Mechanism and responsibility

D. C. Dennett

Mechanism and responsibility

I

In the eyes of many philosophers the old question of whether determinism (or indeterminism) is incompatible with moral responsibility has been superseded by the hypothesis that *mechanism* may well be. This is a prior and more vexing threat to the notion of responsibility, for mechanism is here to stay, unlike determinism and its denial, which go in and out of fashion. The mechanistic style of explanation, which works so well for electrons, motors and galaxies, has already been successfully carried deep into man's body and brain, and the open question now is not whether mechanistic explanation of human motion is possible, but just whether it will ultimately have crucial gaps of randomness (like the indeterminists' mechanistic explanation of electrons) or not (like the mechanistic explanation of macroscopic systems such as motors and billiards tables). In either case the believer in responsibility has problems, for it seems that whenever a particular bit of human motion can be given an entirely mechanistic explanation – with or without the invocation of 'random' interveners – any non-mechanistic, rational, purposive explanation of the same motions is otiose. For example, if we are on the verge of characterizing a particular bit of human motion as a well-aimed kick in the pants, and a doctor can show us that in fact the extensor muscles in the leg were contracted by nerve impulses triggered by a 'freak' (possibly random?) epileptic discharge in the brain, we will have to drop the search for purposive explanations of the motion, and absolve the kicker from all responsibility. Or so it seems. A more central paradigm might be as follows. Suppose a man is found who cannot, or will not, say the word 'father'. Otherwise, we may suppose, he seems perfectly normal, and even expresses surprise at his 'inability' to say 'that word I can't say'. A psychoanalyst might offer a plausible explanation of this behavior in terms of unconscious hatred and desires and beliefs about his father, and a layman might say 'Nonsense! This man is just playing a joke. I suspect he's made a bet that he can go a year without saying "father" and is doing all this deliberately.' But if a neurosurgeon were to come along and establish that a tiny lesion in the speech center of the brain caused by an aneurysm

(random or not) was causally responsible for the lacuna in the man's verbal repertory (not an entirely implausible discovery in the light of Penfield's remarkable research), both the analyst's and the layman's candidates for explanation would have the rug pulled out from under them. Since a mere mechanistic happening in the brain, random or not, was the cause of the quirk, the man cannot have had reasons, unconscious or ordinary, for it, and cannot be held responsible for it. Or so it seems.

The principle that seems to some philosophers to emerge from such examples is that *the mechanistic displaces the purposive*, and any mechanistic (or causal) explanation of human motions takes priority over, indeed renders false, any explanation in terms of desires, beliefs, intentions. Thus Hospers says, 'Let us note that the more *thoroughly* and *in detail* we know the causal factors leading a person to behave as he does, the more we tend to exempt him from responsibility.'[1] And Malcolm has recently supported the view that 'although purposive explanations cannot be dependent on non-purposive explanations, they would be refuted by the verification of a comprehensive neurophysiological theory of behavior'.[2] I want to argue that this principle is false, and that it is made plausible only by focusing attention on the wrong features of examples like those above. The argument I will unwind strings together arguments and observations from a surprisingly diverse group of recent writers, and perhaps it is fair to say that my share of the argument is not much. I will try to put the best face on this eclecticism by claiming that my argument provides a more fundamental and unified ground for these variously expressed discoveries about the relations between responsibility and mechanism.

II

The first step in reconciling mechanism and responsibility is getting clearer about the nature of the apparently warring sorts of explanations involved. Explanations that serve to ground verdicts of responsibility are couched at least partly in terms of the beliefs, intentions, desires, and reasons of the person or agent held responsible. There is a rough consensus in the literature about the domain of such explanations, but different rubrics are used: they are the 'purposive' or 'rational' or 'action' or 'Intentional' explanations of behavior. I favor the term 'Intentional' (from the scholastics, via Brentano, Chisholm, and other revivalists), and shall capitalize it to avoid confusion with 'intend' and

its forms, thereby freeing the latter terms for more restrictive duty. *Intentional explanations*, then, cite thoughts, desires, beliefs, intentions, rather than chemical reactions, explosions, electric impulses, in explaining the occurrence of human motions. There is a well-known controversy debating whether (any) Intentional explanations are ultimately only causal explanations – Melden and Davidson[3] are the initial protagonists – but I shall avoid the center of this controversy and the related controversy about whether a desire or intention could be identical with a physical state or event, and rest with a more modest point, namely that Intentional explanations are at least not causal explanations *simpliciter*. This can best be brought out by contrasting genuine Intentional explanations with a few causal hybrids.

Not all explanations containing Intentional terms are Intentional explanations. Often a belief or desire or other Intentional phenomenon (Intentional in virtue of being referred to by Intentional idioms) is cited as a cause or (rarely) effect in a perfectly Humean sense of cause and effect.

(1) His belief that the gun was loaded caused his heart attack
(2) His obsessive desire for revenge caused his ulcers
(3) The thought of his narrow escape from the rattler made him shudder.

These sentences betray their Humean nature by being subject to the usual rules of evidence for causal assertions. We do not know at this time how to go about confirming (1), but whatever techniques and scientific knowledge we might have recourse to, our tactic would be to show that no other conditions inside or outside the man were sufficient to bring on the heart attack, and that the belief (however we characterize or embody it) together with the prevailing conditions brought about the heart attack in a law-governed way. Now this sort of account may be highly suspect, and ringed with metaphysical difficulties, yet it is undeniable that this is roughly the story we assume to be completable in principle when we assert (1). It may seem at first that (1) is not purely causal, for the man in question can tell us, infallibly or non-inferentially, that it was his belief that caused his heart attack. But this is false. The man is in no better position than we to say what caused his heart attack. It may feel to him as if this was the cause of the attack, but he may well be wrong; his only *knowledge* is of the temporal juxtaposition of the events. Similarly, (2) would be falsified if it turned out that the man's daily consumption of a quart of gin was more than sufficient to produce

his ulcers, however strong and sincere his intuitions that the vengefulness was responsible. We are apt to think we have direct, non-inferential experience of thoughts causing shudders, as asserted in (3), but in fact we have just what Hume says we have: fallible experience over the years of regular conjunction.

These explanations are not Intentional because they do not explain by *giving a rationale* for the *explicandum*. Intentional explanations explain a bit of behavior, an action, or a stretch of inaction, by making it reasonable in the light of certain beliefs, intentions, desires ascribed to the agent. (1) to (3) are to be contrasted in this regard with

(4) He threw himself to the floor because of his belief that the gun was loaded
(5) His obsessive desire for revenge led him to follow Jones all the way to Burma
(6) He refused to pick up the snake because at that moment he thought of his narrow escape from the rattler.

