Attributing Responsibility to the Narrative Self

David L. Thompson

1 Introduction: The Question

The question which my paper will address is best raised by considering a number of concrete scenarios.

Kenneth Parks, a 23-year-old Toronto man, was suffering from severe insomnia caused by joblessness and gambling debts. Early in the morning of May 23, 1987 he arose, got in his car and drove 23 kilometres to his in-laws' home. He stabbed to death his mother-in-law, whom he loved. He then went to the police, and, in a confused state, said that he was afraid he might have killed someone. The defence relied on expert psychiatric opinion to argue that Parks had been in a somnambulist state and so was innocent of any crime. In 1988 the jury agreed and acquitted him. The case was appealed to the Canadian Supreme Court which, in 1992, sustained the jury's acquittal. (https://db0nus869y26v.cloudfront.net/en/R_v_Parks)

Another scenario (hypothetical): consider of the case of a woman mystified to find new clothes in her closet, clothes she has no recollection of buying. Her own signature on her credit card slips, however, makes it clear that she was the one who bought them. This is typical of purported cases of multiple personality disorder. Her body was involved in the purchase, although the actual action was performed by one of her alter-egos. She denies all responsibility for the purchase; the action was performed by someone else.

For a third scenario, consider a person who, under hypnosis, promises to sell his car for one dollar, indeed signs a legally binding contract to do so. The next day, discovering that he had been hypnotized, he repudiates the contract, claiming that he was not responsible for the promise and so is not bound by it.

Finally, as my last scenario, consider a “normal” case. I borrow $20 from a stranger and promise to repay it the next day. When we meet the next day, the stranger says, “Weren’t you the one who promised to repay me $20 today?” I respond, “Oh, Yes! I’m the one” and hand over the cash.

Each of these scenarios involves the identity of a person (or self -- I’ll be using the two terms interchangeably). Clearly, however, in these cases it is not a matter of the identity of a body that is at stake, but rather the question of which actions are those of the person involved, which ones are his or mine, and which commitments the self is responsible for. An action is mine, that is, I “own” it, if I am responsible for it. But what is it to be responsible for an event? The notion of responsibility that I want to examine includes legal and moral responsibility, but is not restricted to them. The concept I want is more fundamental than either, for both moral and legal responsibility depend on what one might call basic responsibility.

When a white pawn is moved in a game of chess we distinguish between the case when it is moved by an earth tremor or by one of the players. Even when one of the players is the cause of the movement, it fails to count as a move in the game if the player’s finger moves it by accident when reaching for another piece. It is a chess move only when the player explicitly or implicitly declares ownership of the move, that is, acknowledges responsibility for it. In this case there is no moral or legal issue at stake: to say I’m responsible for the move is not yet to evaluate the move in any way. It is simply to claim that the move is an action, and to attribute the responsibility for that action to the same self which has made or will make other moves with other white pieces.

The notion of a continuing self responsible for its actions is a prerequisite for any subsequent question about moral or legal responsibility. This is the point that Paul Ricoeur is making when he distinguishes between

description, ascription, and prescription.[1] We can *describ*e an event: the pawn moves; we can *ascribe* an action: that is the player's move in the game; or we can *prescribe* an action: that is, refer to what the person morally ought to do. Without the prior notion of an agent or self to whom we can ascribe basic responsibility, there would be nothing to which we could later attribute moral or legal responsibility.

My aim in this paper is examine what it is to attribute responsibility, in this basic sense, to a self. First I will outline a metaphysical analysis of identity, and then I will offer some criticisms of that position. Finally I will present a positive exposition of the narrative theory of selfhood and show how it might account for the attribution of responsibility to a self.

2 The Metaphysical Way

Let me start by sketching an approach that I think is wrong. Some philosophers believe that to understand the attribution of responsibility we must first establish what it is for an object called the self to remain self-identical over time. Historically, this is how John Locke first introduced us to the issue. Locke claims that "person" is a forensic concept, and that we should only reward or punish someone for an act if they are "identical" to the person who committed the act. I care about reward or punishment, and indeed, my survival in the future, only if it will be the selfsame *me* who survives.

