

# The complex evolution of a simple traffic convention: the functions and implications of habit

Geoffrey M. Hodgson<sup>a,\*</sup>, Thorbjørn Knudsen<sup>b</sup>

<sup>a</sup> *The Business School, University of Hertfordshire, Hatfield, Hertfordshire AL10 9AB, UK*

<sup>b</sup> *Department of Marketing, University of Southern Denmark, Odense Campus, 5230 Odense M, Denmark*

Received 3 August 2001; received in revised form 7 February 2003; accepted 11 April 2003

---

## Abstract

This paper explores the evolution of a simple traffic convention concerning the side of the road on which to drive. This agent-based simulation probes some of the deeper conceptual issues involved in the evolution of conventions, particularly the nature of rational decision-making and its possible reliance upon habit. The simulations show that the systemic convergence to a left/right convention is often improved or sustained by habit, alongside other “intelligent” and calculative attributes of agents. We show that habit formation is part of a possible mechanism of “reconstitutive downward causation” among agents where the preferences of each agent are partly malleable.

© 2004 Elsevier B.V. All rights reserved.

*JEL classification:* A10; B40; B52; C60; D83; Z13

*Keywords:* Agent-based model; Evolution; Conventions; Institutions; Habits; Rationality; Downward causation; Endogenous preferences

---

## 1. Introduction

The evolution of conventions and institutions has become the subject of much analysis, modeling and discussion.<sup>1</sup> We raise here some further analytical and conceptual issues on the basis of a heuristic, agent-based simulation with heterogeneous agents. The general *outcome* of the simulation is relatively uncomplicated because we choose one of the most straightforward of decisions and conventions: whether to drive on the right or on the left of the road.<sup>2</sup> In our model, artificially intelligent “drivers” in “cars” are programmed to

---

\* Corresponding author. Present address: Malting House, 1 Burton End, West Wickham, Cambridgeshire CB1 6SD, UK.

*E-mail address:* g.m.hodgson@herts.ac.uk (G.M. Hodgson).

<sup>1</sup> See, for example, Marimon et al. (1990), Wärneryd (1990b), Young (1993), and Howitt and Clower (2000).

<sup>2</sup> Young (1996) provides an interesting historical account of the evolution of traffic conventions.

negotiate a circular road configuration along with a number of other, similar vehicles. We show that the emergence of a convention is possible but by no means guaranteed. Furthermore, some manipulation of the decision processes through which these “drivers” decide to move to the left or the right provides a basis to consider some of the deeper conceptual issues that are involved in the evolution of conventions, such as the nature of rational decision-making and its possible reliance upon habit.

Each driver is boundedly rational. To negotiate the track and avoid collision, it would seem to be rational for each driver at least to consider *conformity* with the perceived distribution of traffic to the left and right and *avoidance* of cars that are immediately ahead. To these factors, our model adds *habit*.

Any left/right convergence outcome in this model is likely to depend on initial conditions and circumstances. Strong path dependence is likely, but we are more interested in the degree and resilience of any emergent convention than whether it is on the left or the right.

We show that in following or avoiding other traffic in some circumstances, strength of habit and processes of habituation can play a vital role alongside rational deliberation and selection pressure. This outcome not only raises important questions concerning the role of habit in decision-making, but also it challenges the frequent assumption that preference functions should always be taken entirely as exogenously given.

This paper is structured in eight parts. As well as defining some key terms, the second part considers the theoretical background and points to some important differences of view concerning the manner in which institutions and conventions evolve. The heuristic model is presented in the third part. In the fourth part the results of the simulations are reported. The fifth part considers a different model with “inertia” and shows that it does not aid convergence as strongly as habit. The implications of the simulations concerning the concept of habit and regarding the concept of “downward causation” are discussed in the sixth and seventh parts, respectively. The eighth part concludes the essay.

## 2. The evolution of conventions and institutions

We follow widespread practice and define institutions as durable systems of established and embedded social rules that structure social interactions. Language, money, law, systems of weights and measures, traffic conventions, table manners, firms (and other organizations) are all institutions. A convention is a particular instance of an institutional rule (Sugden, 1986; Searle, 1995). For example, all countries have traffic rules, but it is a matter of (arbitrary) convention whether the rule is to drive on the right or on the left.

In a book first published in German in 1871, Menger (1981) pioneered the basic analysis of how institutions evolve. His chosen example was money. Menger saw the institution of money as emanating in an undesigned manner from the communications and interactions of individual agents. Traders look for a convenient and frequently exchanged commodity to use in their exchanges with others. Once such regularities become prominent, a circular process of institutional self-reinforcement takes place. Emerging to overcome the difficulties of barter, money is chosen because it is convenient, and it is convenient because it is chosen.<sup>3</sup>

<sup>3</sup> Despite the apparent simplicity of this monetary argument, analyses, experiments and simulations based upon it are extraordinarily complex (Jones, 1976; Kiyotaki and Wright, 1989; Oh, 1989; Wärneryd, 1989, 1990a;

In this Mengerian approach, individual preference functions are taken as given. Menger thus inspired a central, unifying project in the “new institutional economics”: to explain the existence of political, legal, or social institutions by reference to a model of given, individual behavior, tracing out its consequences in terms of human interactions.<sup>4</sup>

However, theoretical analyses or simulation of the evolution of institutions have proved to be remarkably problematic. For example, in the work of Marimon et al. (1990) an attempt is made to model the emergence of money with artificially intelligent agents. Their results are qualified and partially inconclusive. A single monetary unit does not always readily emerge. Menger’s discursive analysis of an emergent convention has proven to be remarkably difficult to replicate in a computer simulation. Our simulations also show the difficulties in reaching convergence even with a very simple convention.