The man's heart attack in (1) is not made *reasonable* in the light of his belief (though we might say we can now understand how it happened), but his perhaps otherwise inexplicable action in (4) is. Sentence (5) conspicuously has 'led' where its counterpart has 'caused', and for good reason. Doubts about (5) would not be settled by appeal to inductive evidence of past patterns if constant conjunctions, and the man's own pronouncements about his trip to Burma have an authority his self-diagnosis in (2) lacks.

The difference in what one is attempting to provide in mechanistic and Intentional explanations is especially clear in the case of 'psychosomatic' disorders. One can say – in the manner of (1) and (2) – that a desire or belief merely *caused* a symptom, say, paralysis, *or* one can say that a desire or belief led a person to *want* to be paralyzed – to become paralyzed *deliberately*. The latter presumes to be a purely Intentional explanation, a case of making the paralysis – as an *intended condition* – *reasonable* in the light of certain beliefs and desires, e.g. the desire to be waited on, the belief that relatives must be made to feel guilty.

III

Intentional explanations have the actions of persons as their primary domain, but there are times when we find Intentional explanations

(and predictions based on them) not only useful but indispensable for accounting for the behavior of complex machines. Consider the case of the chess-playing computer, and the different stances one can choose to adopt in trying to predict and explain its behavior. First there is the *design stance*. If one knows exactly how the computer's program has been designed (and we will assume for simplicity that this is not a learning or evolving program but a static one), one can predict the computer's designed response to any move one makes. One's prediction will come true provided only that the computer performs as designed, that is, without breakdown. In making a prediction from the design stance, one *assumes* there will be no malfunction, and predicts, as it were, from the blueprints alone. We generally adopt this stance when making predictions about the behavior of mechanical objects, e.g. 'As the typewriter carriage approaches the margin, a bell will ring (provided the machine is in working order)', and more simply, 'Strike the match and it will light'. We also often adopt this stance in predictions involving natural objects: 'When spring comes new buds will burst on these twigs'. The essential feature of the design stance is that we make predictions solely from knowledge of or assumptions about the system's design, often without making any examination of the innards of the particular object.

Second, there is what we may call the *physical stance*. From this stance our predictions are based on the actual state of the particular system, and are worked out by applying whatever knowledge we have of the laws of nature. It is from this stance alone that we can predict the malfunction of systems (unless, as sometimes happens these days, a system is *designed* to malfunction after a certain time, in which case malfunctioning in one sense becomes a part of its proper functioning). Instances of predictions from the physical stance are common enough: 'If you turn on that switch you'll get a nasty shock', and, 'When the snows come that branch will break right off' are cases in point. One seldom adopts the physical stance in dealing with a computer just because the number of critical variables in the physical constitution of a computer would overwhelm the most prodigious human calculator. Significantly, the physical stance is generally reserved for instances of breakdown, where the condition preventing normal operation is generalized and easily locatable, e.g. 'Nothing will happen when you type in your question, because it isn't plugged in' or, 'It won't work with all that flood water in it'. Attempting to give a physical account or prediction of the chess-playing computer would be a pointless and

herculean labor, but it would work in principle. One could predict the response it would make in a chess game by tracing out the effects of the input energies all the way through the computer until once more type was pressed against paper and a response was printed.

There is a third stance one can adopt toward a system, and that is the *Intentional stance*. This tends to be most appropriate when the system one is dealing with is too complex to be dealt with effectively from the other stances. In the case of the chess-playing computer one adopts this stance when one tries to predict its response to one's move by figuring out what a good or reasonable response would be, given the information the computer has about the situation. Here one assumes not just the absence of malfunction, but the rationality of design or programming as well. Of course the stance is pointless, in view of its extra assumption, in cases where one has no reason to believe in the system's rationality. In weather predicting one is not apt to make progress by wondering what clever move the wise old West Wind will make next. Prediction from the Intentional stance assumes rationality in the system, but not necessarily perfect rationality. Rather, our pattern of inference is that we start with the supposition of what we take to be perfect rationality, and then alter our premise in individual cases as we acquire evidence of individual foibles and weaknesses of reason. This bias in favor of rationality is particularly evident in the tactics of chess players, who set out to play a new opponent by assuming that he will make reasonable responses to their moves, and then seeking out weaknesses. The opponent who started from an assumption of irrationality would be foolhardy in the extreme. But notice, in this regard, how the designer of a chess-playing program might himself be able to adopt the design stance and capitalize from the very beginning on flaws in rationality he knew were built into the program. In the early days of chess-playing programs, this tactic was feasible, but today, with evolving programs capable of self-improvement, designers are no longer capable of maintaining the design stance in playing against their own programs, and must resort, as any outsider would, to the Intentional stance in trying to outwit their own machines.

Whenever one can successfully adopt the Intentional stance toward an object, I call that object an *Intentional system*. The success of the stance is of course a matter settled pragmatically, without reference to whether the object *really* has beliefs, intentions, and so forth, so whether or not any computer can be conscious, or have thoughts or desires, some computers undeniably *are* Intentional systems, for they are

systems whose behavior can be predicted, and most efficiently predicted, by adopting the Intentional stance toward them.[4]

This tolerant assumption of rationality is the hallmark of the Intentional stance with regard to people as well as computers. We start by assuming rationality in our transactions with other adult human beings, and adjust our predictions as we learn more about personalities. We do not *expect* new acquaintances to react irrationally to particular topics, but when they do we adjust our strategies accordingly. The presumption that we will be able to communicate with our fellow men is founded on the presumption of their rationality, and this is so strongly entrenched in our inference habits that when our predictions prove false we first cast about for external mitigating factors (he must not have heard, he must not know English, he must not have seen x, been aware that y, etc.) before questioning the rationality of the system as a whole. In extreme cases personalities may prove to be so unpredictable from the Intentional stance that we abandon it, and if we have accumulated a lot of evidence in the meanwhile about the nature of response patterns in the individual, we may find that the design stance can be effectively adopted. This is the fundamentally different attitude we occasionally adopt toward the insane. To watch an asylum attendant manipulate an obsessively counter-suggestive patient, for instance, is to watch something radically unlike normal interpersonal relations. It need hardly be added that in the area of behavior (as opposed to the operation of internal organs, for instance) we hardly ever know enough about the physiology of individuals to adopt the physical stance effectively, except for a few dramatic areas, like the surgical cure of epileptic seizures.

