RMM Vol. 0, Perspectives in Moral Science,
ed. by M. Baurmann & B. Lahno, 2009, 157–175
http://www.rmm-journal.de/

Geoffrey Brennan and Alan Hamlin

Bygones Are Bygones

Abstract:

'Bygones are bygones' might seem to be an analytic truth, lacking any substantive content. Yet, economists think that, when they state that bygones are bygones, they are asserting something interesting and important. Furthermore, others would argue that the statement 'bygones are bygones', when read appropriately, is false. By interrogating the statement 'bygones are bygones' we identify a number of key issues relating to rational choice theory and the treatment of intentions, habits and promises. The more philosophical discussion of the things that economists say (and what they might mean) is particularly appropriate in honoring Hartmut Kliemt, much of whose work has brought philosophy and economics into closer proximity.

1. The Sayings of Economists

Economists think that what they say is important. As it happens, this is not a compliment that they extend to the subjects of their models: with certain notable exceptions, 'speech' is not an aspect of human behavior with which economists have been much concerned. Philosophy differs from Economics in this respect. Philosophers think that what people say—and even what they think—is important. And important in itself and not just for the (other) 'actions' with which it might be associated.[1]

Hartmut Kliemt has an academic identity that spans economics and philosophy. Although formally a Professor of Philosophy, he has collaborated extensively with economists—and even his writings in more mainstream philosophy have a distinctively 'economistic' cast. As the economists might say, he thinks, and even talks, like an economist. But being a philosopher, he is appropriately self-reflective; and from time to time therefore, he must reflect on the sort of things that economists say, and what exactly to make of them.

In fact, there are certain things that economists say, *qua* economists, that are distinctive and notable. The remark—'bygones are bygones'—is one of them: standing alongside 'there ain't no such thing as a free lunch' (a warning that

[1] Cohen (2003, 242f.) puts the point succinctly in relation to political philosophy: "[...] suppose that, like me, you think that political philosophy is a branch of philosophy, whose output is consequential for practice but not limited in significance to its consequences for practice. Then you may like me protest that the question for political philosophy is not what we should do but what we should think, even when what we should think makes no practical difference."

everything has a cost); or 'you can't beat something with nothing' (a standard defense against any critique of economic theory); or 'you can't have your cake and eat it!' (a folk version of the necessity of choice); or 'there's more than one way to skin a cat' (a reference to the ubiquity of substitution possibilities). To take such aphorisms and interrogate them strikes us as a potentially interesting and instructive exercise, revealing the lineaments of economists' prejudices and the strengths and limits of the 'economic way of thinking'. And here, we shall undertake just one small piece of that task.

Taken on its face (as an uninitiated philosopher might take it), the 'bygones are bygones' maxim might well seem uninformative to the point of fatuity. The claim that 'A is A' is hardly rocket science! And to be confronted with an earnest economist who looks you in the eye and asserts this claim, with a certain air of intellectual intensity, might make the puzzled bystander wonder what modern education is coming to. What else, the observer might ask, could bygones *be* if not bygones? Yet, economists think that, when they state that bygones are bygones, they are asserting something interesting and important. They think that many people don't realize that this is true—and that they stand in need of correction—or at least fail to see the full import of the remark.

At the same time, some renegade critics think that the 'bygones are bygones' claim is clearly false. And to be sure, if it turned out that A were *not* A, then that would be something that would be ... well, at least 'philosophically interesting'.[2]

It is perhaps obvious that what economists mean by bygones are bygones involves assigning a different meaning to the first 'bygones' from that assigned to the second. So, we probably should refer to the claim as the B_1B_2 claim—or, in extensive form, 'bygones 1 are bygones 2'. And no less obviously, the task of analysis is to specify exactly what 'bygones 1' and 'bygones 2' really are, and why the former is (or is not) an instance of the latter, and why things might seem otherwise in at least some cases. This might be a tedious and slightly old-fashioned form of linguistic analysis; but it is necessary if we are to be clear as to what economists mean—or what they *might* mean on various readings.

2. Story Telling

A natural way of getting a handle on this interpretative exercise is to begin with how the aphorism is deployed in the class-room. And in this context, B_1B_2 is often associated with a story (parable?) about lost opera/theatre/football tickets. The story goes this way.

You have purchased a ticket to the opera (or theatre, or football) for $100; but when the relevant time comes, you cannot find the ticket. You can replace it easily (by telephone call) but that will cost you (another) $100. Since you previously thought that the ticket was worth one hundred dollars, and nothing in

[2] We recall the paper *I Am Here Now* by G. Vision (1985). A title that has that same puzzling air—finding something interesting in a claim that on its face is an *a priori* truth.

relation to your taste for opera (or theatre or football) has changed, you should, if you are rational (say the economists), just shell out the additional money and buy a replacement ticket. In particular, the fact that the ticket you purchased earlier has been lost should not affect your decision. Bygones are bygones.

There is a complication that ought to be recognized and set aside immediately. The loss of the ticket reduces your income (or wealth): loosing the ticket is as if you had lost $100 in cash—and the fact that you are $100 poorer will make you reduce your expenditure across the board to some extent. Perhaps you will decide that you can no longer afford to go to the opera/theatre/football. To neutralize this complication, suppose not only that you lost the ticket, but also that you found $100 in cash on the sidewalk on the way home that evening. In this event, since income, wealth and tastes are all as before, the economist would argue that you should replace the lost opera ticket: anything else would be irrational!

The underlying idea here is that rationality is distinctively forward-looking: the rational act is understood as that act which will produce the best consequences (from the point of view of the agent)—and those consequences necessarily lie in the present and the future. So, rational action is forward-looking, expedient. Human behavior is, on the rational actor view, drawn forward by the future—not pushed from behind by the past.