The central hypothesis behind the present paper is that there is often more to the emergence of real world institutions than mere matters of convenience and calculation by individual agents. Additional psychological factors intervene. A verbal exposition of this basic idea can be found in the writings of the neglected tradition of “old” institutionalism. For instance, arguing that the evolution of money cannot be understood simply in terms of cost reduction and individual convenience, Wesley Mitchell maintained that money “stamps its pattern upon wayward human nature, makes us all react in standard ways to the standard stimuli it offers, and affects our very ideals of what is good, beautiful and true” (Mitchell, 1937, p. 371). Accordingly, the evolution of money changed the mentality, preferences and thinking patterns of individuals themselves. This does not necessarily mean that Menger’s account is wrong, but that it is sometimes inadequate. At least in some circumstances, it may have to be supplemented by an analysis of how institutions can change individual perceptions and preferences.

The idea of the malleability of individual preferences pervades the “old” institutional economics, from Thorstein Veblen to John Kenneth Galbraith. However, it has not yet been shown why some preference malleability may be necessary for the emergence and sustainability of institutions. In this article we begin to fill this gap by showing how a limited form of preference malleability can improve the possibility and stability of an equilibrium convention.

What is at issue here is the adequacy of the standard account of the emergence of institutions. Just as individuals constitute institutions, individuals may also be partially reconstituted by institutions. Once we raise this possibility, however, we encounter some conceptual problems concerning the specification of such preference endogeneity. It is not our intention to replicate a widely criticized picture of individuals as puppets of institutions, roles or cultural values. To avoid such pitfalls, we have to specify adequately the limits, nature and mechanisms of this reconstitution.

It is here that we come to the concept of habit. The simulation outlined in the next section shows how habit can be significant for institutional evolution, especially in circumstances

---

Hodgson, 1993; Marimon et al., 1990; Duffy and Ochs, 1999). Realizing this, we chose a simpler institution as the object of the present study in which each agent has a choice between only two behavioral options at any stage. Our intention was to illustrate the hypothesized results in the simplest possible institutional set up.

<sup>4</sup> For discussions of the limits of this approach see Field (1979, 1984), Knight (1992), Sened (1997), Hodgson (1998) and Aoki (2001).

of limited information. Circumstances help to form the predispositions of individuals by forming and changing their habits. Of course, several attempts have been made to accommodate a notion of habit within relatively sophisticated rational choice models.<sup>5</sup> In these models, any habit is seen as ultimately an outcome of a rational choice. In contrast, in the pragmatist tradition of Charles Sanders Peirce, William James, George Herbert Mead and John Dewey, any rational deliberation is always seen as grounded on habit. The question then is whether rational choice is the foundation of habit, or whether the reverse is true (Becker, 1992; Hodgson, 1998, 2003, *in press*). The discussion of the simulation, in section six, addresses this dilemma. It is shown that the concept of habit developed in the cited rational choice models is not the same as the concept in our model and in the tradition of pragmatist thought.

Our intention is not to treat habit as some kind of psychological panacea, but to investigate its significance in the “experimental” context of a simulation. The model shows that in some circumstances habit can assist convergence to a left/right convention but it also depends upon, and interacts with, other variables and processes. We do not argue that habit is the only factor involved in convergence, but under frequent conditions it is important when allied with other factors. We also find that in some circumstances habit can be disruptive.

### 3. The simulation model

#### 3.1. *The decision problem and the environment of choice*

In our model,<sup>6</sup> 40 agents drive around a  $100 \times 2$  grid, arranged in a ring, with two lanes and 100 zones. We use the terms “agent,” “driver,” and “car” synonymously. The drivers are unique individuals, born to drive either clockwise or counter-clockwise around the ring, referred to as lengthways movement. Half of the agents drive clockwise and the other half counter-clockwise. No car changes its direction of movement.

At time  $t = 0$ , the drivers are randomly assigned a zone and a position on one of the two sides of the ring. The cars then move in turn. During each move, each driver must decide whether to drive on the left or the right side of the ring when making their next lengthways movement. This left/right movement is the driver’s only choice variable. Each driver performs an incremental lengthways movement, placing itself in the next zone ahead, on either its left or its right lane.

The left/right decision is partly based on information about the traffic in front of the driver. The driver looks 10 increments ahead and counts in that region the number of cars in each lane and the number of cars going in each direction. Based on this information and given its behavioral and cognitive dispositions (defined below), the driver will decide on which side of the ring to drive in its next move. Each car drives around the ring until it is

<sup>5</sup> See, for example, Pollak (1970), Winston (1980), Blanciforti and Green (1983), Philips and Spinnewyn (1984), Becker and Murphy (1988), Alessie and Kapteyn (1991), Becker (1992).

<sup>6</sup> The simulations described here were performed using Matlab software. All random numbers are generated from a multiseed generator with the theoretical lower limit of  $2^{1492}$  before the number will repeat itself.

involved in a collision. A collision occurs when a car moves into a zone occupied by another car that is also in the same lane, irrespective of the direction of movement of the cars. Then both drivers die and new cars and drivers replace them. As a result, the number of cars on the grid is always 40. The replacement routine also ensures that the number of cars moving clockwise and counter-clockwise is always 20.

### 3.2. Behavioral and cognitive dispositions

Initially our objective was to make the drivers as “intelligent” as possible, subject to the constraint of a limited number of cognitive and behavioral variables. After numerous runs with additional cognitive parameters, we found that a highly parsimonious model was very effective.<sup>7</sup> Additional decision parameters had little effect in enhancing the survival of individual cars or the convergence characteristics of the model.<sup>8</sup>

When first placed on the ring, each driver receives a unique set  $\{SSensitivity_n, OSensitivity_n, Avoidance_n, Habitgene_n, Habituation_{n,t}\}$  of five cognitive and behavioral dispositions. The first four of these dispositions are randomly assigned and cannot be changed. These variables are randomly chosen according to normal distribution with mean 1 and standard deviation  $\delta$  (referred to as the mutation variable). Negative numbers are truncated to zero, but there is no upper bound.<sup>9</sup> The only disposition that can be changed during the life of the car is the car’s acquired habits ( $Habituation_{n,t}$ ). Furthermore, for all original or newborn drivers, the initial level of the habituation variable is zero ( $Habituation_{n,0} = 0$ ).