IV

The distinction of stance I have drawn appears closely related to MacKay's distinction between the 'personal aspect' and the 'mechanical aspect' of some systems. Of central importance in MacKay's account is his remarking that the choice of stance is 'up to us', a matter of *decision*, not discovery.[5] Having chosen to view our transactions with a system from the Intentional stance, certain characterizations of events necessarily arise, but that these arise *rightly* cannot be a matter of proof. Much the same distinction, I believe, is presented in a different context by Strawson, who contrasts 'participation in a human relationship' with 'the objective attitude'. 'If your attitude toward someone is wholly objective, then though you may fight him, you cannot quarrel with

him, and though you may talk to him, even negotiate with him, you cannot reason with him. You can at most pretend to quarrel, or to reason, with him.'[6] Both MacKay and Strawson say a great deal that is illuminating about the conditions and effects of adopting the personal or participant attitude toward someone (or something), but in their eagerness to establish the implications for ethics of the distinction, they endow it with a premature moral dimension. That is, both seem to hold that adopting the personal attitude toward a system (human or not) involves admitting the system into the moral community. MacKay says, in discussing the effect of our adopting the attitude toward a particular animate human body,[7]

> At the personal level, Joe will have established some personal claims on us, and we on Joe. We shall not be able rightly to tamper with his brain, for example, nor feel free to dismantle his body. . . . He has become 'one of us', a member of the linguistic community – not, be it noted, by virtue of the particular *stuff* of which his brain is built . . . but by virtue of the particular kinds of mutual interaction that it can sustain with our own – interaction which at the personal level we describe as that of person-to-person.

MacKay is, I believe, conflating two choices into one. The first choice, to ascend from the mechanistic to the Intentional stance, as portrayed by our chess-playing designer, has no moral dimension. One is guilty of no monstrosities if one dismembers the computer with whom one plays chess, or even the robot with whom one has long conversations. One adopts the Intentional stance toward any system one assumes to be (roughly) rational, where the complexities of its operation preclude maintaining the design stance effectively. The second choice, to adopt a truly moral stance toward the system (thus viewing it as a person), might often turn out to be psychologically irresistible given the final choice, but it is logically distinct. Consider in this context the hunter trying to stalk a tiger by thinking what *he* would do if he were being hunted down. He has adopted the Intentional stance toward the tiger, and perhaps very effectively, but though the psychological tug is surely there to disapprove of the hunting of any creature wily enough to deserve the Intentional treatment, it would be hard to sustain a charge of either immorality or logical inconsistency against the hunter. We might, then, distinguish a fourth stance, above the Intentional stance, called the *personal stance*. The personal stance presupposes the Intentional stance (note that the Intentional

stance presupposes neither lower stance) and seems to cursory view at least to be just the annexation of moral commitment to the Intentional. (A less obvious relative of my distinctions of stance is Sellars' distinction between the manifest and scientific images of man. Sellars himself draws attention to its kinship to Strawson: 'Roughly, the manifest image corresponds to the world as conceived by P. F. Strawson. . . . The manifest image is, in particular, a framework in which the distinctive features of persons are conceptually irreducible to features of non-persons, e.g. animals and merely material things.'[8] A question I will not attempt to answer here is whether Sellars' manifest image lines up more with the more narrow, and essentially moral, personal stance or the broader Intentional stance.)

Something like moral commitment can exist in the absence of the Intentional stance, as Strawson points out, but it is not the same; the objective attitude – my design or physical stances – 'may include pity or even love, though not all kinds of love'. The solicitude of a gardener for his flowers, or for that matter, of a miser for his coins, cannot amount to moral commitment, because of the absence of the Intentional. (Parenthetical suggestion: is the central fault in utilitarianism a confusion of gardener-solicitude with person-solicitude?)

Since the second choice (of moral commitment) is like the first in being just a choice, relative to ends and desires and not provably right or wrong, it is easy to see how they can be run together. When they are, important distinctions are lost. Strawson's union of the two leads him to propose, albeit cautiously, a mistaken contrast: 'But what is above all interesting is the tension there is, in us, between the participant attitude and the objective attitude. One is tempted to say: between our humanity and our intelligence. But to say this would be to distort both notions.'[9] The distortion lies in allying the non-Intentional, mechanistic stances with the coldly rational and intelligent, and the Intentional stance with the emotional. The Intentional stance of one chess player toward another (or the hunter toward his prey) can be as coldly rational as you wish, and alternatively one can administer to one's automobile in a bath of sentiment.

Distinctions are also obscured if one makes *communicating with* a system the hallmark of Intentionality or rationality. Adopting the Intentional stance toward the chess-playing computer is not necessarily viewing one's moves as *telling* the computer anything (I do not have to *tell* my human opponent where I moved – he can *see* where I moved); it is merely predicting its responses with the assumption that it will

respond rationally to its *perceptions*. Similarly, the hunter stalking the tiger will be unlikely to try to *communicate* with the tiger (although in an extended sense even this might be possible – consider the sort of *entente* people have on occasion claimed to establish with bears encountered on narrow trails, etc.), but he will plan his strategy on his assessment of what the tiger would be reasonable to *believe* or *try*, given its perceptions. As Grice has pointed out,[10] one thing that sets communication as a mode of interaction apart from others is that in attempting a particular bit of communication with A, one intends to produce in A some response *and* one intends A to recognize that one intends to produce in him this response *and* one intends that A produce this response on the basis of this recognition. When one's assessment of the situation leads to the belief that these intentions are not apt to be fulfilled, one does not try to communicate with A, but one does not, on these grounds, necessarily abandon the Intentional stance. A may simply not understand any language one can speak, or any language at all (e.g. the tiger). One can still attempt to influence A's behavior by relying on A's rationality. For instance, one can throw rocks at A in an effort to get A to leave, something that is apt to work with Turk or tiger, and in each case what one does is at best marginal communication.[11]

Communication, then, is not a separable and higher *stance* one may choose to adopt toward something, but a type of interaction one may attempt within the Intentional stance. It can be seen at a glance that the set of intentions described by Grice would not be fulfilled with any regularity in any community where there was no *trust* among the members, and hence communication would be impossible, and no doubt this sort of consideration contributes to the feeling that the Intentional community (or at least the smaller *communicating* community) is co-extensive with the moral community, but of course the only conclusion validly drawn from Grice's analysis here is a pragmatic one: if one wants to influence A's behavior, and A is capable of communicating, then one will be able to establish a very *effective* means of influence by establishing one's trustworthiness in A's eyes (by hook or by crook). It is all too easy, however, to see interpersonal, convention-dependent communication as the mark of the Intentional – perhaps just because Intentional systems process information – and thus make the crucial distinction out to be that between 'poking at' a system (to use MacKay's vivid phrase) and communicating with it. Not only does this way of putting the matter wrongly confuse the system's perception of

communications with its perception more generally, but it is apt to lead to a moralistic inflation of its own. The notion of communication is apt to be turned into something mystical or semi-divine – synonyms today are 'rap', 'groove', 'dig', 'empathize'. The critical sense of communication, though, is one in which the most inane colloquies between parent and teenager (or man and bear) count as communication. (MacKay himself has on occasion suggested that the personal attitude is to be recognized in Buber's famous I–Thou formula, which is surely inflation.) The ethical implication to be extracted from the distinction of stance is not that the Intentional stance is a moral stance, but that it is a precondition of any moral stance, and hence if it is jeopardized by any triumph of mechanism, the notion of moral responsibility is jeopardized in turn.

