CORRIGIBILITY AS THE MARK OF THE MENTAL

IN this paper I will argue that a common account of what it means for an activity to be conscious is unsatisfactory in that it involves a reference to further conscious activities. It runs as follows: An activity is conscious when the agent who is engaged in that activity knows what she is doing. But for one thing, knowing what one is doing should itself be a conscious activity. It is an even more sophisticated conscious activity than the one we set out to explain in the first place. For another, the knowledge of what one is doing can only be identified as such by specifying its object as a conscious activity. This again will render the explanation circular.

I will then go on to argue that to isolate a mark of the mental, one needs to take the etymological origin of the epithet 'conscious' at face value. I will claim that consciousness is originally and essentially an epistemic version of moral conscience. But moral conscience is, in the traditional sense of this term, not a conscious activity. Neither need consciousness be so, then.

Let me begin by distinguishing conscious activities from other things that may count as mental. Paradigmatic instances of conscious activity are thoughts, perceptions and intentional actions. By 'thoughts' here I mean not belief states, but occurrent, actual acts of thinking. Besides these, there are other mental entities that are not events and that are only dubiously called 'entities' at all—tendencies, habits, capacities and states. Another way of contrasting these two classes of mental entities is to note that conscious activities are of the same type insofar as they have the same meaning or object. They are individuated in terms of their meaning and object. Only in this sense can I think or intend the same twice. In contrast, tendencies, habits, and capacities are not identified in terms of their objects. It is true that two capacities of the same kind will also often have the same object when they are actualized. But on the other hand, the very same capacity can be actualized in different ways. It will then have a different object on different occasions.

In this paper I will focus on conscious activities. Such activities are of a general type by virtue of having a specific object or value. These

activities will be particular tokens, but they will always instantiate a general type. They are, in a way, repeatable tokens. I assume that mental entities other than conscious activities can always be circumscribed in terms of instances of conscious activity. For instance, a tendency is something that eventually leads to some such activity and a state may be described as the result of such an activity. I will not go further into that here.

A well-known and often criticized philosophical account of what it is for an activity to be conscious is the one given by Descartes. Descartes uses the term 'cogitatio' in a rather technical sense. It denotes all kinds of conscious activity, such as sensations, perceptions, and acts of willing and thinking. Descartes introduces 'cogitatio' as a general term for everything that one must be capable of doing in order not to trust one's sensory experience. In order not to trust one's experience, one must have some such experience, understand beliefs about them, and refuse to rely on them in one' actions. In short, one must have experiences, beliefs and intentions. This is how Descartes defines the general class that is comprised of these activities:

> I use the term 'cogitatio' to include everything that happens in us such that we are immediately conscious of it.[1]

This is a surprisingly simple definition for a phenomenon as complex as conscious activity. Let me go through everything in it that might need clarification. Note, first, that Descartes writes that something *happening in us* is a conscious activity (cogitatio) if it is the immediate object of our consciousness. What kinds of things happen in us? Some physical processes happen literally within our bodies, such as C-fiber stimulations, respiration and digestion. But Descartes uses the preposition 'in' in a more general sense. In this sense, predicates are 'in' subjects and actions are 'in' agents.[2] Accordingly, a thought occurs in me because it is a thought that I think, just as an action happens in me by being something that I do.

---

[1] Descartes, *Meditationes*, appendix to the Second Replies, *Oeuvres* VII:160.

[2] Compare the sense in which Descartes says that actions are in the mind, in his Letter to Arnauld from July 29, 1648, *Oeuvres* V:222.

But shouldn't we restrict the word 'cogitatio' to *mental* happenings in us? Unfortunately, Descartes could not do that at this stage, since he will later rely on the notion of a 'cogitatio' in order to circumscribe the mental realm. Therefore, he cannot appeal to the distinction between mental and other events in the definition of 'cogitatio'. Hence, in restricting the kind of knowledge that is referred to as 'being immediately conscious' in the definition, we must not assume beforehand that its objects can only be *conscious* activities. This will in fact turn out to be the case, but we are here using this fact in order to say what a conscious activity is. Conscious activity is defined as whatever activity in us is the immediate object of our consciousness. So far, since we cannot yet appeal to the distinction between mental and physical events, merely physical events in our bodies might turn out to be conscious activities. They happen in us and we may know about them. This is obviously not what Descartes wanted.