Note that the terms “left” and “right” are relative to the driver involved. A car driving clockwise on the right will not collide with a car driving counter-clockwise on the right. The same applies to two cars both on the left, likewise moving in opposite directions. The terms “ahead” and “behind” are also relative to the car and its movement. A car may collide with another car moving in the same direction, but only if that other car is one zone ahead and does not move first, or if that other car is one zone behind and does move first.

- (i) *Same-direction sensitivity*: Each driver looks forward and observes the number of cars going in the same direction as itself up to and including 10 zones ahead, and calculates the proportion of this number driving on the left (or right) hand side of the road. (If no car is going in the same direction as itself, up to and including 10 zones ahead,

<sup>7</sup> Given the relative simplicity of the decision environment, and the effectiveness of our “parsimonious” decision algorithm, it seemed neither necessary, appropriate nor fruitful in this model to incorporate more complex learning procedures such as the “elaboration likelihood model” of Petty and Cacioppo (1986) and the non-linear models of attitude change by Eiser et al. (2001). However, more complex learning algorithms would clearly be appropriate in decision environments involving more learning parameters and behavioral choices than are present in our model.

<sup>8</sup> Earlier versions of this paper included three “inertia” parameters and an additional “avoidance” variable applied to the area *two* zones ahead of the driver. The inertia parameters gave each driver a disposition to continue stubbornly with an inclination it has assumed in the recent past. The discussion of inertia in section five below shows that its effects are generally weaker than those of habit. The effect of the second “avoidance” variable was at best marginal and often insignificant. Accordingly, the more parsimonious model was chosen, with the omission of these parameters.

<sup>9</sup> The probability that a negative number will be drawn is extremely small ( $7.43 \times 10^{-6}$ ). An alternative method of selecting the first four parameters would be to draw them randomly from a uniform distribution in a specified interval. Instead, a normal distribution was selected because it was found that it reduced the death rates in the simulation. Selection along an interval will typically create a larger number of drivers with less fit parameters.

then the proportion is taken as 0.5.) The variable  $SSensitivity_n$  indicates the degree to which driver  $n$  takes account of this ratio in determining its next move. If this variable is high then the car will tend to conform to the pattern of behavior of the cars ahead of itself and moving in its own direction.

- (ii) *Opposite-direction sensitivity*: Each driver  $n$  looks forward and observes the number of cars going in the opposite direction to itself, up to and including 10 zones ahead, and calculates the proportion of this number driving on *their* left (or right) hand side of the road. (Again, if no car is going in the opposite direction as itself, up to and including 10 zones ahead, then the proportion is taken as 0.5.) The coefficient  $OSensitivity_n$  indicates the degree to which car  $n$  takes account of this ratio in determining its next move. As well as a rationale to conform to the convention established by others, there is an incentive to avoid this traffic coming in the opposite direction.
- (iii) *Avoidance*: This coefficient captures a tendency for each driver  $n$  to avoid collision with close, oncoming traffic. Each driver looks forward and observes the number of cars going in both directions, *one* zone ahead, and calculates the number on the left and right-hand side of the road, relative to the driver. Because each car moves in turn, another car that is positioned one zone ahead of driver  $n$  poses a collision danger, regardless of its direction of movement: cars moving in both directions threaten driver  $n$  with immediate collision. Driver  $n$ 's avoidance is captured by the coefficient  $Avoidance_n$ , referring to the situation one zone ahead.
- (iv) *Habit gene*: A driver's habit gene must be distinguished from its habituation. The habit gene is the instinctive tendency that a driver has to take account of its acquired habituation. The habit gene cannot change but habituation can. The role of the habit gene is explained in the discussion of habituation below. Driver  $n$ 's habit gene is captured by the coefficient  $Habitgene_n$ .

Every driver receives a unique personal profile in which the values of the above four behavioral and cognitive variables are randomly assigned. However, the following variable can change through the course of a driver's life.

- (v) *Habituation*: Every driver starts with a habituation variable set initially at zero. As time goes on, this variable will be revised according to the car's movements. For instance, if a car has a history of moving on the left-hand side of the road then the habituation variable is likely to be positive, and if a car has generally moved on the right-hand side of the road then the habituation variable is likely to be negative. A more precise account of the habituation process is given below. The habit gene coefficient expresses the degree to which driver  $n$  takes its habituation into account. Driver  $n$ 's habituation at time  $t$  is captured by the coefficient  $Habituation_{n,t}$ .

### 3.3. Calculation, habituation, decision and movement

Each car is addressed and moves sequentially. With no simultaneous moves, some associated problems of interpretation of the intentions of others are thus avoided. In each period, all drivers in turn make a (subjective) decision based on the (objective) information about the traffic ahead. Again the purpose was to make the drivers as "intelligent" as possible, making use of the most important information for their survival, subject to reasonable computational constraints. As each car can only move one zone ahead, there is no reason to take

account of traffic to its rear. As noted above, at time  $t$ , each driver calculates the following variables:

$S_{L,n}$ : The proportion of all cars, going in the *same* direction as driver  $n$ , up to and including 10 zones ahead, that are driving on the left-hand side of the road, where  $0 \leq S_{L,n} \leq 1$ . If no car is going in the same direction as driver  $n$ , up to and including 10 zones ahead, then  $S_{L,n} = 0.5$ .

$O_{L,n}$ : The proportion of all cars, going in the direction *opposite* to driver  $n$  and up to and including 10 zones ahead, that are driving on *their* left-hand side of the road, where  $0 \leq O_{L,n} \leq 1$ . If no car is going in the opposite direction as driver  $n$ , up to and including 10 zones ahead, then  $O_{L,n} = 0.5$ .

$C_{L,n}$ : The number of very close cars, going in any direction, exactly one zone ahead of driver  $n$ , that are driving on the left-hand side of the road relative to driver  $n$ , where  $C_{L,n} = 0$  or 1.

$C_{R,n}$ : The number of very close cars, going in any direction, exactly one zone ahead of driver  $n$ , that are driving on the right-hand side of the road relative to driver  $n$ , where  $C_{R,n} = 0$  or 1.