So, on this reading, B_1 refers to the collection of events—previous ticket purchase, ticket loss—that occurred in the past. B_2 signifies irrelevance for decision making purposes.[3] So 'bygones are bygones' might be rendered as 'you shouldn't cry over spilt milk'—or perhaps, 'no point in shutting the stable door after the horse has bolted'.

But note three things about B_1B_2, interpreted as 'the past should be treated as irrelevant'. First, it is a normative claim. Second, it is a normative claim that specifically attaches to rationality. And third, it is, on its face, manifestly false. Some remarks about each of these aspects in turn.

Economists routinely insist on a clear separation of the positive and the normative. They think that this separation invokes a good methodological principle. They also think that the positive is the arena where their primary authority and expertise lies. They think they know something about how the world works—and in principle are content to leave discussion of how it ought to work to others (though actually knowing how the world works gives them, they think, some insight into how it might most plausibly be made to work differently, and specifically, to work better).

But principle and practice can diverge. If the economist is accepted as authoritative on 'positive' questions, there is an obvious temptation for him to try to slip normative claims into the 'positive' category so as to boost the authority of those claims. B_1B_2 masquerades as an *a priori* methodological principle for

[3] The claim that B_1 is irrelevant for decision making purposes does not imply that it is irrelevant for all purposes—for example, the loss of the ticket may still be relevant to some overall evaluation of well-being, in exactly the same way that the loss of $100 would be.

social analysis; but it is actually a folk version of a normative principle—and one moreover that is by no means beyond question.

Economists believe that agents are rational—or at least that their actions are best understood in terms of rationality in the sense that agents will generally act as if they were rational. That is a core element of the mainstream paradigm. They think of rationality, first and foremost, as a positive (i.e. descriptive) attribute of human action. They don't think of rational actor theory as a normative theory telling people how they *ought* to behave—or even as a hypothetical normative theory, telling people how they would have to behave if they wanted to be 'fully rational'. On the other hand, they like to astound their students by revealing to them the surprising things that being fully rational entails. But this is rather puzzling. The force of the story about the tickets is that you'd be doing something pretty silly if you decided just to stay at home when you couldn't find your tickets. And the implication is that this silly thing is something that lots of ordinary folk do. The economics student is supposed to have learned something that other folks don't know—and to feel superior to others who haven't had the benefits of an education in what rationality requires. But if that is so, how can rationality be assumed to be a standard feature of ordinary agents?

3. Is B_1B_2 True?

The substantive question, though, is not so much what kind of claim: 'the past should be treated as irrelevant' is. It is rather: is the claim 'true'—or more accurately, is it something that rationality requires?

We think not. Indeed, we think 'obviously not!'

And this should give us pause; because although it is great fun to make fun of economists[4], they are not stupid—and there may be much that a closer interrogation of the claim has to teach us. As George Stigler allegedly used to say: "if you think you have found an error in Adam Smith, look again! It's much more likely that you are wrong!" (Not that, as far as we know, Adam Smith ever asserted that bygones were bygones.)

So consider the obvious refutation.

a) Rational choice theory conceives action as the result of an optimizing process in which the agent's preferences confront the set of feasible options. Changes in the feasible set are indeed the primary factor in explaining (changes in) behaviour. But;

b) actions undertaken now will often alter the feasible set in subsequent periods. So, the building of a large, stone wall in period 1 will require you to walk around it in period 2; the commitment to a particular piece of capital plant in period 1 will lock you into a technology in period 2; saving

[4] We are economists of a kind ourselves. Like certain kinds of ethnic jokes, that are only allowed to be told by members of that ethnic group—making fun of economists is an exercise profitably left to economists.

some portion of your income in period 1 will enable you to increase your consumption in period 2; and so on.

c) Indeed, the feasible set is mainly constructed via a combination of social and physical factors all of which occurred in the past and all of which cast their shadow into the present.

d) Given this obvious fact, how *could* the past be treated as irrelevant?

But this, the economist might reply, proves our point. The past isn't irrelevant insofar as it contributes to the constitution of feasible sets—but that is precisely *because* bygones are bygones! The past is given and fixed; constraints, wherever they came from and whatever the events that shaped them occurred, are *constraints*! [Mmm. Another A = A proposition?] Choice can only be exercised over things that could be other than what they are; and the past cannot be other than it is. Bygones are bygones!

This response is instructive. If we were now to spell out the B_1B_2 claim more fully we might have something like: 'narratives of the past are only relevant to the extent that they result in constraints that bind on present action.' Rationality then requires that any two situations that are identical in terms of the constraints faced must be identical in terms of the choice of action that would be made by a rational individual (that is, a single individual with particular, fixed preferences) regardless of any differences in the 'stories' that might be told about the processes that gave rise to the situation. And this expanded account fits well with the story of the opera/theatre/football tickets, where the situation of the first ticket purchase is held to be identical in all relevant respects to the situation in which the individual faces the decision of whether or not to replace the ticket (so that the details of the 'story' are held to be irrelevant). And since the situations are identical in all relevant respects, the decisions should be the same.

So, on this reading, the economist is not making the sweeping (and obviously false) statement that the past is always irrelevant to decision making, but the more moderate statement that details of the past that cannot be tracked to specific differences in the feasible set faced are irrelevant to decision making.

However, even this reading is not without its difficulties. In general, we would require a non-question-begging way of identifying what counts as a 'specific difference' in a feasible set. Otherwise we are simply left with the statement that 'the irrelevant is irrelevant' which is not much progress over the original 'bygones are bygones'. This is a point that we will return to below.

But our more detailed formulation of B_1B_2 also suggests that it is, or should be, connected to the 'demand-side' and not just the 'supply-side' of rational choice. By this we mean that it is the agent and her beliefs and preferences that are the target for the warning that 'bygones are bygones' rather than the specification of the feasible set.