Perhaps the next element of the definition helps here: that the activity should be *immediately* conscious. For is it not only by looking at a screen or some other device that we may know what happens in our brains without knowing it as our own thoughts? As *physical* events, we will know these events only mediately. Our knowledge of their physical properties will be mediated by observation. Conversely, it seems that insofar as we immediately know what happens in us, we know it as our own conscious activity and not as a merely physical event. This however is not true.

Immediacy should be understood in its most straightforward sense here. For instance, place is immediately adjacent to another place when there is no further place between them. A person is immediately responsible for an event if there is no other person that mediates between her and the event. According to this pattern, something is an immediate object of our consciousness if there is no other object of consciousness that mediates between it and the consciousness of it. A mediate object of our consciousness, in contrast, is only the object of our consciousness because something else is the object of our consciousness.

Now consider my respiration. It certainly happens in me, and I may think of it. It may even be the immediate object of my thought. My thought will then be about my respiration not only by virtue of being

about another thought that is about my respiration. But respiration is not a kind of conscious activity in the required sense. Therefore, not everything happening in me that is the immediate object of my thinking will be a conscious activity (cogitatio).

It may be true that we can have immediate knowledge of our own conscious activities, but not everything that we can know immediately will thereby be our own conscious activity. From this it follows that in his definition, by 'being conscious' Descartes cannot have meant 'thinking about'. Being immediately conscious of something that happens in me cannot be the same as having a thought that has this happening as its immediate object.

More generally, if consciousness were itself a conscious activity, then the definition that Descartes proposes would be circular: He would define 'cogitatio' as an activity of the subject that is the object of one of her 'cogitationes'. If her conscious activity were conscious by virtue of being the object of further conscious activity, then Descartes would even define simple conscious activity, such as having a thought or seeing a tree, by reference to more complicated conscious activity, such as thinking that one thinks or that one sees a tree. This does not look like a promising move.

Some might want to suggest that although consciousness is not itself an activity, it may be a tendency to engage in such activity. For an agent to know what she is doing need not involve her actually thinking about it. It will be enough if she has the tendency to have the appropriate thoughts about it under certain circumstances. Accordingly, one might assume that an activity is conscious if the person who is engaged in that activity tends to engage in second order conscious activity that has her first order conscious activity as its immediate object. But this does not help us either. For in order to specify such a tendency, we have to state what would constitute an actualization of it. The tendency in question is actualized by thinking about something that happens in the person that has the tendency. So far, this might also be a thought about her respiration. But respiration is not an instance of what Descartes calls 'cogitatio', no matter how hard we think about it. Hence, we would have to insist that our consciousness is not just the tendency of having a thought

about something that happens in us, but more precisely the tendency of having a thought about one's own conscious activity. We shall be going round in circles again. In our definition of conscious activity, we still refer to further conscious activity. We will even have to do that twice: we will have to define consciousness as a tendency to engage in (1) conscious activity that has (2) conscious activity as its immediate object.

It should be clear, then, that consciousness is not itself a conscious activity, nor is it a tendency to engage in conscious activity. As a consequence, the thinking substance (res cogitans) is not the same as a conscious mind (res conscia). For being conscious is not a conscious activity, and only conscious activities belong to the res cogitans. Since consciousness is not a kind of cogitatio, it does not belong to the res cogitans. Conscious activities may take place in the mind. Consciousness is not one of them. It does not take place in the mind.

What else is consciousness, if it is not a kind of conscious activity? At this point, it might be tempting to say that consciousness is a basic, inexplicable and irreducible phenomenon. But this would not solve the problem that consciousness does not belong to the res cogitans, for it still won't be a cogitatio. Moreover, we should not deliberately stop making sense at any point anyway. We may do that in axiomatic logic, but as philosophers, we should not assume basic entities at will. Even less should we do so only because our attempts to make ourselves clear happen to fail. That we cannot articulate what we mean by 'consciousness' does not imply that it is a brute and basic fact. If there is any chance of saying something further about any given phenomenon, we should try to do so, even if that leads to circular explanations. By following a circular explanation, we will at least reveal some of the structure of the phenomena under investigation. And this is what philosophy is about.