V

Reason, not regard, is what sets off the Intentional from the mechanistic; we do not just reason about what Intentional system will do, we reason about how they will reason. And so it is that our predictions of what an Intentional system will do are formed on the basis of what would be reasonable (for anyone) to do under the circumstances, rather than on what a wealth of experience with this system or similar systems might inductively suggest the system will do. It is the absence from the mechanistic stances of this presupposition of rationality that gives rise to the widespread feeling that there is an antagonism between predictions or explanations from these different stances. The feeling ought to be dissipated at least in part by noting that the absence of a presupposition of rationality is not the same as a presupposition of non-rationality.

Suppose someone asks me whether a particular desk calculator will give 108 as the product of 18 and 6.[12] I work out the sum on a piece of paper and say, 'Yes'. He responds with, 'I know that it *should*, but will it? You see, it was designed by my wife, who is no mathematician.' He hands me her blueprints and asks for a prediction (from the design stance). In working on this prediction the assumption of rationality, or good design, is useless, so I abandon it, not as false but as question-begging. Similarly, if in response to his initial question I reply, 'It's an IBM, so yes', he may reply, 'I know it's *designed* to give that answer, but I just dropped it, so maybe it's broken'. In setting out to make this prediction I will be unable to avail myself of the assumption that the machine is designed to behave in a certain way, so I abandon it. My

prediction does not depend on any assumptions about rationality or design, but neither does it rescind any.

One reason we are tempted to suppose that mechanistic explanations preclude Intentional explanations is no doubt that since mechanistic explanations (in particular, physical explanations) are for the most part attempted, or effective, only in cases of malfunction or breakdown, where the rationality of the system is obviously impaired, we associate the physical explanation with a failure of Intentional explanation, and ignore the possibility that a physical explanation will go through (however superfluous, cumbersome, unfathomable) in cases where Intentional explanation is proceeding smoothly. But there is a more substantial source of concern than this, raised by MacIntyre.[13]

> Behaviour is rational – in this arbitrarily, defined sense – if, and only if, it can be influenced, or inhibited by the adducing of some logically relevant consideration. . . . But this means that if a man's behaviour is rational it cannot be determined by the state of his glands or any other antecedent causal factor. For if giving a man more or better information or suggesting a new argument to him is a both necessary and sufficient condition for, as we say, changing his mind, then we exclude, for this occasion at least, the possibility of other sufficient conditions. . . . Thus to show that behaviour is rational is enough to show that it is not causally determined in the sense of being the effect of a set of sufficient conditions *operating independently of the agent's deliberation or possibility of deliberation* [my italics]. So the discoveries of the physiologist and psychologist may indefinitely increase our knowledge of why men behave irrationally but they could never show that rational behaviour in this sense was causally determined.

MacIntyre's argument offers no license for the introduction of the italicized phrase above, and without it his case is damaged, as we shall see later, when the effect of prediction is discussed. More fundamental, however, is his misleading suggestion that the existence of sufficient conditions for events in a system puts that system in a strait-jacket, as it were, and thus denies it the flexibility required of a truly rational system. There is a grain of truth in this, which should be uncovered. In elaborating the distinction between stances, I chose for an example a system of rather limited versatility; the chess-playing system is unequipped even to play checkers or bridge, and input appropriate to these other games would reveal the system to be as non-rational and

unresponsive as any stone. There is a fundamental difference between such limited-purpose systems and systems that are supposed to be capable of responding appropriately to input of all sorts. For although it is possible in principle to design a system that can be guaranteed to respond appropriately (relative to some stipulated ends) to any limited number of inputs given fixed, or finitely ambiguous or variable, environmental 'significance', there is no way to design a system that can be guaranteed to react appropriately under *all* environmental conditions. A detailed argument for this claim would run on too long for this occasion, and I have presented the major steps of it elsewhere,[14] so I will try to establish at least comprehension, if not conviction, for the claim by a little thought-experiment about *tropistic behavior.* Wooldridge gives a lucid account of a tropism:[15]

> When the time comes for egg laying the wasp *Sphex* builds a burrow for the purpose and seeks out a cricket which she stings in such a way as to paralyze but not kill it. She drags the cricket into the burrow, lays her eggs alongside, closes the burrow, then flies away, never to return. In due course, the eggs hatch and the wasp grubs feed off the paralyzed cricket, which has not decayed, having been kept in the wasp equivalent of deep freeze. To the human mind, such an elaborately organized and seemingly purposeful routine conveys a convincing flavour of logic and thoughtfulness – until more details are examined. For example, the wasp's routine is to bring the paralyzed cricket to the burrow, leave it on the threshold, go inside to see that all is well, emerge, and then drag the cricket in. If, while the wasp is inside making her preliminary inspection the cricket is moved a few inches away, the wasp, on emerging from the burrow, will bring the cricket back to the threshold, but not inside, and will then repeat the preparatory procedure of entering the burrow to see that everything is all right. If again the cricket is removed a few inches while the wasp is inside, once again the wasp will move the cricket up to the threshold and re-enter the burrow for a final check. The wasp never thinks of pulling the cricket straight in. On one occasion, this procedure was repeated forty times, always with the same result.

The experiment unmasks the behavior as a tropism, rigid within the limits set on the significance of the input, however felicitous its operation under normal circumstances. The wasp's response lacks that free-

wheeling flexibility in response to the situation that Descartes so aptly honored as the infinity of the rational mind. For the notion of a perfectly rational, perfectly adaptable system, to which all input compatible with its input organs is significant and comprehensible is the notion of an unrealizable physical system. For let us take the wasp's tropism and improve on it. That is, suppose we take on the role of wasp designers, and decide to enlarge the subroutine system of the tropism to ensure a more rational fit between behavior and *whatever* environment the wasp may run into. We think up one stymying environmental condition after another, and in each case design subroutines to detect and surmount the difficulty. There will always be room for yet one more set of conditions in which the rigidly mechanical working out of response will be unmasked, however long we spend improving the system. Long after the wasp's behavior has become so perspicacious that we would not think of calling it tropistic, the fundamental nature of the system controlling it will not have changed; it will just be more complex. In this sense any behavior controlled by a finite mechanism must be tropistic.