There are at least three aspects of decision-making that might engage with the B_1B_2 claim 'on the demand side' as we put it, and each of these is of some

interest in the account of rationality. These aspects are: intentions; habits; and promises. We will consider each in turn.

3.1 Intentions

In the stripped down Hume-Davidson version of rational choice theory[5] that economists tend to embrace, the central categories are desire, belief and action.[6] Rationality is a connection between action and desires, subject to beliefs including beliefs about which actions best promote desire satisfaction. Intentions, as such, play no role in this stripped down version—although it has become fashionable in some philosophical circles to give them much prominence[7]. For our purposes here a relatively deflationary account of intention will be sufficient.

Being rational takes time, information and effort. You have to consult all your relevant beliefs and desires—and in at least some cases (perhaps most) you will make fewer mistakes in trying to choose the rational action if you make certain calculations to identify that action. In many cases, those calculations will take time and energy; and so the issue of how you optimally allocate your decision making time and energy will itself be an issue on which rationality requirements bear. It will often not be very satisfactory to leave those calculations to the last minute: the exercise of calculation may well get in the way of your acting most effectively. Besides, there will be periods when your mind is free, when it would cost you almost nothing to contemplate what it would be best for you to do in a situation you know is coming up (or is highly likely to come up). Indeed, such contemplation may be the most pleasurable thing for you to do at that time. So, it will be rational to do your calculation *then*—and thereby form an 'intention' at that point as to what it would be best for you to do in the future choice situation. This conception of an intention involves nothing more than your capacities to anticipate and to remember, plus the idea that rationality requires you to allocate calculation effort inter-temporally in an optimal way. It is an example of the forward-looking nature of rationality since it involves thinking ahead, anticipating future decision making situations and investing now in some calculations of relevance to those situations.

So you do your calculation in advance. You form an intention. And, in the absence of any new information when the time comes, you act in accordance with the intention made earlier. Clearly, it will not be 'rational' for you to act without reference to intentions earlier formed: for then there would be no point in forming them. Sometimes you might decide, on the basis of a whim, or a

[5] For an elegant summary of that notion (and of some of the problems to which it gives rise) see Elster 1986. For a discussion by one of the present authors see Brennan 2007.

[6] Actually, the economists' version usually refers to 'preferences' (some rather under-specified amalgam of beliefs and desires, with the latter suitably aggregated) and to specifications of the structure of those preferences. Action is 'rational' if it involves maximal preference satisfaction and if the preferences exhibit completeness, transitivity, etc.. The Humean-Davidsonian underpinning is, however, what theorists of rationality in the economistic mode tend to invoke when required to give an account of what their notion of rationality really amounts to.

[7] Originally Anscombe 1975, but also Davidson 1979 and more recently Broome and Pillar 2001.

desire to be spontaneous, to act in a manner contrary to your intention. But, on the account we have offered so far, that would not be enough to undermine the rationale for intention formation.[8] Nevertheless, most of the time, it will be better for you to follow your intentions; and on our account the rational person will do so most of the time.

But any intention to act, with 'intention' here understood as a psychological state held as a memory of a prior rational calculation, seems by definition to be a 'bygone' (B_1). It was formed in a period prior to action and makes no 'specific difference' to the feasible set. But if you treated it systematically as irrelevant to action, you would appear to be behaving irrationally. You would be undoing the benefits of the prior optimal calculation-time allocation; and your life will go less well for you as a result. Intentions, it seems, are 'bygones' (B_1); but they are not 'bygones' (B_2).[9]

An objection to this line of argument is that an intention, as we have constructed it, is not really a bygone (B_1) but is simply a timeless calculative fact. Just as when we perform some arithmetic calculation we recall previous calculations and apply their results provided that the circumstances are appropriate; so when we have a choice we may recall previous similar calculations (in the form of 'intentions') as a sort of short-cut. All that we have to do is satisfy ourselves that circumstances are similar and that we have no obvious reason to depart from the earlier 'intention'. It is not the fact that the intention is previously constructed that points to its relevance for current decision making, but the fact that we currently accept that the situation is such that a known calculation applies. Just as when we apply the arithmetic fact that $2+2=4$ in helping to perform some current calculation, it is not the fact that $2+2=4$ derives from the past that is relevant, but the fact that we currently hold it to be both accurate and relevant. On this line of argument, it will be rational to form 'intentions' of this sort, in exactly the way that it is rational to practice arithmetic operations—both will serve us well by providing a ready means of performing future calculations.

Consider the simple case in which I aim to cook dinner this evening, and, in order to achieve this aim, will have to go shopping to buy ingredients. My forward thinking rationality tells me that it is sensible to first identify the menu, and construct a shopping list that identifies the ingredients I need to purchase. This is sensible because it allows me to check the list of ingredients required for any menu by referring to recipes that will not be available to me when I am in the market, and also to check the list of required ingredients against those that I already have available at home. So, my shopping list is a record of a form

[8] So it is not a stipulation of rationality, on this view, that you must always act in accord with intention. Some scholars endow intention with a quasi-commitment status—something perhaps like a promise to oneself to act in the way intended. That is not a line we follow here; and it should be clear that we find it somewhat overblown.