After having gathered this information and calculated the above ratios, the driver then updates its habit function according to the following formula:

$$\text{Habituation}_{n,t} = \text{Habituation}_{n,t-1} + \text{LR}_{n,t} / (K + \text{Moves}_{n,t}),$$

where  $\text{LR}_{n,t}$  is the situation of car  $n$  at time  $t$ , whether it is on the left ( $\text{LR}_{n,t} = 1$ ) or on the right ( $\text{LR}_{n,t} = -1$ ) hand side of the road.  $K$  is an arbitrary positive constant and  $\text{Moves}_{n,t}$  is the total number of moves the driver has undertaken, up to and including the present move at time  $t$ . In addition,  $\text{Habituation}_{n,t}$  is bounded between  $-1$  and 1. Clearly, the tendency to change habit decreases with the number of moves; the habit function is cumulative with a decreasing increment. The driver uses the above equation to update its habituation variable.<sup>10</sup>

To make a decision to go left or right, the value of the following expression is calculated:

$$\begin{aligned} \text{LREvaluation}_n &= w_{\text{Sdirection}} \times \text{SSensitivity}_n \times (2S_{L,n,t} - 1) \\ &+ w_{\text{Odirection}} \times \text{OSensitivity}_n \times (2O_{L,n,t-1}) \\ &+ w_{\text{Avoidance}} \times \text{Avoidance}_n \times (C_{R,n,t} - C_{L,n,t}) \\ &+ w_{\text{Habit}} \times \text{Habitgene}_n \times \text{Habituation}_{n,t}. \end{aligned}$$

The  $w_X$  coefficients ( $w_{\text{Sdirection}}$ ,  $w_{\text{Odirection}}$ ,  $w_{\text{Avoidance}}$ , and  $w_{\text{Habit}}$ ) are fixed, non-negative weights common to all 40 drivers. The weights determine how much the components of

<sup>10</sup> Experiments were performed using different habit functions, with similar but slightly weaker results. Perhaps the main rival alternative habit function would be similarly cumulative, but with geometrically decreasing increments, as in the classic work of Hull (1943). However, in the present context, a Hull function has the disadvantage that the sum of a suitable geometric series of increments (with a geometric coefficient between zero and unity) is always finite. As a result, the indefinite reversibility of an acquired habit from one extreme to the other and back again would not be possible. In contrast, the chosen increments in the habit function in the present work decline at a rate that permits in principle the indefinite reversal of habituation from one extreme value to the other: no sign nor degree of habituation is ever irreversible.

every driver's unique set of cognitive and behavioral dispositions will influence the driver's subjective evaluation and thus its choice to go left or right at time  $t$ . The coefficient  $w_{\text{Habit}}$  is referred to as the "habit weighting." The term  $w_{\text{Habit}} \times \text{Habitgene}_n \times \text{Habituation}_{n,t}$  is referred to as "the strength of habit" of a car.

Note that each term on the right-hand side of the equation above includes two positive elements plus one element with expected values equally distributed around zero, all multiplied together. Hence each term on the right-hand side has expected values equally distributed around zero. As a result, there is no bias to the right or the left in the model.

The subjective evaluation of each car is given by the variable  $LREvaluation_n$ . If  $LREvaluation_n$  is greater than zero then the car intends to move to the left. Otherwise it intends to move to the right. The final element to be taken into consideration is the possibility of error. An error probability variable  $\varepsilon$  is pre-set at the beginning of the simulation. A random number generator is used to determine whether each car, with probability  $\varepsilon$ , makes the move opposite to its subjective evaluation. At this final stage, the left or right inclination of the car in the upcoming move is determined.

The car then moves one increment forward onto the next zone, on the left or right as determined. If there is no other car on the same side of the road and in the same zone, then there is no collision. In each period, all drivers in turn go through these steps.

The drivers in the model are boundedly rational. Taking account of the most important local information, each car responds and maneuvers in order to survive. The decision algorithm combines decision elements that vary according to the cognitive personality of the driver and the global parameter weights. The population of varied decision algorithms itself evolves due to selection pressure, leading to surviving decision algorithms of some fitness value.

### 3.4. Replacement of colliding drivers

If there is neither birth nor death, then the pool of fixed characteristics among the population cannot change. At least, a small amount of death and replacement is necessary to select the combinations of fixed cognitive and behavioral dispositions that are conducive to survival. However, this means that a replacement routine is necessary for new cars and its form can influence the outcomes in the model. It should be emphasized, however, that the overwhelming majority of deaths generally occur in the early, transition phase of the simulations.

If there is a collision, then—regardless of blame or circumstances—the two drivers die and are replaced by two new cars and drivers. The weights  $w_X$  are common to all agents and also used by the newborn drivers. However, the two newborn drivers require a new set of four fixed cognitive and behavioral dispositions  $\{SSensitivity_n, OSensitivity_n, Avoidance_n, Habitgene_n\}$ . These were chosen randomly in the same manner as the cars in the population at the beginning of the simulation, with the habituation level  $\{\text{Habituation}_{n,t}\}$  always set initially to zero.

Their cognitive and behavioral characteristics being determined, each new car is allocated to a random position on the track. However, to reduce the frequency of immediate collisions, no new car is allocated to a zone occupied by another car.

### 3.5. Design adjustments and parametric searches

Experiments were performed with different values of the mutation variable  $\delta$ . Although a convention emerged with many runs with a higher value, a relatively low value of 0.2 was chosen in order to achieve a lower and more plausible degree of mutation. Different values of  $K$  in the function above for  $\text{Habituation}_{n,t}$  were also tried. Clearly, as  $K$  decreases to zero, the left/right choice by the car in its first move will increasingly dominate its strength of habit. The outcomes were relatively insensitive to variations in this coefficient, but habit had a slightly improved positive effect on convergence with values of  $K$  in the region of 10. This value ensured that habituation adjusted at a significant but modest rate.

The decision horizon is the number of zones ahead that a driver scans to estimate the traffic pattern. This data affects the driver's 'conformist' calculations concerning same-direction and opposite-direction sensitivity. As summarized in [Appendix B](#) below, a number of simulations were performed with various decision horizons greater than 10 zones ahead, including the possibility that drivers see all 100 zones of the entire ring. It was found that habit significantly improves convergence for values of the horizon from zero up to and including 25 zones. The maximum habit effect appears with a horizon of 10, which is the value used in the standard runs reported in the main text below.