Parfit continues this tradition when he asks whether, if a teletransporter disassembles me and re-creates me atom by atom on Mars, the new me is numerically identical with the earthling or only qualitatively

1 "If ascribing is not describing, is this not by virtue of a certain affinity, which remains to be clarified, with prescribing?" Ricoeur, *Oneself as Another*, 99.

similar. If the latter, then I have not been transported to Mars, but have died. If I have not survived, however, then presumably the Martian me has no more responsibility for the actions or commitments that the earthling me had made than my qualitatively identical twin would have. Whether I am responsible or not depends on the prior question about whether I have remained numerically self-identical.[2] From this point of view, responsibility depends on and presupposes identity

This approach – the one I am challenging – might be called the "metaphysical" approach. This position thinks of the self as an objective entity which remains identical with itself over time and the approach conceives of responsibility for an action as a property or accident of that entity. To put it in logical terms, responsibility for an action is a property which is predicated of a subject. Whether a proposition attributing this property to a subject is true or not depends on how the world is in itself; there is in reality a fact of the matter upon which the truth-value of the proposition depends. Locke argues that the metaphysical substrate to which the property of responsibility is attributed is neither the body, nor a spiritual substance, such as a Cartesian mind.[3] Whether such an objective self is mental, physical or something else is a secondary issue; the crucial point for the metaphysical approach is that it is the substrate that accounts for the self's enduring identity.

This approach is inherited from Descartes' substantialism. Descartes holds that, while the body is a substance, the mind cannot be identified with it; the mind is an independent substance in its own right. But because of his epistemological concern about certainty, Descartes has a peculiar notion of this substantiality. He claims to be certain of his existence at the instant when he thinks about himself, but, since memory is subject to doubt, it is

2 Parfit, *Reasons and Persons,* 197 and following.
3 Locke, *Essay*, Chapter XXVII.

only by reliance on the self-constancy of God that his duration over time is assured. For Descartes, no substance has the power to maintain itself in existence from one point of time to the next, so only the continuous re-creative intervention of God ensures the appearance of continuous self-identity over time.[4] Such a "punctualist" approach raises a fundamental question which will dominate discussion for the subsequent three centuries: if at a point in time God re-creates Descartes anew and creates in him "memories" of experiences at previous points in time, is the appearance of numerical identity merely an illusion, since the new Descartes is only qualitatively similar to the previous one? Parfit, for example, places the question in the technological context of teleportation rather than Descartes' theological context, but his concern that my Martian self might be only qualitatively identical with my earthly self and not numerically identical with it, continues the Cartesian metaphysical approach.

This kind of punctualism also underpins the notion of mental state that has developed in recent decades in much of the philosophy of mind. Analytic philosophers of mind have investigated the relationship between brain processes and mental states. Functionalism, which I consider the most plausible approach, understands mental events as functional states which might conceivably be realized in multiple, alternative brain processes. This cognitive model explains action in terms of the current mental beliefs and desires of the agent. In principle the approach could be used to account for mental states in nonhuman animals, although there has been significant debate about the status of beliefs in organisms without language. The focus of this approach has been on individual mental states, time-slices of life, as it were, and their relationship to brain processes. What makes the approach punctualist is that each individual current mental state is considered primarily in its relationship to brain processes while the dependence of these

4 Descartes, Meditation III, para. 31.

states on the ongoing life of a self dispersed in time is seldom considered. Indeed I would like in my vocabulary to distinguish a functionalist notion of *mind* from the alternative concept of *self* which I'm examining.

It is not impossible for this punctualist metaphysical identity approach to account for the problems of responsible selfhood raised by the scenarios at the beginning of my paper. It might find some way of establishing that Kenneth Parks, the sleepwalking killer, was not the same person asleep as when awake, and so the waking Parks was not responsible for the actions of the sleeping Parks. I suppose one could say that he did not have any intention, since he was asleep. Or perhaps one could claim that he didn't have the belief that it was his mother-in-law he was killing. Or one could devise some system of excuses for actions by sleeping selves. But all of these seem to me to be a*d hoc*, and forces the phenomena into a Procrustean bed because of dogmatic attachment to a philosophical theory. The narrativists think they have a better way for understanding responsibility and criticize the metaphysical identity approach on this basis.