[9] Of course, it might be suggested that we have merely assumed that the formation of intentions is itself rational in order to challenge the idea that the B_1B_2 claim is required by rationality. It might equally be suggested that the B_1B_2 claim is entailed by rationality in order to show that forming intentions is irrational. Either way, we would conclude that rational intentions of the type we describe and the B_1B_2 claim cannot coexist within an account of rationality.

of intention—I go to the market intending to buy that list of items. But once I arrive at the shop, I might spot, by chance, an item that I had not previously considered and revise my menu and my shopping plan. How might we analyse this sequence of events? We offer the following account: at the initial stage, the formulation of a shopping list and the intentions associated with it, are motivated by a desire to ensure that the feasible set of options to be confronted at later stages is non-empty, and that this set contains at least one acceptable option. The shopping list/intention provides assurance that we will be able to produce a meal, and that we will be able to do so without undue cost incurred at either the shopping or the cooking stage. In this way, the intention may be seen as aimed at making a 'specific difference' to the feasible set. In this case, the intention is not best understood as pre-empting or committing choice—there is no sense in which I should feel obliged to buy the items on my list if a better alternative presents itself. In this way, it is both rational to form the intention (so as to provide assurance of feasibility) and it may also be rational to depart from the intention without feeling any loss. The intention provides a sort of default option or back-stop, but does not commit future decision making at all.

This understanding of an intention seems to us to fit reasonably well with the more detailed formulation of B_1B_2. The shopping list idea of an intention may have value as a means of conveying information through time, and in providing assurance of feasibility and so is not a bygone (B_1). But still it is neutral with respect to commitment.

But this discussion also points to a potential second dimension of an intention—the idea that an intention may carry independent weight as a reason for acting in the manner intended. Put loosely an intention is a sort of promise to yourself (or yourself at a different date) and you have reason to honour such 'promises'. If an intention is just a remembered calculative outcome or shopping list that carries no independent weight but is simply a neutral input into current decision making then it seems that such an intention may not be a bygone (B_1); whereas if we understand an intention as carrying independent weight, predisposing rather than simply informing your eventual decision, then the B_1B_2 claim is more seriously challenged.

This second sense of 'intention' is perhaps captured in the idea of a 'resolution'—where I resolve to act in a certain way, with the idea that such an intention/resolution will itself act to change my behaviour relative to the counterfactual that lacks the resolution but is otherwise identical. Here we are in territory that includes questions of akrasia, endogenous and higher-order preferences and the like, and this is territory that is not well explored by economists, so that whether their attachment to 'bygones are bygones' extends (or is intended to extend) into this realm is unclear. But if we were to argue along similar lines to those employed in the case of 'specific differences' in the feasible set, we might suggest that, *if* it is possible for an agent to 'intend' in this second sense—so as to successfully commit, constrain or alter her preferences in the future—then such an intention would constitute a relevant 'specific difference' on the demand side. In this case then, leaving aside the difficult question of the mechanism

by which such an intention/resolution might operate, we might further restrict our understanding of the 'bygones are bygones' claim so that it might now read something like: details of the past that cannot be tracked to specific differences in either the feasible set faced or the structure of preferences are irrelevant to decision making. This still narrower interpretation of the phrase points to the idea that since rational decision-making analyses any instance of a decision into just two component elements—the feasible set and the preferences—any impact from the past must work through its impact on one (or both) of these.

Our simple discussion of intentions has emphasized two aspects or dimensions of the underlying idea: one which views an intention as a pre-calculation that is informative but does not commit or constrain (although it may provide assurance of feasibility), the other which views intention as resolution, at least attempting to commit or constrain future action. We have suggested that these two interpretations carry very different implication for the B_1B_2 claim and, to the extent that our actual understanding of the idea of an intention includes both of aspects discussed, that fact renders the relationship between intentions and the B_1B_2 claim somewhat complex. There is a sense in which the truth of the claim can be maintained—but only by restricting its domain quite significantly.

3.2 Habits

Humans are creatures of habit. We form habits both unintentionally as a matter of course from repeated behaviour; and intentionally or consciously, as when we 'learn' to play a piece of music on the piano or train ourselves for the triathlon. Generally, in the psychology literature, it is assumed that repetition of various tasks or actions creates 'tracks' in the neural pathways of the brain which in turn allow much more rapid and unconscious behaviour patterns. In some cases, the habits in question appear to be matters of biochemical dependence, as is the case with popularly described 'addictive substances'. But whatever the precise mechanism, it does seem that certain actions increase the desire or capacity to undertake similar actions on future occasions. And in that sense, any explanation of current behaviour will involve some reference to behaviour in the past.

Again there are 'fact' and 'norm' aspects to this phenomenon. To the extent that habit acquisition is a fact, habits acquired from past action effectively change the feasibility set—not necessarily in the sense that they rule certain actions out as 'infeasible', but in the slightly weaker sense that they change the relative 'prices' of different actions in the present, when the 'prices' in question are defined to include the cost of overriding habits. In the face of the 'fact' aspect of habits, to say that 'bygones' (past actions) are bygones (irrelevant) is just to deny the facts about habits and the structure of our brains: it seems a hopeless claim!

But, of course, habits can be broken: old habits can be replaced by new ones, or eroded. Is there any presumptive normative force in the notion that we should

seek to minimize the role that habit plays in our conduct, or eschew any activities that threaten to become habit-forming? We think the answer is clearly not—or at least, not if we want our life to go as well as possible. There are good habits and bad habits. And in general one might wish to promote and entrench good habits while eroding and replacing bad habits. Many of the highest accomplishments we recognize—artistic, athletic and intellectual—depend on the cultivation of relevant habits both on the 'supply' and the 'demand' side. Skills are cultivated as a matter of muscular habit. And the disposition to work long hours in the development of those skills—to become, for example, a semi-compulsive scholar[10]—is also a matter of habit (not to say addiction!). To be sure, choice at each point in time will be influenced by whatever habits we have (good and bad), and we should recognize this fact and attempt to keep an eye on the habits and inclinations we develop and rationally assess their overall consequences for likely future choices so as to engage in whatever habit forming or re-forming activities might seem to be appropriate.

