The Identity of the Self over Time is Normative

David L. Thompson

1. INTRODUCTION: THE PROBLEM

What is to be a human person? Since the cognitive revolution half a century ago, the analytic philosophy of mind has interpreted the question as the mind-body problem: how are mental states that have cognitive or semantic content related to their concomitant brain states or causal neural processes? Let me call this the vertical problem. Functionalism seems to offer the most convincing account of this relationship: the mind is not the brain; the mind is what the brain *does*.

Functionalism by itself, however, has two striking limitations. First, it has trouble accounting for consciousness – a problem I will be ignoring in this paper. Secondly, it tends to neglect the integration of mental states over time; let me call this the horizontal problem. What is the temporal structure of the one life a person leads from birth to death? I will refer to this structure as the *self*.

My central claim in this paper is that the temporal unity of a self is an external unity, a unity imposed from outside the individual – beyond both brain processes and functions – by the institutional context of social norms.

I proceed, after rejecting the empiricist approach, first by examining the nature of commitment as a social institution, using promising as my leitmotif, and then, secondly, by inquiring about the internal structures that an individual would need to have to be able to live up to these commitments. I rely on two analogies, one biological, and one from computers, to elucidate these individual structures. Finally, I draw the institutional and individual analyses together in my proposed contextual account of selfhood.

2. INADEQUACY OF THE EMPIRICIST ACCOUNT

So what kind of account of the integration of self over time one give give? One approach – let me call it the empiricist approach since it originates with Locke – claims that attempts to account for self-identity over time by appeal to the unity of a substance, whether physical or mental, must be rejected on the grounds that if I have no consciousness of past events, I cannot be held responsible for them.[1] So, rather than assuming a substantial self, empiricism takes as its starting point the existence of mental events – ideas, impressions, mental states, etc. – which can be defined in isolation from each other and without any initial reference to a unifying principle. It then looks for some kinds of relations between these events which will, secondarily, unify them into the one self. Locke – at least as traditionally interpreted – understands the unifying relationship to be one of memory: two states are states of the same self just in the case when one state is the memory of the other. Since I can forget what happened to me, however, Parfit considers Locke's direct memory connections to be inadequate and offers instead a series of overlapping memory chains that he calls psychological continuity.[2] For example, a soldier, victorious in battle, remembers being flogged as a boy, but has forgotten the whipping by the time he retires as a general. (The example is Reid's.[3]) Nevertheless, as long as the general remembers the victory, the boy and the general are the same self.

How does psychological continuity work? For instance, what relationship exists between the soldier's mental state of remembering the whipping and his mental state of perceiving the victory, the state that the general will later remember? The two states occur at the same time, and in the same space, namely the space of the soldier's brain; that's what overlapping means. But how does their simple contiguity assure us of the temporal unity of the self? That billiard ball A hits B and coincidentally, at the same time on the same table, ball C hits D, hardly unites A and D in any significant way. Perhaps Parfit is taking for granted some stronger conception of the synchronous unity of the self that he is levering into diachronic unity. But on the surface it appears that all he is appealing to is synchronous contiguity.

[1] Locke, *Essay*, Chapter XXVII.

[2] Parfit 205

[3] Reid, *Essays* III. 6, p. 276

In any case it is not clear that simple memory connections themselves escape pure contiguity either. Locke's approach has been criticized as circular on the grounds that our current mental state is a real memory only if it is of *my* previous mental state, which presupposes the temporal unity of the self that Locke is attempting to establish. Parfit tries to rescue Locke from this circularity by insisting on the existence of the right kind of causal connection between the remembered and remembering states. However, if we understand causation as Humean constant conjunction, then again we are relying on a kind of contiguity. But whatever account we give of causation, empiricism understands the temporal integration of the self as based on contingent, empirically discoverable relations (contiguity, causality) between independently definable, atomic mental states.

Psychological continuity includes psychological connections other than memory. Parfit mentions, for instance, the connection between an intention and the action that fulfills it,[4] and casually refers to continuity of character. Nevertheless, his primary paradigm for a connection is that between a mental event and a later recollection of the same event, so his account focuses almost exclusively on memory. If his leitmotif were different, his conclusions might be less convincing. The relationship between promise making and promise fulfillment, for example, cannot so easily be analyzed in terms of independently definable, empirically related events. Let me examine how the promise paradigm might lead us to a very different account of selfhood.