After all the design adjustments were complete, the values of the first three non-negative weights  $\{w_{\text{Sdirection}}, w_{\text{Odirection}}, \text{ and } w_{\text{Avoidance}}\}$  were considered by searching through their multidimensional parametric space, with progressively decreasing increments of search, with  $w_{\text{Habit}}$  always set at zero. The three positive weights were always normalized according to the rule that their average was unity. The convergence performance, death rates and other aspects of the model were monitored during these searches. At each search point, a sample of at least 100 repeated simulations were made to obtain mean values. Also at each point, error was increased uniformly from zero to 0.02, across the set of 100 or more samples. This search of parameter space identified the point of maximum convergence  $\{w_{\text{Sdirection}} = 1.4, w_{\text{Odirection}} = 0.9, \text{ and } w_{\text{Avoidance}} = 0.7\}$  with  $w_{\text{Habit}} = 0$ .<sup>11</sup>

## 4. Simulation results for the standard model

### 4.1. Preliminary remarks

The principal aim of the simulations is to gauge the degree of left/right convergence in multiple runs of the model, exploring different points of parameter space and assessing the impact of different levels of habit and error.

Generally, when an equilibrium outcome emerges, the resulting convention, whether drive-to-the-right or drive-to-the-left, can be highly sensitive to initial conditions. Once the system begins to swing decisively and permanently one way or the other and a convention begins to emerge, then it can become locked into a process that is the cumulative result of tiny initial movements (Arthur, 1994).

<sup>11</sup> Searches in parameter space confirmed that this was a global rather than a local maximum. However, the region of convergence optimization is almost flat, making accuracy to more than one decimal place superfluous.

However, two factors can disturb this process of convergence to a left/right convention. The first, and more ubiquitous, is error. The effects of error can be particularly disruptive in the early phases of this process. However, even in later phases, errors can trigger deaths that lead to replacements that are ill-adapted for the road conditions, leading to further collisions, and so on. It is possible for such processes of positive feedback to destroy an established convention.

The second disturbing factor emerges under specific conditions only. It is prevalent in a relatively small neighborhood of parameter space. In some circumstances agile drivers can evolve, typically with a low but positive level of the habit gene. These drivers are sufficiently agile to avoid the traffic ahead, by moving repeatedly from one side of the road to the other if required. A “cycling” pattern can occur, when cohorts of agile drivers repeatedly move safely and laterally to avoid other oncoming groups. There may be a degree of local convergence in each group, but the conventions in different groups may be different. If there are no further collisions then replacement and mutation through death cannot occur. Consequently, a unanimous convention will not emerge among the population as a whole.

4.2. Some illustrative simulations

Illustrative results from two different runs are displayed in Fig. 1. The vertical scale measures the average inclination of the cars to the left or right. The horizontal scale measures the number of car moves enacted through time. In Figs. 1 and 2, a value of unity on the vertical scale would correspond to the unanimous use of either the right or the left-hand side of the road by all cars. A value of 0.5 on the vertical scale would indicate that the cars were equally distributed on the right or the left. The expected value at the start of the run is 0.5. With a run of 20,000 car moves, the mean of the 20,000 vertical values is computed.

In order to compare results whether the drivers happened to converge on the left or the right, we used the following standardization procedure. If this average is less than 0.5, then it is subtracted from 1, ensuring that the overall convergence outcome (C) is always greater

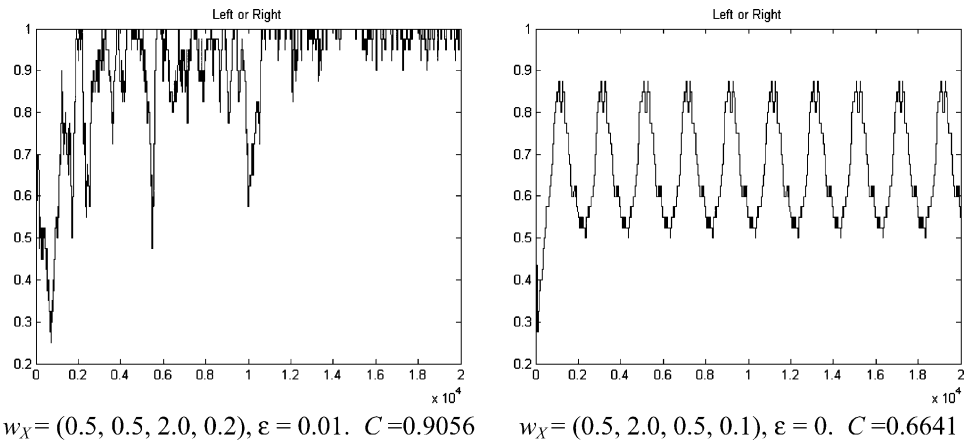


Fig. 1. Two illustrative runs.

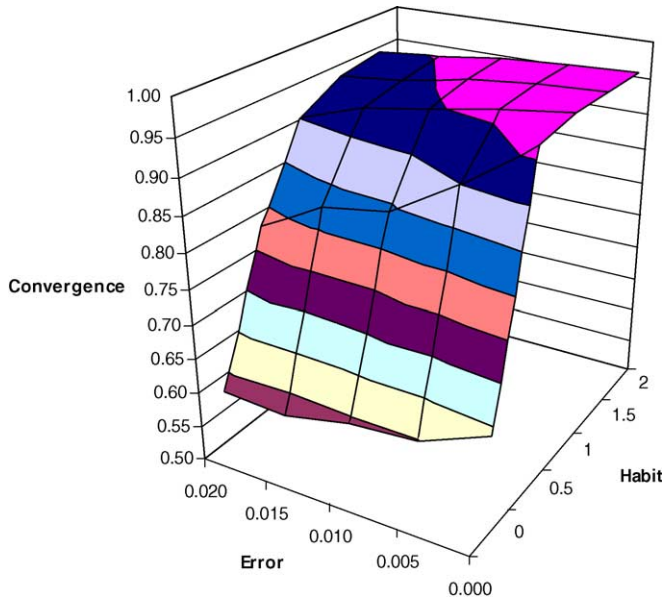


Fig. 2. Degrees of convergence with 200 runs for each level of habit and error  $w_X = (1.4, 0.9, 0.7, w_{Habit})$ .

than or equal to 0.5 and less than or equal to 1. With this measure of convergence success, higher values indicate a greater degree of convergence. A figure of 0.95 would indicate that on average, for one entire run, 95 percent of the cars were on one particular side of the road.