In Adam Smith's account of the division of labour, and the increasing returns that arise from specialization, habituation is seen to play an important role in shaping increases in productivity across the whole range of human activities. And of course, as every good economist knows, it is the division of labour and the specialization it allows that constitute the source of the "*greatest improvements in [...] productive powers*" and consequent increases in "*the wealth of nations*".

Explanation in economics, on one influential reading,[11] actually requires agent preferences to be stable: on this view, to 'explain' some change in the state of the world in terms of changes in individuals' tastes is not to explain anything. Proper explanation involves changes in feasibility sets and/or relative prices (or incentives), *with preferences fixed.* Of course, the objects over which the fixed preferences are defined may be quite abstract,[12] so as to allow for rationally explicable trade-offs between alternative mechanisms ('tastes') by which those more abstract preferences might be best satisfied. But the fixity in those preferences (at whatever level) is an important general presumption of the 'rational actor method' in practice and it is difficult to see how such fixity can be assumed without some appeal to the forces of habit.

So it just seems a mistake to think it a bad thing in general that people should develop habits and a good thing for them to bend their energies to resisting habit formation at every turn. On this view, bygones are not bygones and it would be a bad thing if people treated them as such.

The possession of habits is an important part of making human behaviour predictable not just to the economist—but also to the participants in economic and social life. Of course, life in the absence of habits would not be a matter of random behaviour—but still habits enhance predictability. Here we must return to the distinction between types of habit: the unintended, perhaps even

[10] Any association between this description and the honouree of this collection is, of course, entirely coincidental.

[11] Becker's specifically, as for example laid out in Becker 1993, or Stigler and Becker 1977.

[12] For example see Stigler and Becker 1977.

unconscious habit; and the consciously intended developed habit. These two types of habit may pose different challenges to rational choice theory and hence to the B_1B_2 claim.

Intended habits open up consideration of second-order rationality; that is, the idea that rationality is not merely concerned with choosing the appropriate act on each occasion, but also with choosing the relevant mode of choice which will be employed to select particular actions—what we have previously termed modal dispositions.[13] If I decide, as a matter of second-order consideration, that the best way for me to make decisions whenever I am offered the choice between tea and coffee is to forgo any detailed calculation of the relative desirability of the two beverages and simply choose coffee in the mornings and evenings, but tea in the afternoons, then while my tea and coffee drinking behaviour may seem to be habit—based, it is also at a slightly deeper level rational.

By contrast, unintended habits seems to stand in clear opposition to the basic idea of rational choice as an alternative explanation of behaviour. Put crudely, if a wide range of behaviour is habitual, and habits are not themselves intended or chosen, there seems little work for rational choice theory to do. There are two broad responses to this line of thought: first it might be that even if habits are not intended or chosen, they might still arise in a manner that is consistent with the rational choice style of explanation at some level; second we might question the claims regarding the extent and importance of unintended habits for at least some types of behaviour. These two responses are mutually compatible but only the first seems to raise interesting theoretical questions, so we will focus our attention on this line of argument.

Unintended habits emerge as regularities in behaviour, so that the question must be what conditions the emergence and continuation of such habits? The most promising area in which to look for an answer might seem to be models of social evolution—an area in which Hartmut Kliemt has made significant contributions. At the most basic level, an evolutionary model consists of two primary moving parts—a process that generates variation, and a selection mechanism. A selection mechanism might normally be considered in terms of some notion of 'fitness'—that is, the behaviour to be selected (and established as a habit) will be that which, of those behaviours available, maximizes 'fitness', however 'fitness' might be defined in a particular case.

Now, it is easy to see that this type of approach to evolutionary modeling might produce explanations that are consistent with rational choice theory. If an evolutionary engine produces as its outcome behaviour that maximizes fitness, this will engine will be consistent with a rational choice model that operates *as if* behaviour is directly chosen to maximize fitness. So, assuming that a similar notion of fitness can be motivated in both the evolutionary and rational choice frameworks, and assuming that we can interpret the idea of maximization similarly in the two contexts, we might suggest that the two types of models will overlap in their range of explanation.

[13] See Brennan and Hamlin 2000; Hamlin 2006. Note the link to intentions given the discussion above.

The idea here then is that unintended habits might be assimilated to the case of intended habits by appeal to the idea of the evolution of habits. Of course, this relaxes the idea of rational choice away from the literal idea that habits are rationally chosen, towards the idea that habits that evolve will tend to be those that would have been chosen by an appropriately rational individual; but this shift to hypothetical rather than actual choice remains comfortably within the broader rational choice tradition.

To the extent that intended habits can be accommodated within rational choice theory, and unintended habits can be assimilated within a rather more expansive notion of rational choice theory, the problems posed by habits for the B_1B_2 claim are essentially similar to those posed by the idea of intentions in the sense of resolutions or commitments. Habits become an essentially rational means of linking past and current (as well as future) preferences together and, as such, can be interpreted similarly to preferences. If the enriched reading of the B_1B_2 claim is taken to mean that the only ways in which the past can impact on present decision making are via impacts on the feasible set or on preferences, then our discussion of habits suggests that they provide another way in which preferences might be impacted. As before, this allows us to maintain the reading of the B_1B_2 claim but only by further reinterpretation of the underlying rational choice model. We have to give up the strict idea that rational choice explanation takes place only on the side of the feasible set; and accept that there are interesting, if revisionist, aspects of preferences that are themselves subject to scrutiny within the more broadly rational choice tradition.