Anyway, consciousness does not even at first sight appear to be fundamental and inexplicable. When we compare conscious activity with the behavior of lower animals, we do not have the impression that something *simple* makes the difference. The difference does not appear to consist in the addition of some basic an inexplicable factor. It seems natural to say that an activity is conscious if and only if the agent knows

what she is doing, in some sense of 'knowing'. This knowledge is a rather sophisticated achievement, and it should be subject to further investigation.

Consciousness, as I have said, is a kind of knowledge of what one is doing. I will now go on to claim that this knowledge is not a further conscious activity of the agent, but an ideal knowledge from an impersonal point of view. An activity is conscious when it is subject to an ideal and objective evaluation.

Explicitly stating such an evaluation would be a second order conscious activity. It would be a thought about another thought. As such, it would an instance of what Descartes calls 'cogitatio', and it would be subject to further objective evaluation. It might be mistaken.

However, that an activity is subject to an evaluation does not imply that it actually gets evaluated at any time. An ideal and objective evaluation may also apply to something without in fact being applied. In general, having a value is logically independent from being evaluated. Things that have a value need not be evaluated by anyone, and things that are evaluated by someone need not have a value. That something has a value also does not mean that any actual person *tends* to evaluate it. Again, no one need actually be inclined to evaluate something that has a value, and people do indeed have a habit of attributing values to things that have no value. Hence, that an activity has a value does not correspond to any particular activity of evaluating it or any related tendency.

I will nonetheless speak of an evaluation, namely an objective and ideal one. This evaluation is the one that a thing that has a value actually deserves. It need not be actual. Since it need not be actually carried out, the objective evaluation of an activity is not an activity. It rather corresponds to a perspective or an aspect according to which the activity has its value. I claim that this objective and ideal evaluation is what 'being conscious' in Descartes' definition refers to. That an activity is conscious simply means that it has a certain ideal and objective value. I understand that this claim still requires some further argument and clarification at this stage.

A consciousness that is the objective, but not necessarily actualized evaluation of an activity would fit neatly into Descartes' definition of

'cogitatio'. He would then define conscious activity as something happening in us that is immediately subject to an objective evaluation. Conscious activity is an activity that has a certain value. More specifically, the value of an intentional action may be a moral value, and the value of a conscious thought may be a truth-value. If Descartes really uses 'consciousness' in the sense outlined so far, he assimilates truth-values to moral values. Truth-values are the moral values that specifically apply to such activities as thought and perception. Just as we ought to avoid morally bad behavior in general—this is what 'bad' means—, we ought to avoid false beliefs. This in turn can be taken to be the meaning of 'false'.

One reason to attribute such a notion of consciousness to Descartes is that it enables us to make sense of his definition of conscious activity. We can see on this basis how he manages not to define conscious activity in terms of further conscious activity. Another reason for attributing this notion of consciousness to him is that before Descartes, the Latin term 'conscientia' was indeed used in exactly the required sense. It is a well-known fact about the etymology of the word 'consciousness' that the corresponding Latin term originally referred to moral conscience. It is commonly assumed that Descartes was the very first author to use the word in a different sense. But there is no good reason for such an assumption. There rather are good reasons against it. As I have demonstrated above, when we do not interpret Descartes according to the traditional notion of conscientia, we end up with a circular definition of conscious activity.

Let me briefly recall some facts about the traditional meaning of the term 'conscientia'. Moral conscience comes in a pair with the Final Judgment, and both were popularized by St. Paul more than by anyone else. In his Letter to the Romans, Paul writes that the conscience (συνείδησις) of the gentiles testifies the natural law that is written in their hearts. On the Final Day, they will be judged according to this law. When John Locke defines 'person' by reference to consciousness, he explicitly refers to this picture:

> But in the great Day, wherein the Secrets of all Hearts shall be laid open, it may be reasonable to think, no one shall be made to answer for what he

> knows nothing of; but shall receive his Doom, his Conscience accusing or excusing him.[3]

Pauline conscience is thus an intuitive knowledge of the lawfulness of one's own actions. On the basis of this knowledge, we will eventually be held responsible for everything we have done. The according judgment will be pronounced in the Final Day, when our thoughts 'accuse and excuse each other'.[4] The Final Judgment is the objective and ideal, but not yet actual evaluation of our actions to which I have referred above.