What conclusion should be drawn from this about human behavior? That human beings, as finite mechanical systems, are not rational after all? Or that the demonstrable rationality of man proves that there will always be an inviolable *terra incognita*, an infinite and non-mechanical mind beyond the grasp of physiologists and psychologists? It is hard to see what evidence could be adduced in support of the latter conclusion, however appealing it may be to some people, since for every awe-inspiring stroke of genius cited in its favor (the Einstein-Shakespeare gambit), there are a thousand evidences of lapses, foibles, bumbling and bullheadedness to suggest to the contrary that man is only imperfectly rational. Perfection is hard to prove, and nothing short of perfection sustains the argument. The former alternative also lacks support, for although in the case of the wasp we can say that its behavior has been shown to be *merely* mechanically controlled, what force would the 'merely' have if we were to entertain the notion that the control of man's more versatile behavior is merely mechanical? The denigration might well be appropriate if in a particular case the mechanical explanation of a bit of behavior was short and sweet (consider explanations of the knee-jerk reflex or our hypothetical man who cannot say 'father'), but we must also consider cases in which the physiologist or cybernetician hands us twenty volumes of fine print and says, 'Here is the design of this man's behavioral control system'. Here is a case where

the philosopher's preference for simple examples leads him astray, for of course any *simple* mechanistic explanation of a bit of behavior will disqualify it for plausible Intentional characterization, make it a mere happening and not an action, but we cannot generalize from simple examples to complex, for it is precisely the simplicity of the examples that grounds the crucial conclusion.

The grain of truth in MacIntyre's contention is that *any* system that can be explained mechanistically – at whatever length – must be in an extended sense tropistic, and this can enhance the illusion that mechanistic and Intentional explanations cannot coexist. But the only implication that could be drawn from the *general* thesis of man's ultimately mechanistic organization would be that man must, then, be imperfectly rational, in the sense that he cannot be so designed as to *ensure* rational responses to all contingencies, hardly an alarming or counter-intuitive finding; and from any *particular* mechanistic explanation of a bit of behavior it would not follow that that particular bit of behavior was or was not a rational response to the environmental conditions at the time, for the mere fact that the response *had* to follow, given its causal antecedents, casts no more doubt on its rationality than the fact that the computer *had* to answer '108' casts doubt on the arithmetical correctness of its answer.

What, then, can we say about the hegemony of mechanistic explanations over Intentional explanations? Not that it does not exist, but that it is misdescribed if we suppose that whenever the former are confirmed, they drive out the latter. It is rather that mechanistic predictions, eschewing any presuppositions of rationality, can put the lie to Intentional predictions when a system happens to fall short of rationality in its response, whether because of weakness of 'design', or physically predictable breakdown. It is the presuppositions of Intentional explanation that put prediction of *lapses* in principle beyond its scope, whereas lapses are in principle predictable from the mechanistic standpoint, provided they are not the result of truly random events.[16]

VI

It was noted earlier that the search for a watershed to divide the things we are responsible for from the things we are not comes to rest usually with a formulation roughly harmonious with the distinction drawn here between the Intentional and the mechanistic. Many writers have

urged that we are responsible for just those events that are our intentional *actions* (and for their foreseeable results), and a great deal has been written in an effort to distinguish action from mere happening. The performing of actions is the restricted privilege of rational beings, persons, conscious agents, and one establishes that something is an action not by examining its causal ancestry but by seeing whether certain sorts of talk about *reasons* for action are appropriate in the context. On this basis we exculpate the insane, with whom one is unable to reason, unable to communicate; we also excuse the results of physical *force majeure* against which reason cannot prevail, whether the force is external (the chains that bind) or internal (the pain that makes me cry out, revealing our position to the enemy). This fruitful distinction between reason giving and cause giving is often, however, the source of yet another misleading intuition about the supposed antagonism between mechanism and responsibility. 'Roughly speaking', Anscombe says, 'it establishes something as a reason if one argues against it.'[17] One is tempted to go on: a reason is the sort of thing one can argue against with some hope of success, but one cannot argue against a causal chain. There is of course a sense in which this is obvious: one cannot argue with what has no ears to hear, for instance. But if one tries to get the point into a form where it will do some work, namely: 'the presentation of an argument cannot affect a causal chain', it is simply false. Presentations of arguments have all sorts of effects on the causal milieu: they set air waves in motion, cause ear drums to vibrate, and have hard to identify but important effects deep in the brain of the audience. So although the presentation of an argument may have no detectable effect on the trajectory of a cannonball, or closer to home, on one's *autonomic* nervous system, one's perceptual system is designed to be sensitive to the sorts of transmissions of energy that must occur for an argument to be communicated. The perceptual system can, of course, be affected in a variety of ways; if I sneak up behind someone and yell 'flinch, please!' in his ear, the effects wrought by my utterance would not constitute an action in obedience to my request, not because they were effects of a cause, but because the intricate sort of causal path that in general would have to have existed for an Intentional explanation to be appropriate was short-circuited. An Intentional system is precisely the sort of system to be affected by the input of information, so the discovery in such a system of a causal chain culminating in a bit of behavior does not at all license the inference: 'since the behavior was caused we could not have argued him out of it', for a prior attempt

to argue him out of it would have altered the causal ancestry of the behavior, perhaps effectively.

The crucial point when assessing responsibility is whether or not the antecedent inputs achieve their effects as inputs of information or by short-circuit. The possibility of short-circuiting or otherwise tampering with an Intentional system gives rise to an interesting group of perplexities about the extent of responsibility in cases where there has been manipulation. We are generally absolved of responsibility in cases where we have been manipulated by others, but there is no one principle of innocence by reason of manipulation. To analyze the issue we must first separate several distinct excusing conditions that might be lumped together under the heading of manipulation.

First, one may disclaim responsibility for an act if one has been led to commit the act by deliberately false information communicated by another, and one might put this: 'he manipulated me, by forging documents'. The principle in such cases has nothing to do with one's Intentional system being tampered with, and in fact the appeal to the deliberate malice of the other party is a red herring.[18] The principle invoked to determine guilt or innocence in such cases is simply whether the defendant had reasonably good evidence for the beliefs which led to his act (and which, if true, would have justified it presumably). The plain evidence of one's senses is normally adequate when what is at issue is the presentation of a legal document, and so normally one is absolved when one has been duped by a forgery, but not, of course, if the forgery is obvious or one has any evidence that would lead a reasonable man to be suspicious. And if the evidence that misled one into a harmful act was produced by mere chance or 'act of God' (such as a storm carrying away a 'Stop' sign) the principle is just the same. When one is duped in this manner by another, one's Intentional system has not been tampered with, but rather exploited.