There is one further possibility that should be mentioned here. So far we have limited concern to the case of individual choice—although we have allowed extension of that individual over time to consider intentions and habits. The multi-person or social dimension raises further concerns. For example, it could conceivably be the case that it might be good for *me* if everyone else was a habit-former but I was not: habits may be 'public' rather than 'private' goods and subject to the incentive to free-ride. And, if this were the case, there would be a new tension introduced—while it might be rational for me to develop habits when I consider the purely inter-temporal aspect, it would not be rational for me to form habits when I consider their purely inter-personal aspect (although it might still be rational to encourage others to form habits). We mention this possibility but offer no resolution here—though we will return to more strategic, inter-personal considerations below. On the general matter of habits it seems plausible that, within appropriate limits, being a predictable person is the price of entry into social life, and the resolute habit-resister is unlikely to be an attractive trading-partner/friend/spouse. We can probably trust evolution to quietly remove such types from the landscape, or at least limit their influence.

3.3 Promises

A further arena in which bygones might be thought to be bygones—and one which typically operates inter-personally—relates to the force of promises. And indeed this is a matter of considerable significance both descriptively for the operation of markets and similar institutions, and normatively. Whether or not the world is one in which promises are normally kept, without recourse to the additional resources of an enforcement mechanism such as contract law, makes a significant difference to how social and economic interactions proceed—and to the kinds of institutional arrangements that are justifiable.

The basic idea here is that when you come to consider the act that would be required to fulfill a promise, the promise itself is a 'bygone'. All that confronts you at the point of action are the benefits and costs of that action *as perceived from now onwards*. So you will act *as if* no promise had been made (or as if the promise had no force): this, so the story goes, is what is rationally required. In particular, if it is beneficial to you at that point for you to exploit the promisee, you will exploit her, whether you promised otherwise or not. Of course, this is problematic because, if this is so, then no-one will trust you and you will never be in a position to exploit anyone.[14]

Consequently, it will be desirable for you to be able to make credible promises—ones that you will rationally fulfill—because then you will earn the fruits of mutually beneficial arrangements otherwise denied you. But this fact is not enough in itself to make your promises binding. As Bentham tellingly remarked, "*the* demand *for rights are no more rights than hunger is bread!*" (with 'credible promises' standing in for 'rights' here). The rational desire to be able to make binding promises does not make all promises rationally binding: to assume so is to assume that Ulysses can be bound to the mast whether there is a mast at hand or not!

It is clear that any promise to act in the future in a particular way will, when the time to act appears, be a 'bygone' in the sense that it occurred in the past. But this does not make it irrational to abide by that promise—at least, not unless rationality is defined in unremittingly 'objective pay-off' terms. For anyone familiar with Hartmut Kliemt's work[15] (or for that matter our own[16]), the 'trust/reliance' predicament will be something of a cliché. But it will be useful to go over this old ground here just to emphasize one or two points that might not be obvious. So in Figure 1, we show the basic reliance predicament, in which player A must choose in period 1 whether to rely, R, or not (NR). If she chooses not to rely the interaction ends (or perhaps, fails to start). If R is chosen, player B must choose in period 2 whether to exploit (E) that reliance or not (NE). In this interaction, it is rational for player B to choose E, because

[14] The idea that purely verbal promises—or other statements that are not associated with costly commitment devices—are merely 'cheap talk' and will have no impact is discussed in Farrell and Rabin 1996.

[15] Güth and Kliemt 1994; 2000.

[16] Brennan and Hamlin 2000, ch's 3, 5.

E offers B a larger payoff than NE. Player A knows this; and, since NR offers player A a higher payoff than does E, A will choose not to rely.

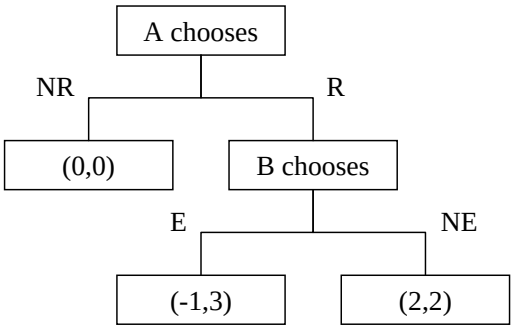

Figure 1

Now suppose that B promises A that B will play NE in period 2. If 2 could make this undertaking binding, then both 1 and 2 would be better off than in NR. But is the [R, NE] outcome genuinely accessible? Whether it is or not *does not depend* on the mere fact that the promise actually occurred in the past, once B comes to act. It depends instead on whether B is a sufficiently 'trustworthy type'. A trustworthy type here is someone who considers it 'intrinsically' wrong to break promises. And we can depict this conveniently, by postulating that a trustworthy type would endure a subjective loss of amount g (for guilt) if she were to break her promise. In the example given, if this subjective cost g exceeds a value of 1, then player B will fulfill her promise because the payoff to B under NE now exceeds that under E. Player 1 will know this and so [R, NE] will be the equilibrium outcome.

Now, it is true that g is an entirely subjective payoff and not necessarily fully observable by A. And player B can certainly do better by *pretending* to be trustworthy, provided that B can convince A of his trustworthiness. So whether it would be 'rational' for B to become trustworthy will depend on whether his trustworthiness can be recognized in at least enough cases to make his acquiring the character trait worthwhile. Suppose those conditions are met—for B, at least. Then the extended game illustrated involves a 'pre-move' in which B either promises (P) or does not (NP). If she does not, the game is as before, with equilibrium NR. If she does promise, then the game is the left branch of the extended game in Figure 2, with equilibrium [R:NE].

Note that given the existence of g, player B is behaving entirely *rationally*: it is just that objective payoffs do not reflect the full structure of the interaction. Player B is only apparently 'irrational' if the crucial subjective element in overall payoff is ignored. So, as far as rationality goes, it will be: rational for B to acquire the trait of trustworthiness; rational for B to make a promise; rational for A to trust B; and rational for B to choose NE over E.

