states of the immune system as having intentional objects, while denying that these are conscious states. This might, of course, entail that these states are not beliefs and desires as such, but what this would show is that teleology does not demand beliefs and desires.

What have we achieved in this paper? We started with the assumption (shared, for instance, by Bennett and Nagel) that it is legitimate to view as goal-directed some processes occurring in nonconscious systems. We then argued that, if this assumption is correct, then it is legitimate to view certain properly functioning systems as directed toward abnormal goals, provided that these systems are of a certain complexity. The abnormality of these goals puts them in need of explanation, and we argued that such an explanation could sometimes be found in the falsity of the information internally represented by the system. This indicates that these systems have internal representations of propositions, some of which are the goals toward which the system is oriented.

Our account sheds light on the work done by goal-ascriptions even in more normal cases—it indicates that these ascriptions permit the use of logical relations between intensional objects to predict and explain the behavior of purely physical systems.

MOHAN MATTHEN

University of Alberta

EDWIN LEVY

University of British Columbia

COMMENTS AND CRITICISM

FACT, FICTION, AND FITNESS: A REPLY TO ROSENBERG*

ALEXANDER ROSENBERG begins his recent article on the concept of fitness with the remark that “debates about the cognitive status of the Darwinian theory of natural selection should have ended long ago.”[†] I agree that this obsession needs to be overcome. But Rosenberg repeats some of the old mistakes and invents epicycles on others. In this comment I will not be able to circumscribe fully the range of topics that an adequate

* I wish to thank the University of Wisconsin, Madison, and the National Science Foundation for financial support and Alexander Rosenberg and Philip Kitcher for comments on an earlier version.

[†] “Fitness,” this JOURNAL, LXXX, 8 (August 1983): 457–473. Page references in the text are to this article.