3 Narrativist criticism of the metaphysical approach

From the viewpoint of the narrativist position there are a number of criticisms which can be made of the approach that I have outlined. The first is that the metaphysical position treats the person as an object which remains identical in itself through time, regardless of the person's knowledge. This criticism simply draws out the implications of an argument already presented by John Locke. Locke argues that personal identity, as a forensic concept, must not be confused with material identity as a physical object, with biological identity as an organism or with spiritual identity as a substantial mind. His argument uses thought experiments of body exchange, or of the soul's reincarnation, to appeal to the intuitive injustice of punishing a person for crimes they have no memory of or are not conscious of. But the

essence of his argument is to overthrow *any* attempt to conceive of the self as an object or substance of any kind. No matter what kind of object we might conceive the self to be, we can always devise a thought experiment in which lack of memory or consciousness makes that object's identity irrelevant from a forensic point of view, that is, with respect to selfhood.[5] Whatever the notion of self is trying to illuminate, any appeal to substantial identity will fail to enlighten us.

A similar criticism is offered by Schechtman. Her claim is that punctualists are asking the wrong question. Punctualists ask the "reidentification question:" under what conditions can we identify person-stage A with a later person-stage B. Narrativists, on the other hand, ask the characterization question: when can we attribute a characteristic to a self? The logical structure of the two questions is quite different. The narrativist question does not attempt to conjoin two entities, but to conjoin a property to a self. Under what conditions can a person be said to be responsible for a previous action? When can we attribute to someone the characteristic of "being committed," for example, being bound by a promise? This is quite a different question than the reidentification question, so we should not be too surprised that the kind of answers offered are also different. Schechtman claims that, as philosophers, we became interested in the question of selfhood in the first place because of four concerns:

1. We are interested in our own survival;
2. We need an account of moral responsibility;
3. We need to explain the unique concern we have about our own life;
4. We want to know to whom compensation, that is, reward and punishment, is due.[6]

---

5 "Self depends on consciousness, not on substance. Self is that conscious thinking thing,--whatever substance made up of, (whether spiritual or material, simple or compounded, it matters not)--which is sensible or conscious of pleasure and pain, capable of happiness or misery, and so is concerned for itself, as far as that consciousness extends." Locke, Essay, 156. (Ch XXVII, para 17.)

6 Schechtman, Constitution, 2.

Only the characterization question, she claims, engages seriously with these issues. The punctualists' reidentification question invariably disappoints because it attempts to treat selfhood as an object in-itself, a fact of the matter, rather than as a feature of experience.

A different line of criticism can be directed at the belief/desire explanation of action. This approach isolates a particular moment or action in the life of an individual and explains the action on the basis of the conditions present at that moment. Indeed Parfit makes it an explicit thesis of his "reductionist" theory that there are events – including mental events – that can be specified independently of reference to the ongoing life of any person.[7] But this punctualist approach seldom if ever makes sense in a human context. If I repay you $20 because of the belief that I owe it to you and the desire to be honest, then this explanation of the action, and indeed the very action itself, only makes sense in a wider temporal context. We need to understand that yesterday I borrowed $20 from you and promised then that I would repay it, and to understand that I am concerned to preserve my reputation of honesty in the future. Without this temporal context the movement of hands and paper would not even *be* an action of "repayment." I cannot just, out of the blue one morning, wake up owing you $20 and repay you. Even if, during sleep some kind of post-hypnotic suggestion was implanted in me that I had promised you $20, I must attribute that promising to myself rather than to the hypnotist – mistakenly as it happens – in order to believe that I owe the money to you. To understand action we need to incorporate the notion of enduring responsibility into our conceptual scheme, over and above momentary states of belief and desire. Actions, beliefs and desires only make sense in the context of a unified life.[8]

---

7 Parfit, Reasons, 340.

For example, to understand my action of writing a paper for a conference it is not enough to explain my intention as a psychological function realized in certain brain processes, though this is not incorrect. My intention cannot occur out of the blue, but is only *this* intention if writing such a paper is in my repertoire, is among my competencies. If I had not been trained as an academic, if I didn't understand the point of a conference, if I didn't have some minimal grasp of how to structure a talk, I could not be said to have the intention of writing such a paper. So Parfit's reductionism is wrong: the individual intention, the momentary mental state, can only be that kind of state within the context of my personal history no matter what brain process it is realized in. Similarly there is also a social and institutional context which has constructed the notion of a conference, the norms for academic talks, and disciplinary boundaries within which the content of my paper makes sense.