Hence the degree of convergence ( $C$ ) for  $m$  moves with  $c$  cars is the total number of moves in which a car is on the left/right, divided by  $c \times m$ . In this calculation, all cars are considered for each individual move of every car through the entire run. The choice of left or right in this definition is made to ensure that  $C \geq 0.5$ . Hence  $0.5 \leq C < 1$ .

In the run displayed on the left of Fig. 1 there is imperfect and incomplete convergence to one side of the track. The small amount of error slightly disturbs the emergent convention and prevents complete convergence. The death rate (not illustrated) is fairly steady and does not greatly subside. In the run displayed on the right of Fig. 1, complete convergence to a left/right convention is prevented by a minority cohort of 18 cars that defy convention and have a disposition to drive on the other side. When they meet oncoming cars they are able to maneuver to avoid collision. Partly because the error coefficient is zero in this case, no collisions or deaths occur after the first few moves. Consequently, no further evolution of the model in this run is possible and the “cycling” pattern becomes permanent.

Many thousands of distinct runs were tried. In some simulations, the habit weighting ( $w_{Habit}$ ) took the values of 0, 0.5, 1.0, 1.5 and 2.0 in turn. The purpose was to show the effects of increasing weightings to the habit term in the decision function for every car. For each level of  $w_{Habit}$ , the error probability  $\varepsilon$  took the values of 0.000, 0.005, 0.010, 0.015 and 0.020. This meant that 25 combinations of different levels of  $w_{Habit}$  and  $\varepsilon$  were explored. We tried 200 runs, each with 20,000 car moves, for each of the 25 different com-

binations of the values for  $w_{\text{Habit}}$  and  $\varepsilon$ . This meant a total of 5000 runs and 100 million car moves.<sup>12</sup>

We found that the degrees of convergence, the death rates, the effects of error, and the impact of habit can vary substantially, depending on the values of the three parameters  $\{w_{\text{Sdirection}}, w_{\text{Odirection}}, \text{ and } w_{\text{Avoidance}}\}$ . In some regions of parameter space, with a given level of error, increases in the overall strength of habit in the population as a whole (formed by the terms  $w_{\text{Habit}} \times \text{Habitgene} \times \text{Habituation}$ ) can often help to improve the speed of convergence to a left/right convention. In addition,  $w_{\text{Habit}}$  can sometimes help the system cope with error and subvert “cycling” behavior. In other parts of parameter space, the impact of habit on convergence is small or negative.

However, it is important to emphasize that convergence is never achieved by the force of habit alone. Furthermore, convergence can sometimes occur with low or zero levels of habit. Crucially, habit helps convergence only when it is combined with selection pressure on the fixed “instincts” in the population of cars.

The results of a multiple simulation with different levels of  $w_{\text{Habit}}$  and  $\varepsilon$  are reported in Fig. 2 above. The three weights in this model are from the point in parameter space where convergence is maximized with zero habit. The vertical axis on Fig. 2 shows the degree of convergence to a left/right convention. The higher the value the greater the degree of convergence.

A striking outcome displayed in Fig. 2 is the sensitivity of convergence to the habit weighting ( $w_{\text{Habit}}$ ) and strength of habit. As  $w_{\text{Habit}}$  increases, at least from zero to unity, mean convergence levels improve for all levels of error ( $\varepsilon$ ). Habit generally improves convergence.<sup>13</sup>

#### 4.3. Multiple simulations in parameter space

The simulation reported in the previous section is from a point in parameter space where the convergence is maximized with zero habit. The question is raised whether other parts of parameter space exhibit the same positive habit effect, and if so, to what degree.

The first three parameter weights define locations in parameter space. The normalization procedure (where the sum of these weights is always averaged to unity) reduces this to two dimensions. Fig. 3 shows the effect of increasing the habit weight ( $w_{\text{Habit}}$ ) from zero to unity in parameter space. At each point in parameter space, the probability of error was increased uniformly from zero to 0.02, across a set of 200 samples. The increase in the habit weight ( $w_{\text{Habit}}$ ) had a positive effect on convergence over the whole of parameter space.

The *habit effect* is defined as the degree of convergence with  $w_{\text{Habit}}$  as unity, minus the degree of convergence with  $w_{\text{Habit}}$  as zero. In other words, the habit effect is  $C_1 - C_0$ , where  $C_0$  and  $C_1$  are the degrees of convergence for zero and unitary values of the habit weight. Fig. 3 shows values of three  $w_X$  coefficients ( $w_{\text{Sdirection}}, w_{\text{Odirection}}, w_{\text{Avoidance}}$ ) and the habit effect in the whole parameter space.

<sup>12</sup> Experiments with a greater number of car moves are reported in Appendix B for this article in the Elsevier website. Convergence to the left or right was monitored in all runs, confirming that the model had no bias towards one side of the road rather than the other.

<sup>13</sup> More data from this set of simulation results are presented in Appendix A, available on the Elsevier website for this article.