First, the event of fulfilling a promise cannot exist and cannot qualify as a fulfillment except by reference to a previous act of making a promise. Secondly, the relationship between promise making and promise fulfilling is not a causal relationship nor is it based on simple sequence or contiguity. Imagine that I programmed my computer to say yesterday, "I promise to check your e-mail every five minutes tomorrow." Today, when I inquire about my e-mail, I simply get a blank look on the screen. Maybe the problem is memory! So I reprogram the computer to not only make the promise, but also to store that fact in its memory. The next day, instead of the blank screen, I get a statement acknowledging that it had stated the previous day, “I promise....” But my e-mail still doesn't arrive! Simply remembering a promise-statement is not enough for the computer to be obliged. But note that even if I program the computer so that whenever it makes such a statement then it actually *does* check my e-mail regularly, that

[4] Parfit, *Reasons* 205

action is not the fulfillment of a promise but only a causal effect of my programming. The relationship of promise-making to promise-fulfillment is not an empirical one, that is, it is not one of contiguity, simple sequence, or causation. We need a different analysis.

3. THE COMMITMENT PARADIGM

My own analysis is based on commitment rather than memory. I focus primarily on promising which I use as a concrete icon or stand-in for commitments in general. The project of taking promising as the paradigm for the temporal unity of the self appears, on the surface, to have a similar structure to Parfit's memory-based project. In the memory paradigm, it is the link between the two events which accounts for the identity of the self at the two times; in a parallel way in the promising paradigm, it is the link between promise making and keeping that produces the unity of the self over time. In the memory model, only if a current person-stage is in memory-continuity with an earlier person-stage, are they stages of the same person. In the promise model, only if a current event is the fulfillment of an obligation created by an earlier promise, is the promise-maker and the promise-keeper the same self. For both analyses, to avoid circularity, it is crucial that the temporal unity of the self not be presupposed. Parfit, for example, ties himself in knots attempting to relate "quasi-memories" to the events they represent without presupposing that they are already "mine."[5] My own position, therefore, must show that a current self is the same as a previous one only in so far as it is obligated by a previous promise. Let me first discuss the nature of promising itself and then talk about what kind of individual is capable of such a commitment.

a. Promising as an institution

How does promising work? I used a computer analogy above to illustrate why the relation between promise-keeping and promise-making is neither a causal relation nor one based on contiguity. If, as in the computer case, causal processes in my brain made me hand you $100 whenever I said, "I promise to give you $100 tomorrow," the previous day, it would still not count as keeping a promise. That an act has the status of fulfilling a promise is not the effect of a cause. In this, the relationship is very different for how Parfit conceives of memory. So what is the relationship?

[5] Parfit, *Reasons*, 220

Promise making is a performative which establishes an obligation in accordance with a social convention. Promising is an institution which is governed by the rules of a language game. Consider a chess analogy: if my king is threatened and can't move, this is checkmate, and I lose the game. There is nothing logically necessary about this rule nor does it express a causal relation. Rather, it is a convention we have adopted and agreed to live by. Similarly, promising is a convention we have adopted and, once we have adopted it, whenever an individual makes a promise the act establishes an obligation to fulfill it. The institution of promising is the context or horizon within which the acts of promise-making and of promise-fulfilling are defined and the relation between them established.

Let me refer to this way of thinking as *externalism.* What makes a piece of metal be a dollar is the external context of its use, not some internal property of the material object. Thousands of years ago, humans discovered gold, but we cannot say that humans ever discovered money in this way. Humans *invented* money; they established a convention that treated pieces of metal as coins.

From this perspective, we cannot say that promise-making and promise-keeping pre-existed the establishment of the convention of promising. They are not isolated, atomic, separately definable events that were empirically discovered and then later incorporated into a unified structure of promising. It is the promising convention that makes them what they are. Their relationship is not a *contingent* one, as if we just happened to discover that one follows the other, nor is it a *causal* one, for it is not that the promise-making causes the second action to be a promise-keeping.