One attempt to respond to these criticisms of the metaphysical identity approach might be to supplement the punctualist notion of selfhood by finding some way to incorporate the concept of responsibility. More interesting is to consider the possibility that the theory is fundamentally flawed in a way that *ad hoc* fixes cannot repair. The better alternative is think of responsibility as not so much an incidental property of the self as that by which the self is constituted. Where the metaphysical approach considers numerical identity to be a precondition for responsibility, the most central idea I want to examine in this paper is that the exact opposite is the case: instead of a self-identical self being a condition for responsibility, it is the narrative attribution of responsibility which sets up a continuing self in the first place.

---

8 MacIntyre refers to the "concept of a self whose unity resides in the unity of a narrative which links birth to life to death as narrative beginning to middle to end." MacIntyre, *After Virtue*, 220.

4 Positive Account of Responsible Selfhood

Enough of criticizing the punctualists! What kind of positive theory of the self do narrativists offer that might better account for the problems of responsibility in the scenarios with which I started this paper? First we need an account of narrative theory; then we can move on to see how it accounts for responsibility.

It would be unwise to assume that all so-called "narrativists" are in agreement with each other. There is no one canonical narrative theory of selfhood: there are a cluster of theories, not always in agreement with each other, that all employ narrative as a means of understanding selfhood. Ricoeur maintains that the self has a structure which is analogous to that of a fictional or historical story. McIntyre seems to hold that the self actually *is* a kind of narrative. Schechtman says that a self must be capable of exhibiting a narrative even if that narrative is unconscious or available only to others. Dennett seems to understand a self as a fictional character within a narrative generated by an impersonal brain processes.[9]

All these theories have two central features in common, that I will consider in turn. First, a narrative mode of explanation places each action into the context of the history of events leading up to it and of the future projects towards which the self is oriented. Secondly, and this is the feature I wish to focus on, to be a self is to integrate one's life over time, making commitments to the future and accepting responsibility for the past.

First, as an example, consider a narrative explanation of Mary attending a human rights demonstration. The explanation might note that Mary has been horrified by speaking with victims of torture, she has a vision of a future world in which human rights would be respected, and she believes the demonstration which she has seen advertised will help. The

9 Dennett, *Consciousness*, 412-430

notice advertising the demonstration could be offered as one cause of Mary's action, but the narrative approach goes beyond this causal account, however correct, to explain *why* this notice acted as a cause of Mary's action, even though other people might have been unmoved by it. Similarly, giving a reason for her action -– the value she places upon human rights and a future torture-free world –– while it offers a correct, teleological explanation of her action, fails to show why this is a value to which she is committed. The narrative explanation goes beyond any simple appeal to cause, to reason or to a belief/desire complex by placing each of these in the historical context of Mary's life. The integration of her action into a story about how she has come to be the person she currently is, and what her values and future aspirations are, appears to be the essential requirement for the account to be labelled narrative.

But, secondly, this narrative history should not be understood as a series of impersonal, objective events but rather of actions for which the self is responsible. What is it to be responsible?

4.1 What Is Responsibility?

What is responsibility? Let me recall that the concept involved is not that of legal or moral responsibility, but that basic responsibility by which I accept ownership of all my actions, by which I claim them to be mine and attribute them to myself. Paul Ricoeur takes promising as an icon for all such commitments made by a self, so let me start by investigating promising. I am bound by a promise only in so far as I currently interpret my previous experience as the making of a promise. If I come to believe that the words uttered yesterday were the result of hypnosis, undue pressure, the influence of a drug, or similar circumstances, then I do not consider myself bound. I am bound by a promise only insofar as I currently interpret it as a promise. I could not in fact, objectively, be bound by a promise when I

believe myself not to be. There is no gap here between reality and my knowledge of reality; it only *is* a promise if I *know* it to be a promise. Like all responsibility and commitment, the condition of being bound by a promise can never be simply discovered; it is *attributed*. But what does that mean?