Our moral conscience is thus a knowledge about the moral value of our actions that we ourselves need not yet possess. It will be revealed to us on the Last Day when our conscience will be published, as St. Augustine has it in one of his *Sermons*.[5] In the Christian worldview, it is easy to say who has this knowledge until then. God is the ideal observer and evaluator of our intentional actions. We need not think of his evaluation as something that actually takes place, but only as something that will take place under the ideal circumstances of the Final Judgment. In other words, God's knowledge of the value of our actions is not a conscious activity. God does not actually think. The reason for this is simply that God could not have any false belief, nor could he do anything morally wrong. This means that his activities are not subject to any further evaluation. Hence, according to the Cartesian definition, they are not conscious activities (cogitationes). Moral conscience is God's ideal and objective knowledge and evaluation of our actions. Accordingly, consciousness may be said to be God's ideal and objective knowledge and evaluation of our thoughts.

Let me now, on the basis of this understanding of 'consciousness', address the main topic of this essay. According to the Cartesian definition as I have interpreted it so far, conscious activities are conscious insofar as they are subject to a certain kind of evaluation. This is not yet a satisfactory account, since everything may be evaluated in some respect. For instance, I may attribute a value to my respiration. Who

---

[3] *Essay Concerning Human Understanding* II,xxvii,22.
[4] *Romans* 2,14–6.
[5] *Sermo* 252,7,7, *Patrologia Latina* 38:1176.

knows what values God attributes to things? It says that everything God made is good. So does not everything have some value? We will have to further specify the kind of evaluation that is relevant here.

I will do this by claiming that the purpose of attributing a value to an activity in the relevant sense must be to *correct* an agent. Accordingly, an activity is conscious insofar as it has a value whose attribution constitutes a sanction. When we call actions morally good or bad, this involves a critique of the agent. The point in evaluating actions as good or bad is to encourage further good acting and to discourage further bad acting. The same applies to thoughts and their specific value. That a thought is false means that I should not entertain further thoughts of the same type.

In the beginning of his letter to the Romans, Paul insists that we are subject to the very same standards that we use in judging others. Abelard emphasizes this point in his commentary on *Romans* 2,15.[6] He concludes that the moral conscience to which Paul refers primarily testifies the validity of the Golden Rule. By acting in a certain way, that is, we acknowledge rules that we also must apply to other agents. Conversely, when we evaluate the actions of others according to a certain standard, then we must apply this standard to our own actions. Now if conscience is the implicit knowledge of the Golden Rule, then one of its primary objects must be the comparability and equivalence of actions. But two actions are equivalent in this sense if they have the same value. Knowing and acknowledging the Golden Rule thus amounts to identifying actions in terms of their common value. The Golden Rule tells us to abstract from particular agents and consider only their action. We shall consider actions as things that can be done by any agent in the same way. In a word, following the Golden Rule is one way of considering actions to be repeatable.

Abelard is well known for stressing the related point that actions can only be attributed a moral value in terms of the intention from which they result. More specifically, Abelard claims that the moral value of an action depends on the description under which the agent consents to it. Further, he explains what it means to consent to something by reference to a tendency. We consent to an action if we are 'inwardly ready to do it

[6] *Corpus Christianorum, Continuatio Mediaevalis* 11, p. 86.

when given the chance'.[7] Hence, by depending on the agent's consent, the value of an action ultimately depends on the tendency from which that action results.[8]

The reason why the intention with which an agent does something is important is that it may lead to further actions of the same kind. Therefore, according to Abelard, a particular action has its moral value only insofar as it indicates a general tendency of the agent to do further things of the same kind. In contrast, actions that do not result from any general tendency have no moral value. They are just accidental and not intentional. The reason is simply that there would be no point in sanctioning them. If an action does not result from a general tendency that might also be actualized in the future, then we will not achieve anything positive by blaming the agent. But encouraging and discouraging future actions is the whole point of moral evaluation. This is why we generally describe actions in terms of actualized intentions.

Hence, when the Golden Rule tells us to consider actions to be repeatable, this does not imply that we shall pay no attention whatsoever to the agent. The converse is closer to the truth. Since we must evaluate an action in terms of the tendency of the agent to continue doing things of the same type, we cannot evaluate an action without evaluating an agent. This is why we punish people for things they have done in the past.