Consider the money analogy: while it is the financial institution of exchange that makes the disk of metal into money, it would be inappropriate to say that the exchange context *caused* the dollar. The pressing machines in the mint can be said to cause there to be a disk of metal. The disk being money, however, depends on the social context, but this dependence is not a causal relationship. That an act is the fulfillment of a promise depends on there being a previous act of promise-making and it also depends on the context of the convention of promising, but neither dependence is causal.

Another way of explaining the distinction between an internalist and externalist account is by looking at biological functions. Five centuries ago people came to Newfoundland to fish for cod. Only recently have scientists discovered that Arctic codfish can survive freezing ocean temperatures only because they produce a particular glycoprotein in their bloodstreams that functions as an antifreeze. So an individual codfish, call her Charly, survived the recent winter because she had this glycoprotein in her blood. The presence of the glycoprotein has been explained by tracking down the gene in her DNA that codes for this protein. There is a causal chain leading back from this DNA to her ancestors from which she inherited it through some 500,000 generations. Some great great-grandmother cod suffered a mutation, perhaps from cosmic rays, which created the gene for the glycoprotein. This causal chain provides an internalist, causal account that explains why Charly survived last winter.

Note, however, that this story does not account, does not even attempt to account, for the glycoprotein in Charly's blood having the *function* of being an antifreeze. There is no mention of function in the causal chain story. The function is not the effect of a cause as the presence of the glycoprotein is. To account for the function of the glycoprotein we must appeal to a much wider context of factors external to the causal chain. We will need to mention ice ages 2.5 million years ago, the absence of this glycoprotein in the ancestral grandmother’s siblings and their offspring, the consequent differential reproduction rates of these two clades of codfish, and many similar evolutionary adaptive factors. This is an externalist account that the bestowal of functionality on the basis of a context which is external to the direct causal chain that explains the current presence of the glycoprotein.

Now I propose that we think of promise fulfillment as like a function, but an institutional function rather than a biological one. There might be a causal chain between a statement and a later behavior: there is such a chain when I program my computer so that whenever it says, "I promise to check your e-mail every hour," it actually starts to do so. But the existence of such a causal chain, no matter how we rearranged it internally, would not make this behaviour the fulfillment of a promise. Its status as promise-keeping is attributed to it by the external context of the social institution of promising. This is the first part of my argument: promising, and by extension, all commitments, must be understood in a contextual or externalist manner.

b. The Individual: Internal Preconditions

The second part of my analysis examines what internal properties an individual must have for it is to be capable of participating in this – or any -- social institution. How can an individual follow the rules of the game? I look at two analogies, one from biology, and one from computing.

i. Biological analogy: sense organs

First, consider biological sense perception. In a famous passage and diagram, Descartes developed a mechanical account of perception in which a peripheral sense organ, such as a tactile sensor in the foot, is affected by physical stimuli from the world and the effects – the sensations – are transmitted faithfully and unchanged through the nerves to the brain where, perhaps in the pineal gland, the mind perceives them. Let me call this theory the *Faithful Transmitter Hypothesis*. This hypothesis dominated philosophy, usually implicitly, for over three centuries, but it is fundamentally wrong on at least four counts.

1. Peripheral sense organs are not passive but are already pre-tuned by evolution to attend to certain physical stimuli rather than others. In particular cases, sense organs are often actively searching for stimuli that they have pre-categorized in advance. We feel, in the dark, for the handle of a door, or we scan a crowd of faces expecting to see a friend.

2. Far from being faithful transmitters, sense nerves actively process the information they are transmitting. They assign spatial parameters, for instance, to visual or tactile inputs. The visual system compares input from two retinas to produce the perception of a single three-dimensional object. A facial recognizer does not present us with pink or brown surfaces or shapes, but with the familiar face of a friend.

3. What is perceived is not normally a neutral object or situation. The result delivered to perception is categorized, not in physical terms, but with categories that have evolved to meet the pragmatic needs of the organism, for example, the face of a con-specific. What is more, the object perceived is imbued with values significant to the perceiver.
A predator is perceived as dangerous. Those who go to horror movies know that

monsters *appear* threatening; we do not perceive them initially as neutral and then make a secondary value judgment about them. This is true even in the simplest cases. If I place my finger on a hot stove, I don't just receive the information that my finger is burning. My experience is rather of something which is painful and noxious and about which I should do something immediately. That is, the information has been evaluated prior to it being perceived.