The distinction between attribution and discovery can perhaps be elucidated by an analogy with the legal system. Consider the question of whether same-sex couples can legally marry. Some people appear to think that there is a fact of the matter that politicians and parliaments should acknowledge. I think this makes no sense. This is a matter for decision: it is up to parliaments to create laws on the subject. Such laws might be wise or unwise, fair or unfair, but they cannot be false by failing to correspond with some pre-existing reality.

Another analogy is with political institutions. The Governor General has the power to sign acts of Parliament into law. Nobody *discovers* that she had this power all along; rather she is granted this power through an institutionalized appointment process. The power is attributed to her.

In a way analogous to the legal or political process, the condition of being responsible is attributed to me by a kind of interpretive process and is not the discovery of a pre-existing fact of the matter. My current responsibility for any past commitment is a matter of me attributing an action to my self, that is, interpreting my status today on the basis of my acknowledgement of yesterday's action as mine. To understand myself as bound is for me to accept the previous commitment as mine.

The metaphysical approach must claim that being the same self today and yesterday is a fact of the matter, whether anyone knows or acknowledges that fact or not, and it is this objective identity that underpins responsibility. The narrativist position is that one being bound by a commitment, for example, a promise, depends on one's interpretation, on whether one attributes the act of commitment to one's self or not.

The metaphysical stance is realist. The essential feature of metaphysical realism is that an object is what it is, and has the properties that it has, prior to and independently of any perception, knowledge, or recognition of the object. The object is what it is "in itself" without regard to any subjective intervention. From this metaphysical point of view it make sense to say that one is responsible even though one may be unaware of the fact or does not acknowledge it, even privately. This is what narrativist attribution is denying.

If we push this narrative analysis through to its conclusion, it is not just responsibility which is attributed but the very status of being a self. The narrativist position is that sameness of self is of the order of attribution: to declare myself bound by a promise today is to attribute the status "myself" to the originator of yesterday's promise. Responsibility is not so much a property of a prior, independently existing self as that by which the self is constituted. In the case of Kenneth Parks, his mother-in-law's death is not attributed to him, and so as a consequence, it was not him, the self he is now, that performed the action. One's unity over time is not a metaphysical self-identity, but an interpretive process in which one constitutes oneself as the same. It is the metaphysical self that Schechtman is rejecting when she insists, "An identity in the sense of the characterization question, is not, I claim, something that an individual has whether she knows it or not, but something she has because she acknowledges her personhood and appropriates certain actions and experiences."[10] (Schechtman, Constitution, 95).

The self is not a metaphysical object with objective attributes but a subjective entity for which the condition of being bound is essentially linked to self interpretation. If today I attribute yesterday's experiences and actions to *me* this should not be understood as a discovery of objective

10 Schechtman, *Constitution*, 95

properties that some pre-existing self has. Rather the process of attribution creates the unity of the self in the first place.

4.2 How Could Responsible Selfhood Originate?

One way of understanding how attribution creates the self is by examining the conditions under which a new self might come into being. Computers are currently not considered to be responsible for their actions, that is, they are not selves. William Bechtel, however, has discussed the conditions under which we might in the future hold a computer responsible for its actions. He argues that if a computer is programmed in such a way that it does what it's told, then the responsibility for its actions rests with the programmer not the computer. For the computer to be responsible, it must be an intentional system which has beliefs based on symbols which have meaning for the computer system. "[Such symbols] should be thought of as having meaning for the computer when the way the computer system uses these symbols is adaptive for it in its environment and is being shaped by the environment through some form of selection."[11] If however the computer learns to program itself, as he suggests a neural net does, then it *is* conceivable that we could hold it responsible for its actions. The main factor here seems to be progressive learning within the history of the individual. If on the basis of its past interactions with the world, a computer reconfigures its own program so as to approach the world differently in the future, then, the computer itself can be held responsible for these future actions. He says:

> What we are claiming in attributing responsibility to an intentional system is that it was because the agent was of this kind that *it* made the decision it did. This explains how the decisions of an intentional system stem from it; what remains is to account for the respect in

11 Bechtel, "Attributing Responsibility to Computer Systems." 302

> which the decision was under the system's control. The sense in which the decision of such a system is under its control is that it responds the way it does because of the way it adapted to its environment. Had it evolved differently, then it might have decided differently. Moreover, it had within it the capacity to learn and so adapt to its environment in different ways.[12]

Bechtel's position amounts to the claim that such an adaptive computer would be its own programmer, so that the responsibility for its current programmed responses rests with itself in so far as it designed its own program in the past. In that sense it is self programmed, and so self responsible.