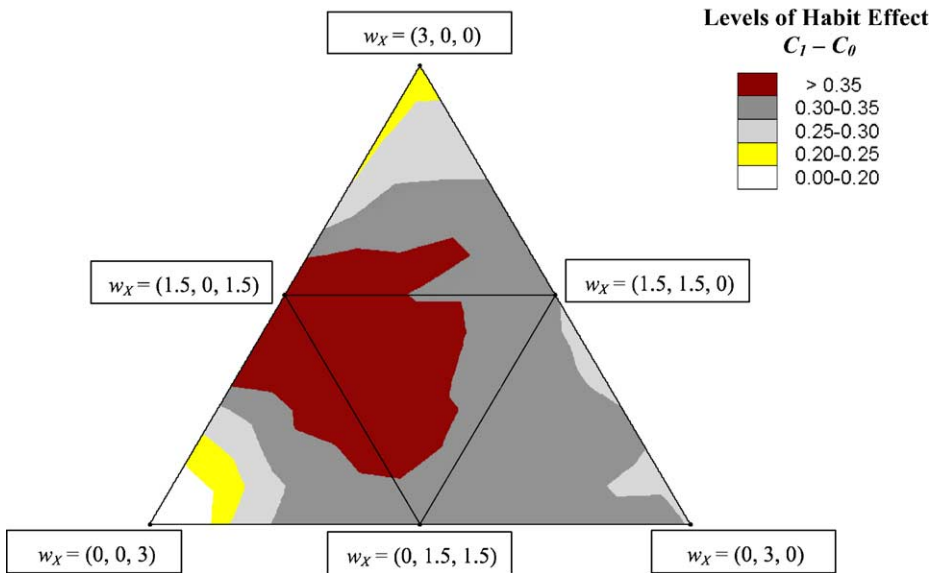


Fig. 3. The habit effect in parameter space.

In another set of simulations, thirty random points were chosen within the parameter space. At each point,  $w_{\text{Habit}}$  took the successive values 0, 0.5, 1 and 1.5, and  $\varepsilon$  took the successive values of 0, 0.01 and 0.02, providing 12 observations at each point and a total of 360 observations overall. At each observation 100 standard runs of the model were made (20,000 car moves) and average figures obtained. The following paragraphs identify the variables that emerge as statistically significant effects in the parameter space.<sup>14</sup>

Based on the data from the 360 observations in parameter space, habit emerged as the most significant factor determining the degree of convergence. A dummy regression (equivalent to ANOVA) was estimated (Regression 1 in Table 1), using  $w_{\text{Habit}}$  as the independent variable and convergence as the dependent variable. According to this regression, the adjusted  $R^2$  was 0.62 and the standardized coefficient was 0.79 ( $t = 24.09$ ) for habit. That is, the linear effect of habit generally increases convergence throughout the parameter space, and the linear effect of habit explains 62 percent of the variation in convergence. A further regression (Regression 2 in Table 1) was estimated, including the randomly generated  $w_X$  coefficients ( $w_{\text{Sdirection}}$ ,  $w_{\text{Odirection}}$ ,  $w_{\text{Avoidance}}$ ) and  $\varepsilon$  as independent variables. Adding these effects marginally increased the explained variation in convergence to 65 percent. Whereas  $\varepsilon$ ,  $w_{\text{Odirection}}$ , and  $w_{\text{Avoidance}}$  generally decrease convergence,  $w_{\text{Sdirection}}$  has no general significant effect. In summary, higher levels of habit significantly improve convergence and higher levels of error have a significant and opposite effect.

The mean death rate is the expected number of drivers to die when all 40 cars move once, averaged over the whole run. As can be seen in Table 1, avoidance emerges unsurprisingly as the most important effect for the reduction of death rates, but the habit effect

<sup>14</sup> We consistently use a significance level of 0.05.

Table 1  
Regression analyses in parameter space

Independent Variables	Dependent variable											
	Convergence (C)						Mean death rate					
	Regression 1			Regression 2			Regression 3			Regression 4		
	Beta	<i>t</i>	<i>P</i>	Beta	<i>t</i>	<i>P</i>	Beta	<i>t</i>	<i>P</i>	Beta	<i>t</i>	<i>P</i>
Constant		80.91	0.00		46.47	0.00		13.03	0.00		4.52	0.00
$w_{\text{Habit}}$	0.79	24.09	0.00	0.79	25.31	0.00	-0.15	-2.86	0.00	-0.15	-4.61	0.00
Error ( $\epsilon$ )				-0.15	-4.89	0.00				0.57	17.56	0.00
$w_{\text{Sdirection}}$				0.00	-0.01	1.00				0.38	11.35	0.00
$w_{\text{Odirection}}$				-0.08	-2.39	0.02				-0.13	-3.67	0.00
$w_{\text{Avoidance}}$				-0.13	-3.99	0.00				-0.36	-10.48	0.00
Adjusted $R^2$	0.62			0.65			0.02			0.62		

Each regression is based on 360 observations, each an average of 100 simulations of 20,000 iterations; *t* is the *t*-statistic; *P* is the *P*-value; and Beta is the standardized coefficient.

also marginally decreases death rates throughout the parameter space. Thus, habit generally improves convergence, but not at the cost of increased death rates.<sup>15</sup>

No other variable emerged in general to improve convergence in our simulations. For instance, while the avoidance coefficient can help the drivers to survive, it does not significantly assist convergence.

We conclude that in this boundedly rational situation, where drivers cannot see the whole of the ring, habit emerged as the single most significant factor improving convergence. If drivers can see further ahead (see Appendix B), habit still has a positive effect. In addition, when the decision horizon is greater than 10, and hence there is more information concerning the traffic ahead, “conformist” factors related to the  $w_{S\text{direction}}$  and  $w_{O\text{direction}}$ , coefficients become significant and more important in aiding convergence. The relative importance of habit is inversely related to omniscience.

## 5. Replacing habit by inertia

Are there alternative mechanisms to habit that can aid convergence? To address this question we considered a modified model where habit is replaced by what we call inertia. As well as the  $SSensitivity_n$ ,  $OSensitivity_n$  and  $Avoidance_n$  coefficients, we added  $Inertia0_n$ ,  $Inertia1_n$ , and  $Inertia2_n$ . As before, for each driver, these coefficients cannot be changed and are randomly assigned (in a normal distribution with mean 1 and standard deviation  $\delta$ , with negative numbers truncated to zero). With inertia coefficients, each driver  $n$  may be equipped with a disposition to continue stubbornly with an inclination it has assumed in the recent past. Each driver may take into account its current (left/right) position in time  $t$  and—with a two-period memory—it may also take into account its actions at times  $t - 1$  and  $t - 2$ . Driver  $n$ 's inertia with respect to times  $t$ ,  $t - 1$  and  $t - 2$  is captured by the three inertia coefficients, respectively:  $Inertia0_n$ ,  $Inertia1_n$  and  $Inertia2_n$ .