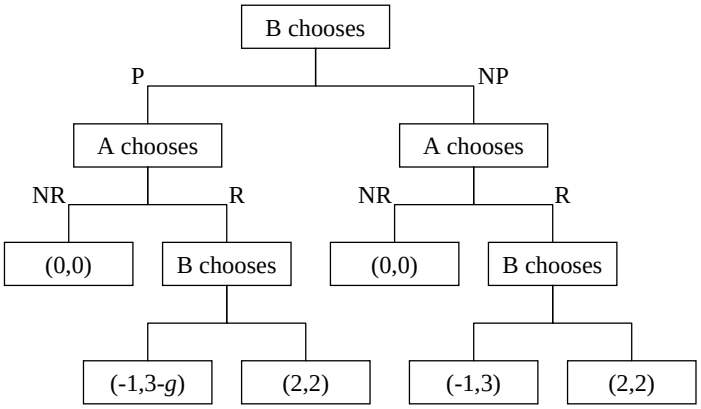

Figure 2

Two aspects of this account are worth noting. The first is that it makes the rationality of fulfilling promises contingent on the presence of the *g*-factor that B will suffer if he does not keep his promise. Some commentators think that this gets the ontology of morality wrong: B keeps her promise (and more generally behaves morally) because it is the right thing to do, not because she will feel terrible if she breaks her word. It is morality that causes 'right-doing'—not guilt. In one sense, that claim seems to us to be entirely correct. If B came to think that in a particular instance, it would not be wrong to break his promise then he would indeed break it: the *g* factor would simply disappear. It does not however follow from this fact that the guilt plays no direct motivating role: indeed, it would be odd that we should have evolved to feel guilt if it played no role in disciplining our actions. In any event, our *g*-factor does not need to be thought of exclusively as guilt: it could be some combination of a natural distaste for 'doing the wrong thing' and of 'guilt' as more commonly understood.[17]

The other aspect of the interaction worth noting involves implications for the normative. Economists are used to drawing a distinction between motivation and justification in relation to invisible hand mechanisms involving narrow self-interest. Here some of that same distinction arises in relation to normativity. It is tempting to think that the primary justification for trustworthiness lies in its capacity to increase the objective pay-offs available to players—morality *solves* the predicament. But of course, it can only do this by introducing extra normative considerations—normative considerations that take on a life of their own. Trustworthy persons consider that goodness in the world is not reducible to

[17] For example when one remembers doing something wrong, one feels guilty by virtue of that recollection—one does not *remember* feeling guilty (or if one does, that is something else). The portion of guilt relevant for g is the present discounted value of all the future negative feelings that doing the wrong thing will generate. There can remain a strictly instantaneous preference for 'doing the right thing'!

objective payoffs: there is something more. And there *needs* to be that something more if objective payoffs are to be maximized.

The more important point for our purposes here, however, lies not so much in normative interpretation but in the relevance of bygones. Whether B has made a promise or not is a critical piece of the account—at least for that subset of the population who have already become committed to promise-keeping. And there is nothing intrinsically irrational about keeping promises earlier made. All one can say is that the mere desire to maximize pay-offs is insufficient as a direct route to achieving the maximizing of pay-offs. The trust predicament is a real one and the passage of time plays an important role in it; but it does not show that bygones are bygones.

Finally, a word about the relation between the Ulysses example and the promising case. When Ulysses hears the Sirens, he desires to leap into the water—to come as close to the source of the sound as is possible: this desire is an all-things-considered desire but the related action is thwarted by his being bound to the mast. The promise-keeping case seems different in its basic description: here, the actor has been so constituted that he does *not* desire, all things considered, to break his promise. To be sure, there is a *pro tanto* desire to exploit; but that *pro tanto* desire is defeated by another contrary desire—driven by directly 'moral' considerations. This descriptive difference may not be behaviourally relevant in the case in hand. But it is counterfactually relevant—because in the promising case behaviour is responsive to relative prices in a way that Ulysses is not. Ulysses remains bound to the mast whatever additional temptations arise. But the promise keeper can fall victim to temptation: if the cost of promise-keeping (or the gain from promise-breaking) were to increase sufficiently, his behaviour will change. It is this piece of counterfactual reality that leads us to model trustworthiness (and other moral requirements) as a demand-side phenomenon, rather than as a constraint.[18] In this sense, some 'bygones' (mast-tying) may be more bygone than others (promises)!

4. Equilibrium in Co-ordination Games

One (but only one) of the things at stake in the bygones issue appears to be the capacity of rational players to form expectations of others' behaviour in coordination settings. Consider the simplest 'left-right' coordination rule in determining which side of the road to drive on. Traditionally, it has been thought—in a symmetric game of the kind illustrated in Table 1—that rational players can readily settle on a rule or norm of driving based on the actions that others have taken

[18] Whether morality is understood as a constraint or a competitive desire seems to be part of what is at stake in Sen's distinction, first aired in his famous *Rational Fools* paper (1977), between "commitment" (constraint) and "sympathy" (special desire). It is not entirely clear for what purpose Sen seeks to deploy this distinction—and the idea of modeling moral considerations as constraints sits uncomfortably with his treatment of rights—but something like the demand-side/supply-side distinction does seem to be at stake.

in the past. So, if in an iterated version of this interaction, players have played 'left' for the last ten iterations, there seems to be some rational presumption that they will continue to play 'left'. But part of the thrust of the B_1B_2 notion, the notion that rationality is ruthlessly forward-looking, carries the implication that learning of this sort is not 'rational' so that the norm of, say, [left, left] will not emerge as an equilibrium of the iterated game. In a world in which bygones are truly bygones each iteration of the game occurs as if it were the first. So neither player is able to induce from any history of past plays that the purely rational other will act as he has done in the past. So there will be no rational reason for either player to assume that the other will continue to play 'left' and so no rational reason to play left himself.