4. Finally our perceptual systems do not deliver us an internal mental image of what is perceived; the object perceived is located in the world, not in the mind or the brain. While internal processes such as low blood sugar or stimulation by sexual pheromones have important roles to play, what we actually perceive is appetizing food or the attractiveness of a potential sex partner. These are in the world, not in our minds.

So we can conclude that the Faithful Transmitter Hypothesis is an incorrect account of sense perception.

I propose that we think of a promise made by an individual as analogous to the creation of a new sense modality in that individual, provided we do not analyze it by the faithful transmitter account. If my character is such that I habitually fulfill my promises, then making a promise categorizes in advance how I will experience the world. From here on I will perceive my world, in particular my social world, in a different manner. Yesterday I promised you $100. Today when I meet you the situation I experience is pre-categorized as a promise-keeping opportunity. I experience you not as a physical object but as a promisee. Nor do I experience this situation as neutral or simply factual; it comes pre-evaluated as imbued with a sense of obligation. The situation calls on my trustworthiness, and carries a demand for me to fulfill yesterday’s promise. It is not just that a previous internal mental event of promising gets recalled by memory – although that may also happen – it is that today’s events appear differently to me, for my promise has restructured my way of experiencing the world. Given my habitual character, the world in which I live presents itself to me as already evaluated; in particular, some situations are experienced as involving obligations.

What I am claiming is that when an individual becomes capable of making a commitment that restructures its future way of experiencing the world in this way, then it has developed the ability to participate in the institution of promising. In other words, the individual is now a “self.” The structure of a self’s integration over time is the structure that allows the making of a commitment today to determine how it will perceive and act tomorrow. If the way one perceives, evaluates, and acts can be called one’s nature or character then a self’s character today is a product of its past commitments. The narrative approach has this much right: who a self is must be understood on the basis of its history.

It is this restructuring capacity that the memory paradigm of selfhood neglects. Locke and Parfit conceive of memories as inert traces or objects possessed. The retired general still “possesses” his memory of the military victory, but has “lost” the memory of being flogged as a boy. They treat a memory as a mental content, like a ball on a billiard table. A promise, however, is not a passive object sitting in the mind: it is more like a new mental form or structure that changes the future perceptual and responsive capacities of the self.

The most fundamental way that the promise paradigm differs from the memory account is that it is normative. An individual restructured by a promise experiences the promise-fulfilling situation as obligatory: her experience includes an evaluation of the situation. This stems from the normativity of any rule-governed situation: some possible responses are right and some are wrong. Indeed, to say that I bound myself by yesterday’s promise is precisely to say that I committed myself in advance to evaluating any rejection of your request for the $100 today as “wrong.” If I didn’t bind my future evaluations in this way, then I didn’t make a promise. When a young child says, “I promise...” the behaviour is not a promise precisely because it cannot bind tomorrow’s situation. The child will wake up tomorrow unburdened by obligations from yesterday’s events. When the child grows up and learns how to be a self, however, it will accept to be bound by values and obligations that stem from its actions yesterday. In other words, it will have taken on the temporal structure in which it accepts to be the same self from day to day. It is this commitment structure that makes a being an enduring self.

ii. Computer analogy: Promising as re-programming

My second analogy for understanding the internal preconditions needed for a self is based, not on biological sensation, but on computer programs. An algorithmic computer is programmed to accept a certain kind of input, to process that input in a prearranged way, and then to produce an appropriate output. Similarly, as a psychological structure, the brain has a set of perceptual capacities that allow it to recognize and interpret the world in which it lives. It then processes what it has perceived and, on the basis of the values that it has previously adopted, it performs an appropriate action. Some organisms have only genetic programming, but human organisms can also be reprogrammed by teaching. An individual capable of being a self, however, has in addition to these, the ability – a learned ability – to reprogram itself by making commitments. The promise an individual made yesterday preprogrammed its input systems in such a way that it perceives and interprets today's situation as one in which it is obliged to fulfill the promise. (Of course other factors might lead it to break the promise, but in so far as it has developed an honest character, the previous promise-making will weight heavily in determining its action or output.) Each promise is, as it were, a mini-reprogramming that controls, when the situation arises, the processing of the new input and the appropriate output response. If an individual is the kind of organism that has this capacity for such reprogramming, then it has what it needs to be attributed the status of a self.