According to Paul, moral conscience testifies the possibility of treating actions in terms of their moral value. Abelard adds that actions have their moral value only insofar as they result from intentions. Intentions belong to agents. But conversely, an intention can only be specified by referring to a certain kind of action. It is the intention-to-do-such-and-such. Hence, actions have their moral value by virtue of being of a kind that will eventually be instantiated again. They have their moral value insofar as they are repeatable. They are sanctioned only as such. When we correct an agent, we want to alter his intention, such that he will tend to act differently in the future. I have argued that in the definition that Descartes gives of 'cogitatio', consciousness functions in a way very

[7] *Ethics*, ed. Luscombe p. 14,17–19.
[8] *Ethics* p. 12.

similar to Pauline moral conscience. It does not concern the moral value of our actions, but also the specific value of our thought, which may be their truth, coherence, or originality.

Let me therefore conclude with the following thesis: For an activity to be conscious is to be *corrigible*. The evaluation that is relevant here takes the following form: "should there be more of this?" It would make no sense to try to correct a particular, non-repeatable action. A particular action happens only once, and when it has happened, there is nothing about this past particular happening that could be corrected. When we sanction an agent for what she has done, we are interested in correcting the intention and tendency that has led to her past actions and that may lead to actions of the same kind in the future. An activity is thus conscious insofar as it may be sanctioned with a view to its repetition. In a word, it is conscious insofar it may be corrected.

Some readers may want to object at this point that Descartes distinguished the mental from the physical by claiming the exact opposite: that we cannot be mistaken about the contents of our minds. Hence it seems that for Descartes, the mark of the mental is incorrigibility and not corrigibility. That is, conscious activities are unlike other activities in that certain knowledge claims about them cannot be overridden.

However, this is not even an accurate account of what Descartes explicitly says. Descartes does not claim that we cannot be mistaken about the contents of our minds. In the *Regulae*, he treats the inner and outer sense on the same level. Both may deceive us.[9] And in a letter to Mersenne, he approves of a saying attributed to St. Ambrose and Augustine: 'our thoughts are not within our power, and they often confuse our minds, leading it where we do not want to be'.[10] The point that Descartes makes is not that as a matter of fact, we cannot be mistaken about our own thoughts and actions. It is rather that our own consciousness of what we do cannot be superseded by any other source of knowledge. Our knowledge of what we are doing may be mistaken, but still it must not be overridden. Anyway, overriding self-knowledge is not what

[9] *Regula* 12, *Oeuvres* X:422–3.
[10] Descartes, *Oeuvres* III 248–9; Ambrose, *De Fuga Saeculi* 1, *Corpus Scriptorum Ecclesiasticorum Latinorum* 32,2:163.

we do when we correct someone. Rather, in order to correct an agent, it is important that he remains the agent of his actions. Once we deny the power of the agent over her own actions, we cannot any longer expect to change her future behavior by changing her intentions. It is indeed a regulative principle of moral evaluation that first-person contemporaneous reports about actions must never be thrown out. But that does not mean that we cannot correct agents with regard to their actions. The contrary is true.

I have argued that conscious activities are conscious insofar as they are subject to an evaluation with a view to their repetition. This evaluation need not actually take place; that is, it need not itself be a conscious activity. The point is rather that conscious activity has a value, whether this value actually is attributed to it or not. It has this value insofar it results from a general tendency of the agent to engage in further activity of the same type. The point is not that being subject to the appropriate kind of evaluation is more basic than being conscious. This may be an interesting point to make, but it would need more argument. On the face of it, we would rather say that we evaluate actions and thoughts only because they are conscious. This might constitute a circle, but it will be benign one. It will still be a defining feature of conscious activities that they can be evaluated with a view to correction. But from this it already follows that incorrigibility is not the mark of the mental, not even for Descartes. We individuate activities as conscious activities, that is, in terms of their specific value, because we are interested in correcting people's behavior. Conscious activities are instances of types because they are individuated with a view to their repetition. Other mental entities may be defined in terms of conscious activities. Therefore, corrigibility is the mark of the mental.[11]

[11] This paper was written under the auspices of the Wolfgang Paul Program of the Alexander von Humboldt Foundation and the project "Forms of Life" sponsored by the Volkswagen Foundation. Some formulations used here were deliberately assimilated to formulations that Richard Rorty has used in an essay with a very similar title.