But this position, as Bechtel himself points out, is open to an objection: the system had no choice about what it learned, so is the question of ultimate responsibility not simply pushed further back? Bechtel's response is that human beings are in the same boat, yet we do not deny responsibility to them: "We do not inquire further as to whether they [responsible humans] choose what they learned."

I think Bechtel's response is inadequate. It may be true that we do not normally inquire about whether humans choose what they learned but this is only because we assume that their learning was itself a responsible act. If we believe that an individual's "learning" took place in a brainwashing camp, was due to membership in a cult, or was hypnotically induced, then we would indeed inquire further. Normally learning is done by a self who accepts, rejects, or interprets the information fed to it on the basis of the values, information, and world view it already has. That is, the self is more or less responsible for what it learns from any given situation. In cases where "learning" takes place in a non-responsible manner, then we are likely

12 Bechtel, "Attributing Responsibility to Computer Systems." 305

to interpret the future acts of such individuals as “programmed” responses for which they are then, indeed, not held responsible.

A more creative response to Bechtel’s problem is to think of the constitution of a self as a progressive, bootstrapping, operation. A baby has no responsibility for its actions but as it grows older the kind of character which is has learned – not responsibly – takes on an ever increasing role in interpreting the information it receives from the world, so that its learning becomes progressively more responsible. Hence the structure of its future interpretations of, and responses to, the world become more and more the result of its own prior responsible actions of learning.

Indeed, the self should be understood precisely as such a structure, that is, as a mode of interpretation of perception and as a set of values and dispositions on the basis of which it responds to the world, and it is the previous commitments by the self which are responsible for this current structure. If you ask me for $20 today, how I interpret the request, and how I respond to it, would be determined by the promise I made to you yesterday to pay you that sum. By yesterday’s promise I have, as it were, programmed myself to respond in a particular way today. The kind of self I am today, one whose actions are bound by a promise, is this kind of self because of the promise I gave yesterday. “Learning,” in the sense I need here, is not simply the accumulation of information, the addition of data which a program may then process, but rather a way of changing the program itself so that I respond to the world as a different kind of person. A system which responds to its interactions with the world by changing its own modes of response in the future is what I am calling a responsible self, and, I agree with Bechtel, there is no reason in principle why a future computer system should not qualify.

5 Conclusion

The self, then, to conclude, is not a metaphysical, self-identical object but a mode of organization of an organism, a temporal structure in which the past, present and future are held together as a unity by the key relationship of responsibility. Responsibility is not an objective relationship, but is based on the way in which the current self interprets previous events, attributes them to itself and so commits itself for the future.

David L. Thompson
Philosophy, Memorial University of Newfoundland
Visitor, Victoria University of Wellington
2004

Bibliography

Bechtel, William, “Attributing Responsibility to Computer Systems,” *Metaphilosophy* 16:296-306, 1985

Dennett, Daniel C., *Consciousness Explained,* Little, Brown & Co. 1991

Descartes, Rene. *Meditations on First Philosophy*. Any edition.

Locke, John, *An Essay Concerning Human Understanding.* Oxford University Press. 1975

MacIntyre, Alasdair, *After Virtue:A Study in Moral Theory*. Third Edition. University of Notre Dame Press, Notre Dame, Indiana. 2007

Parfit, D., *Reasons and Persons.* Oxford: Clarendon Press. 1984

Ricoeur, Paul. *Oneself as Another [Soi-meme comme un autre]* translated by Kathleen Blamey. The University of Chicago Press. 1992

Schechtman, Marya, *The Constitution of Selves*. Cornell University Press. 1996.