This computer analogy enables us to put my main question a slightly different way: under what conditions would a computer be able to partake in the social institution of promising, and so become a self? 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. If it says, “I promise...”, then, if this is considered a commitment at all, it is a commitment that binds the programmer. If however the computer becomes capable of programming itself, in other words, if it becomes its own programmer, then we might well consider it a worthy participant in our social games. The chess strategies used by Deep Blue to beat Kasparov were not programmed into the computer by the programmers. Rather Deep Blue learned from experience with earlier matches to reprogram itself to follow its own novel tactics, tactics that the programmers understood no better than Kasparov. Responsibility for the actual moves can hardly be laid at the feet of the IBM programmers. The main factor here seems to be progressive learning within the history of the individual. If, on the basis of its past adaptive interactions with the world, a computer reconfigures its own program so as to approach the world differently in the future, then the computer itself could be held responsible for these future actions.

So, to complete my second analogy, in so far as an individual's brain reprograms itself is this way, then the individual can be attributed responsibility for the situation it finds itself in, namely, facing a world in which some action is perceived as obligatory. This is why the simple memory of making a commitment is inadequate as a basis for the self's unity over time: it is not the past event as a recorded content of the mind that is needed here, but the event in so far as it has reprogrammed the individual. An individual that has a temporal structure that allows it to make commitments that bind it in the future can be attributed the status of a "self."

4. THE RELATIONSHIP BETWEEN INSTITUTION AND INDIVIDUAL

So far, I have examined the nature of promising as a social institution and have offered two analogies – biological sense organs, and computer reprogramming – to help us understand the kind of internal structure that an individual human organism would have to have to be able to participate in this institution. Let me now bring these two parts of the analysis together.

It is the ability of the individual organism to restructure itself or reprogram itself that is the internal precondition which allows the individual to participate in the external, social institution of promising and thereby to have the status of selfhood bestowed on it. Let me first remind you that I have been taking promising as a stand-in for, or example of, any commitment whatsoever. All social relations, legal contracts, marriages, as well as activities such as taking a job, writing a check, joining a political party, raising a family, and so on, all have the structure of making a commitment at one time and accepting subsequent obligations at a later time. These are normative, rule-governed institutions which an organism can participate in only if it is capable of performing present acts which bind it in future situations. My claim is that to be a self is to have a role in these institutions and that playing such a role requires a set of mental or psychological mechanisms.

The mechanisms are not themselves the self. Remember our dollar coin. That it is made of metal is an internal property of the disc, but that it is money is externally defined by the economic context of exchange. When someone points out that a puff of air or a soap bubble could not be a dollar because the material substrate must be made of something enduring, they are not thereby claiming that it is the hardness of the metal which makes it a piece of money. My claim is that it is not the internal capacities of the individual that make it a self; these capacities are only internal preconditions which enable social institutions to attribute the status of selfhood to the individual. One final analogy: Sylvia Cartwright is Governor General of New Zealand and while a minimum of maturity, intelligence, and emotional stability on her part may be preconditions for her being able to fulfill the role, it is not these properties that make her a governor general. Being Governor General is an externally attributed status. It is her role in a political institution that includes the Queen, Prime Minister, Parliament and so on, that defines her status. My position is that being a self is similarly an externally attributed status and cannot be reduced to the necessary internal preconditions in the organism. Since commitment is a normative, rule-governed activity, my account of the self is non-reductive. A self is a contextually defined entity.

4. CONCLUSION

By self I mean the structure that unifies a life over time. I have claimed that the empiricist account of selfhood looks for patterns of unity that are internal to the individual and that are based on contingent relations of contiguity or causality between pre-existing, independently definable mental events. In opposition, I have used the notion of commitment – illustrated by the promise paradigm – to propose a contextual or externalist account of the self as a normative being whose temporal structure is unified by the social rules of commitment and responsibility.

David L. Thompson
Memorial University of Newfoundland
Visitor at Victoria University of Wellington
2005

Bibliography

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

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

Reid, T. Essays on the Intellectual Powers of Man—A Critical Edition. Edited by Derek R. Brookes. Edinburgh, UK: Edinburgh University Press. 2002 (Original work published in 1785.)