The cases of concern to us are those in which one's behavior is altered by some non-rational, non-Intentional interference. Here, cases where a person's body is merely mechanically interposed in an ultimately harmful result do not concern us either (e.g. one's arm is bumped, spilling Jones's beer, or less obviously, one is drugged, and hence is unable to appear in court). One is excused in such cases by an uncomplicated application of the *force majeure* principle. The only difficult cases are those in which the non-rational, non-Intentional interference alters one's beliefs and desires, and subsequently, one's actions. Our paradigm here is the idea – still fortunately science fiction

– of the neurosurgeon who 'rewires' me and in this way inserts a belief or desire that was not there before. The theme has an interesting variation which is not at all fictional: the mad scientist might discover enough about a man's neural *design* (or program) to figure out that certain inputs would have the effect of reprogramming the man, quite independent of any apparent sense they might have for the man to react to rationally. For instance, the mad scientist might discover that flashing the letters of the alphabet in the man's eyes at a certain speed would cause him (in virtue of his imperfectly rational design) to believe that Mao is God. We have, in fact, fortuitously hit upon such ways of 'unlocking' a person's mind in hypnotism and brain-washing, so the question of responsibility in such cases is not academic. Some forms of psychotherapy, especially those involving drugs, also apparently fall under this rubric. Again it should be noted that the introduction of an evil manipulator in the examples is superfluous. If I am led to believe that Mao is God by a brain hemorrhage or eating tainted meat, or by being inadvertently hypnotized by the monotony of the railroad tracks, the same puzzling situation prevails.

Philosophers have recognized that something strange is going on in these cases, and have been rightly reluctant to grant that such descriptions as I have just given are fully coherent. Thus Melden says,[19]

> If by introducing an electrode into the brain of a person, I succeed in getting him to believe that he is Napoleon, that surely is not a rational belief that he has, nor is he responsible for what he does in consequence of this belief, however convinced he may be that he is fully justified in acting as he does.

Why, though, is the man not responsible? Not because of the absurdity of the belief, for if a merely negligent evidence-gatherer came to believe some absurdity, his consequent action would not be excused, and if the electrode-induced belief happened to be true but just previously unrecognized by the man, it seems we would still deny him responsibility. (I do not think this is obvious. Suppose a benevolent neurosurgeon implants the belief that honesty is the best policy in the heads of some hardened criminals; do we, on grounds of non-rational implantation, deny these people status in the society as responsible agents?) The non-rationality, it seems, is not to be ascribed to the *content* of the belief, but somehow to the manner in which it is believed or acquired. We do, of course, absolve the insane, for they are *in general*

irrational, but in this case we cannot resort to this precedent for the man has, *ex hypothesi*, only one non-rational belief. Something strange indeed is afoot here, for as was mentioned before, the introduction of the evil manipulator adds nothing to the example, and if we allow that the presence of one non-rationally induced belief absolves from responsibility, and if the absurdity or plausibility of a belief is independent of whether it has been rationally acquired or not, it seems we can never be sure whether a man is responsible for his actions, for it just may be that one of the beliefs (true or false) that is operative in a situation has been produced by non-rational accident, in which case the man would be ineligible for praise or blame. Can it be that there is a tacit assumption that no such accidents have occurred in those cases where we hold men responsible? This line is unattractive, for suppose it were *proved* in a particular case that Smith was led to some deed by a long and intricate argument, impeccably formulated by him, with the exception of one joker, a solitary premise non-rationally induced. Our tacit assumption would be shown false; would we deny him responsibility?

A bolder skepticism toward such example has been defended by MacIntyre: 'If I am right the concept of causing people to change their beliefs or to make moral choices, by brain-washing or drugs, for example, is not a possible concept.'[20] Hampshire, while prepared to countenance causing beliefs in others, finds a conceptual difficulty in the reflexive case: 'I must regard my own beliefs as formed in response to free inquiry; I could not otherwise count them as beliefs.'[21] Flew vehemently attacks MacIntyre's proposal:[22]

> If it did hold it would presumably rule out as logically impossible all indoctrination by such non-rational techniques. The account of Pavlovian conditionings in Aldous Huxley's *Brave New World* would be not a nightmare fantasy but contradictory nonsense. Again if this consequence did hold, one of the criteria for the use of the term *belief* would have to be essentially backward-looking. Yet this is surely not the case. The actual criteria are concerned with the present and future dispositions of the putative believer; and not at all with how he may have been led, or misled, into his beliefs.

Flew's appeal to the reality of brain-washing is misplaced, however, for what is at issue is how the results of brain-washing are to be coherently described, and MacIntyre is right to insist that there is a conceptual

incoherency in the suggestion that in brain-washing one causes beliefs, *tout simple*. Elsewhere[23] I have argued that there *is* an essential backward-looking criterion of belief; here I shall strike a more glancing blow at Flew's thesis. Suppose for a moment that we put ourselves in the position of a man who wakes up to discover a non-rationally induced belief in his head (he does not know it was non-rationally induced; he merely encounters this new belief in the course of reflection, let us say). What would this be like? We can tell several different stories, and to keep the stories as neutral as possible, let us suppose the belief induced is false, but not wild: the man has been induced to believe that he has an older brother in Cleveland.

In the first story, Tom is at a party and in response to the question, 'Are you an only child?' he replies, 'I have an older brother in Cleveland.' When he is asked, 'What is his name?' Tom is baffled. Perhaps he says something like this: 'Wait a minute. Why do I think I have a brother? No name or face or experiences come to mind. Isn't that strange: *for a moment* I had this feeling of conviction that I had an older brother in Cleveland, but now that I think back on my childhood, I remember perfectly well I was an only child.' If Tom has come out of his brainwashing still predominantly rational, his induced belief can last only a moment once it is uncovered. For this reason, our earlier example of the impeccable practical reasoning flawed by a lone induced belief is an impossibility.

In the second story, when Tom is asked his brother's name, he answers 'Sam' and proceeds to answer a host of other obvious questions, relates incidents from his childhood, and so forth. Not *one* belief has been induced, but an indefinitely large stock of beliefs, and other beliefs have been wiped out. This is a more stable situation, for it may take a long time before Tom encounters a serious mismatch between this large and interrelated group and his other beliefs. Indeed, the joint, as it were, between this structure of beliefs and his others may be obscured by some selective and hard to detect amnesia, so that Tom never is brought up with any hard-edge contradictions.