	2's choice	
1's choice	left	right
left	(10,10)	(0,0)
right	(0,0)	(10,10)

Table 1

It would, of course, be possible to modify the definition of rationality in an apparently modest way to accommodate this problem. We could for example insert into the definition of rationality a clause in the spirit of Newton's laws of motion, in the following way:

> *A is rational if A's behaviour remains unchanged unless A's beliefs or the relative prices of things that A desires (incentives) change.*

At least at first cut, this would seem to secure the possibility of the emergence of stable norms in coordination predicaments, like that in Table 1. But specifications of this kind seem both too weak and too strong—too weak because even the smallest change in beliefs might be sufficient to alter behaviour in a way that might put at risk the stability of coordination equilibria; and too strong because it would rule out mixed strategies (in which behaviour changes according to optimally determined random factors), to say nothing of preference changes as such. Becker (1993) may be right that changes in tastes are not so much an 'explanation' in economics as a confession that we *have* no explanation—but it seems excessive to declare all (unexplained) changes of tastes irrational![19] Moreover, at least in cases like Table 1, player 1 only has to believe that there is a slightly higher probability that 2 will choose what he chose last time for it to be rational for 1 to continue to play the strategy that matched on the previous plays (and likewise for 2): it is not necessary that the rational player choose the equilibrium strategy with certainty.

[19] Becker more than most economists has of course been an exponent of the view that many changes in tastes can be accounted for within a broad economic model (i.e. by virtue of changes relative prices at an appropriately abstract level). See for example Becker 1996, *Accounting for Tastes*.

However, one reading of the claim that bygones are bygones is itself an extreme claim: it asserts that the past is irrelevant *in its entirety*. And that seems to us to deny inter-temporal connectedness of a kind that is totally familiar to ordinary persons and ought to be accessible to rational ones!

5. The Bottom Line

If all this is so, then the question seems to be: what is right, if anything, about the B_1B_2 claim? We think merely this: that causal chains are time-dependent. It is not possible for an action undertaken now to *cause* an event in the past.[20] The consequences of an act undertaken now all lie in the future or the present—and rational action is directed exclusively at those consequences.[21] This fact is not by any means sufficient to make the past irrelevant—but it is a constraint on the kinds of considerations about the past that might be entertained. However, since no-one ever thought that actions undertaken now could directly cause the past, it is not entirely clear precisely what this 'constraint' delivers. What it does *not* deliver are any strong claims (normative or positive) about habits, intentions or promises—either in the making or the keeping. In each of these cases, we have argued that rational choice theory—suitably interpreted and applied can make good sense of the relevant inter-temporal and inter-personal connections in terms of relevant changes in either the set of feasible alternatives or the set of relevant preferences (interpreted to include appropriate commitments and dispositions). Unreasonably narrow conceptions of rational choice theory give rise to difficulties in understanding the force of the B_1B_2 claim, but these difficulties may be resolved by taking what we regard as more reasonable positions with respect to the defining characteristics of rational choice theory.

On balance, we are inclined to think that 'bygones are bygones' is a remark that economists would be well-advised to give up. It seems entirely clear that it confuses non-economists! And it is at least arguable that it confuses the economists themselves.

[20] Of course this is not to deny that the anticipation of an act at time t_1 can be causally relevant to an act at the earlier time t_0—but simply to distinguish between the anticipation and the act itself.

[21] Even this claim needs to be interpreted carefully. I may be able to bring about the 'consequence' that a promise I made in the past was indeed a promise, if I currently fulfill it. If (certain) 'actions' are denumerated in terms that extend over time, then arguably an agent can *now* bring about the truth of a claim about past actions.

References

Anscombe, G. E. M. (1957), "Intention", *Proceedings of the Aristotelian Society* 57, 321–332.

Becker, G. (1993), *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, Chicago: University of Chicago Press.

— (1996), *Accounting for Tastes*, Boston: Harvard University Press.

Brennan, G. (2007), "The Grammar of Rationality", in: F. Peter and B. Schmid (eds.), *Rationality and Commitment*, Oxford: Oxford University Press, 105–123.

— and A. Hamlin (2000), *Democratic Devices and Desires*, Cambridge: Cambridge University Press.

— and — (2008), "Revisionist Public Choice Theory", *New Political Economy* 13(1), 77–88.

Broome, J. and C. Piller (2001), "Normative Practical Reasoning", *Proceedings of the Aristotelian Society*, Supplementary Volumes 75, 175–216.

Cohen, G. A. (2003), "Facts and Principles", *Philosophy & Public Affairs* 31(3), 211–245.

Davidson, D. (1970), "Mental events", in: L. Foster and J. Swanson (eds.), *Experience and Theory*, Amherst: University of Massachusetts Press, 79–101.

Elster, J. (1986), *Rational Choice*, New York: New York University Press.

Farrell, J. and M. Rabin (1996), "Cheap Talk", *The Journal of Economic Perspectives* 10, 103–118.

Güth, W. and H. Kliemt (1994), "Competition or Co-operation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes", *Metroeconomica* 45(2), 155–187.

— and — (2000), "Evolutionarily Stable Co-operative Commitments", *Theory and Decision* 49, 197–221.

Hamlin, A. (2006), "Dispositional Politics and Dispositional Politics", in G. Eusepi and A. Hamlin (eds.), *Beyond Conventional Economics: The Limits of Rational Behaviour in Political Decision Making*, Cheltenham: Edward Elgar, 3–16.

Sen, A. (1977), "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory", *Philosophy and Public Affairs* 6(4), 317–344.

Stigler, G. and G. Becker (1977), "De gustibus non est disputandum", *The American Economic Review* 67(2), 76–90.

Vision, G. (1985), "I Am Here Now", *Analysis* 45, 198–199.