In the third story, Tom can answer no questions about his brother in Cleveland, but insists that he believes in him. He refuses to acknowledge that well-attested facts in his background make the existence of such a brother a virtual impossibility. He says bizarre things like, 'I know I am an only child and have an older brother living in Cleveland.' Other variations in the story might be interesting, but I think we have touched the important points on the spectrum with these three

stories. In each story the question of Tom's responsibility can be settled in an intuitively satisfactory way by the invocation of familiar principles. In the first case, while it would be *hubris* to deny that a neurosurgeon might some day be able to set up Tom in this strange fashion, if he can do it without disturbing Tom's prevailing rationality the effect of the surgery on Tom's beliefs will be evanescent. And since we impose a general and flexible obligation on any rational man to inspect his relevant beliefs before undertaking important action, we would hold Tom responsible for any rash deed he committed while under the temporary misapprehension induced in him. Now if it turned out to be physically impossible to insert a single belief without destroying a large measure of Tom's rationality, as in the third story, we would not hold Tom responsible, on the grounds of insanity – his rationality would have been so seriously impaired as to render him invulnerable to rational communication. In the second story determining responsibility must wait on answers to several questions. Has Tom's rationality been seriously impaired? If not, we must ask the further question: did he make a reasonable effort to examine the beliefs on which he acted? If the extent of his brainwashing is so great, if the fabric of falsehoods is so broad and well-knit, that a reasonable man taking normal pains could not be expected to uncover the fraud, then Tom is excused. Otherwise not.

With this in mind we can reconsider the case of the hardened criminals surgically rehabilitated. Are they responsible citizens now, or zombies? If the surgeon has worked so delicately that their rationality is not impaired (perhaps improved!), they are, or can become, responsible. In such a case the surgeon will not so much have implanted a belief as implanted a suggestion and removed barriers of prejudice so that the suggestion *will be* believed, given the right sort of evidential support. If on the other hand the patients become rigidly obsessive about honesty, while we may feel safe allowing them to run loose in the streets, we will have to admit that they are less than persons, less than responsible agents. A bias in favor of true beliefs can be detected here: since it is hard to bring an evidential challenge to bear against a true belief (for lack of challenging evidence – unless we fabricate or misrepresent), the flexibility, or more specifically rationality, of the man whose beliefs all seem to be true is hard to establish. And so, if the rationality of the hardened criminals' new belief in honesty is doubted, it can be established, if at all, only by deliberately trying to shake the belief!

The issue between Flew and MacIntyre can be resolved, then, by noting that one cannot directly and simply cause or implant a belief, for a belief is essentially something that has been *endorsed* (by commission or omission) by the agent on the basis of its conformity with the rest of his beliefs. One may well be able to produce a zombie, either surgically or by brainwashing, and one might even be able to induce a large network of false beliefs in a man, but if so, their persistence *as beliefs* will depend, not on the strength of any sutures, but on their capacity to win contests against conflicting claims in evidential showdowns. A parallel point can be made about desires and intentions. Whatever might be induced in me is either fixed and obsessive, in which case I am not responsible for where it leads me, or else, in MacIntyre's phrase, 'can be influenced or inhibited by the adducing of some logically relevant consideration', in which case I am responsible for *maintaining* it.

VII

I believe the case is now complete against those who suppose there to be an unavoidable antagonism between the Intentional and the mechanistic stance. The Intentional stance toward human beings, which is a precondition of any ascriptions of responsibility, *may* coexist with mechanistic explanations of their motions. The other side of this coin, however, is that we *can* in principle adopt a mechanistic stance toward human bodies and their motions, so there remains an important question to be answered. *Might* we abandon the Intentional stance altogether (thereby of necessity turning our backs on the conceptual field of morality, agents, and responsibility) in favor of a purely mechanistic world view, or is this an alternative that can be ruled out on logical or conceptual grounds? This question has been approached in a number of different ways in the literature, but there is near unanimity about the general shape of the answer: for Strawson the question is whether considerations (of determinism, mechanism, etc.) could lead us to look on everyone exclusively in the 'objective' way, abandoning the 'participant' attitude altogether. His decision is that this could not transpire, and he compares the commitment to the participant attitude to our commitment to induction, which is 'original, natural, non-rational (not *ir*rational), in no way something we choose or could give up'.[24] Hampshire puts the point in terms of the mutual dependence of 'two

kinds of knowledge', roughly, inductive knowledge and knowledge of one's intentions. 'Knowledge of the natural order derived from observation is inconceivable without a decision to test this knowledge, even if there is only the test that constitutes a change of point of view in observation of external objects.'[25] In other words, one cannot *have* a world view of any sort without having beliefs, and one could not have beliefs without having intentions, and having intentions requires that one view *oneself*, at least, Intentionally, as a rational agent. Sellars makes much the same point in arguing that 'the scientific image cannot replace the manifest without rejecting its own foundation'.[26] Malcolm says, 'The motto of the mechanist ought to be: One cannot speak, therefore one must be silent.'[27] But here Malcolm has dropped the ball on the goal line; how is the mechanist to *follow* his 'motto', and how *endorse* the 'therefore'? The doctrine that emerges from all these writers is that you can't get there from here, that to assert that the Intentional is eliminable 'is to imply pragmatically that there is at least one person, namely the one being addressed, if only oneself, with regard to whom the objective attitude cannot be the only kind of attitude that is appropriate to adopt'.[28] Recommissioning Neurath's ship of knowledge, we can say that the consensus is that there is at least one plank in it that cannot be replaced.

Caution is advisable whenever one claims to have proved that something cannot happen. It is important to see what does not follow from the consensus above. It does not follow, though Malcolm thinks it does,[29] and there are some things in the world, namely human beings, of which mechanism as an embracing theory cannot be true, for there is no incompatibility between mechanistic and Intentional explanation. Nor does it follow that we will always characterize some things Intentionally, for we may all be turned into zombies next week, or in some other way the human race may be incapacitated for communication and rationality. All that is the case is that we, *as persons*, cannot *adopt* exclusive mechanism (by eliminating the Intentional stance altogether). A corollary to this which has been much discussed in the literature recently is that we, as persons, are curiously immune to certain sorts of predictions. If I cannot help but have a picture of myself as an Intentional system, I am bound, as MacKay has pointed out, to have an *underspecified* description of myself, 'not in the sense of leaving any parts unaccounted for, but in the sense of being compatible with more than one state of the parts'.[30] This is because no information system can carry a complete true representation of itself (whether this

representation is in terms of the physical stance or any other). And so I cannot even in principle have all the data from which to predict (from any stance) my own future.[31] Another person might in principle have the data to make all such predictions, but he could not tell them all to me without of necessity falsifying the antecedents on which the prediction depends by interacting with the system whose future he is predicting, so I can never be put in the position of being obliged to believe them. As an Intentional system I have an epistemic horizon that keeps my own future as an Intentional system indeterminate. Again, a word of caution: this barrier to prediction is not one we are going to run into in our daily affairs; it is not a barrier preventing or rendering incoherent predictions I might make about my own future decisions, as Pears for one has pointed out.[32] It is just that since I must view myself as a person, a full-fledged Intentional system, there is no complete biography of my future I would be right to accept.

All this says nothing about the impossibility of dire depersonalization in the future. Wholesale abandonment of the Intentional is in any case a less pressing concern than partial erosion of the Intentional domain, an eventuality against which there are no conceptual guarantees at all. If the growing area of success in mechanistic explanation of human behavior does not in itself rob us of responsibility, it does make it more pragmatic, more effective or efficient, for people on occasion to adopt less than the Intentional stance toward others. Until fairly recently the only well-known generally effective method of getting people to do what you wanted them to was to treat them as persons. One might threaten, torture, trick, misinform, bribe them, but at least these were forms of control and coercion that appealed to or exploited man's rationality. One did not attempt to adopt the design stance or the physical stance, just because it was so unlikely that one could expect useful behavioral results. The advent of brainwashing, subliminal advertising, hypnotism and even psychotherapy (all invoking variations on the design stance), and the more direct physical tampering with drugs and surgical intervention, for the first time make the choice of stance a genuine one. In this area many of the moral issues are easily settled; what dilemmas remain can be grouped, as MacKay has observed, under the heading of treating a person as less than a person *for his own good*. What if mass hypnosis could make people stop wanting to smoke? What if it could make them give up killing? What if a lobotomy will make an anguished man content? I argued earlier that in most instances we must ask for much more precise descriptions of the

changes wrought, if we are to determine whether the caused change has impaired rationality and hence destroyed responsibility. But this leaves other questions still unanswered.

Tufts University, Medford, Mass.

Notes

1 J. Hospers, 'What Means This Freedom?' in S. Hook (ed.), *Determinism and Freedom in the Age of Modern Science* (New York: Collier, 1958), p. 133.
2 N. Malcolm, 'The Conceivability of Mechanism', *Phil. Review*, 1968, p. 51.
3 A. I. Melden, *Free Action* (London: Routledge & Kegan Paul, 1961); D. Davidson, 'Actions, Reasons and Causes', *J. Phil.*, 1963, pp. 685–700.
4 For a more detailed analysis of the concept, see my 'Intentional Systems', *J. of Philosophy*, 25 February 1971, pp. 87–106, where in particular the notions of rationality of design and Intentionality of information-processing systems are discussed at length.
5 D. M. MacKay, 'The Use of Behavioral Language to Refer to Mechanical Processes', *Brit. J. Phil. Sci.* XIII, 1962, pp. 89–103. See also H. Putnam, 'Robots: Machines or Artificially Created Life?', read at A.P.A. Eastern Div. Meeting, 1964, subsequently published in S. Hampshire (ed.), *Philosophy of Mind* (New York: Harper & Row, 1966), p. 91.
6 P. F. Strawson, 'Freedom and Resentment', *Proc. Brit. Acad.*, 1962, reprinted in Strawson (ed.), *Studies in the Philosophy of Thought and Action* (Oxford University Press, 1968), p. 79.
7 MacKay, *op. cit.*, p. 102.
8 W. Sellars, 'Fatalism and Determinism', in K. Lehrer (ed.), *Freedom and Determinism* (New York: Random House, 1966), p. 145. A Flew, 'A Rational Animal', in J. R. Smythies (ed.), *Brain and Mind* (London: Routledge & Kegan Paul, 1968), pp. 111–35, and A. Rorty, 'Slaves and Machines', *Analysis*, April 1962, pp. 118–20, develop similar distinctions.
9 Strawson, *op. cit.*, p. 80.
10 H. P. Grice, 'Meaning', *Phil. Review*, 1957; 'Utterer's Meaning and Intentions', *Phil. Review*, 1969.
11 J. Bennett, in *Rationality* (London: Routledge & Kegan Paul, 1964), offers an extended argument to the effect that communication and rationality are essentially linked, but his argument is vitiated, I believe, by its reliance on an artificially restrictive sense of rationality – a point it would take too long to argue here. See my 'Intentional Systems', *loc. cit.*, for arguments for a more generous notion of rationality.
12 Cf. L. W. Beck, 'Agent, Actor, Spectator, and Critic', *Monist*, 1965, pp. 175–9.
13 A. C. MacIntyre, 'Determinism', *Mind*, 2957, pp. 248f.
14 *Content and Consciousness* (London: Routledge & Kegan Paul, 1969).
15 D. Wooldridge, *The Machinery of the Brain* (New York: McGraw-Hill, 1963), p. 82.

16 In practice we predict lapses at the Intentional level ('You watch! He'll forget all about your knight after you move the queen') on the basis of loose-jointed inductive hypotheses about individual or widespread human frailties. These hypotheses are expressed in Intentional terms, but if they were given rigorous support, they would be in the process be recast as predictions from the design or physical stance.
17 G. E. M. Anscombe, *Intention* (Oxford: Blackwell, 1957), p. 24.
18 Cf. D. M. MacKay, 'Comments on Flew', in Smythies, *op. cit.*, p. 130.
19 Melden, *op. cit.*, p. 214.
20 Quoted by Flew, *op. cit.*, p. 118.
21 S. Hampshire, *Freedom of the Individual* (New York: Harper & Row, 1965), p. 87.
22 Flew, *op. cit.*, p. 120.
23 *Content and Consciousness.*
24 Strawson, *op. cit.*, p. 94.
25 This is of course an echo of Strawson's examination of the conditions of knowledge in a 'no-space world' in *Individuals* (London: Methuen, 1959).
26 W. Sellars, *Science, Perception and Reality* (London: Routledge & Kegan Paul, 1963), p. 21.
27 Malcolm, *op. cit.*, p. 71.
28 J. E. Llewelyn, 'The Inconceivability of Pessimistic Determinism', *Analysis*, 1966, pp. 39–44. Having cited all these authorities, I must acknowledge my own failure to see this point in *Content and Consciousness*, p. 190. This is correctly pointed out by R. L. Franklin in his review in *Austr. J. Phil.*, September 1970.
29 Malcolm, *op. cit.*, p. 71.
30 D. M. MacKay, 'On the Logical Indeterminacy of a Free Choice', *Mind*, 1960, pp. 31–40; 'The Use of Behavioral Language to Refer to Mechanical Processes', *loc. cit.*; 'The Bankruptcy of Determinism', unpublished, read June 1969, at University of California at Santa Barbara.
31 Cf. K. Popper, 'Indeterminism in Quantum Physics and Classical Physics', *Brit. J. Phil. Sci.*, 1950.
32 D. F. Pears, 'Pretending and Deciding', *Proc. Brit. Acad.*, 1964, reprinted in Strawson (ed.), *Studies in the Philosophy of Thought and Action*, *loc. cit.*, pp. 97–133.