To make a decision to go left or right in this modified model, the value of the following expression is calculated:

$$\begin{aligned} LREvaluation_n = & w_{S\text{direction}} \times SSensitivity_n \times (2S_{L,n,t} - 1) + w_{O\text{direction}} \\ & \times OSensitivity_n \times (2O_{L,n,t} - 1) + w_{Avoidance} \times Avoidance_n \\ & \times (C_{R,n,t} - C_{L,n,t}) + w_{Inertia0} \times Inertia0_n \times LR_{n,t} + w_{Inertia1} \\ & \times Inertia1_n \times LR_{n,t-1} + w_{Inertia2} \times Inertia2_n \times LR_{n,t-2}. \end{aligned}$$

<sup>15</sup> Further analyses showed a non-linear relation between habit and convergence. Increasing the habit effect generally increases the degree of convergence but at a decreasing rate. For low levels of error, the habit effect is less important. To obtain better estimates of the variance explained, a mixed effects design (ANCOVA) included  $w_{\text{Habit}}$  and  $\varepsilon$  as fixed effects and the randomly generated  $w_X$  coefficients ( $w_{S\text{direction}}$ ,  $w_{O\text{direction}}$ ,  $w_{Avoidance}$ ) as covariates. Convergence was included as the dependent variable. According to the ANCOVA analyses the habit effect explains 84 percent of the variation in convergence. Adding the  $w_X$  coefficients ( $w_{S\text{direction}}$ ,  $w_{O\text{direction}}$ ,  $w_{Avoidance}$ ) and  $\varepsilon$  as independent variables marginally increased the explained variation in convergence to 88 percent (main effects). Including all interaction effects in a full factorial design further increased the explained variation in convergence to 94 percent. The explanatory power of habit thus remains as the generally most important convergence improving effect throughout parameter space.

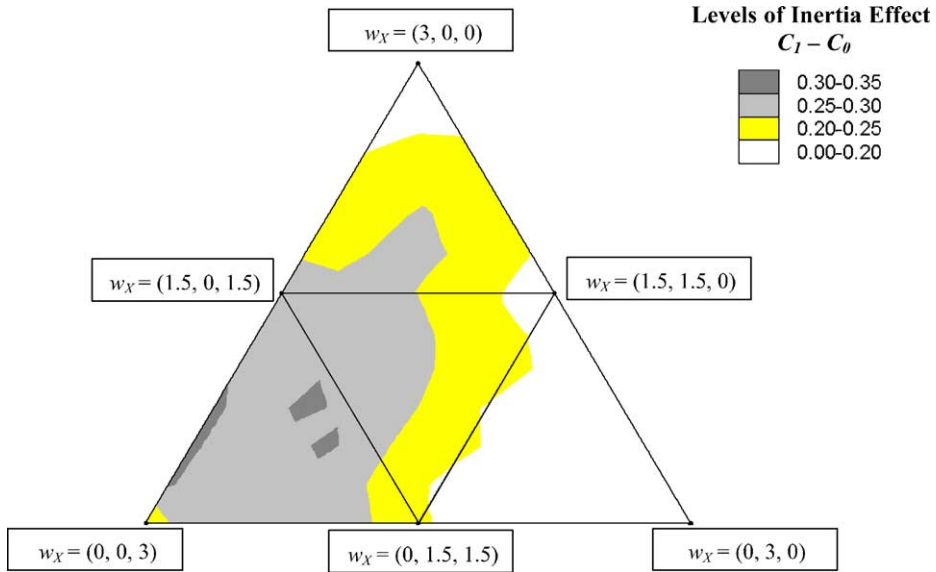


Fig. 4. The inertia effect in parameter space.

Hence the three inertia terms replace habit in the former model. An error factor is included, as before. In this and all other respects the models are identical.

Runs of this model showed that inertia has generally a significantly weaker convergence improving effect than habit.<sup>16</sup> Fig. 4 shows the effect in parameter space of increasing the inertia weights ( $Inertia0_n$ ,  $Inertia1_n$  and  $Inertia2_n$ ) from a state where they are all zero to a state where they take the values 0.17, 0.33 and 0.50, respectively.<sup>17</sup> Note that these three values sum to unity, so the maximum absolute value of the sum of the inertia terms in the left/right decision computation is equal to the maximum absolute value of habit in the standard model. As before, at each point in parameter space, error was increased uniformly from zero to 0.02, across a set of 200 samples. The addition of inertia had a positive effect on convergence over the whole of parameter space. The *inertia effect* is defined and measured as the degree of convergence with  $Inertia0_n = 0.17$ ,  $Inertia1_n = 0.33$ , and  $Inertia2_n = 0.50$ , minus the degree of convergence with zero inertia. In other words, the inertia effect is  $C_1 - C_0$ , where  $C_0$  and  $C_1$  are the respective degrees of convergence for zero and the aforementioned positive values of the inertia weights. Fig. 4 shows the whole parameter space, with values of three  $w_X$  coefficients ( $w_{Sdirection}$ ,  $w_{Odirection}$ ,  $w_{Avoidance}$ ). Comparing Fig. 4 with Fig. 3, it is clear that for most of the parameter space the habit effect is greater than the inertia effect.

It is also useful to consider the conceptual similarities and differences between habit and inertia. Inertia values reflect a memory vector of behavior in a finite set of present

<sup>16</sup> A statistical analysis of many runs, supporting this conclusion, is available from the authors on request.

<sup>17</sup> The inertia values of 0.17, 0.33 and 0.50 were determined by maximizing the degree of convergence ( $C$ ) with  $w_{Sdirection} = 1.4$ ,  $w_{Odirection} = 0.9$ ,  $w_{Avoidance} = 0.7$ ,  $w_{Habit} = 0$ , and with  $Inertia0_n + Inertia1_n + Inertia2_n$  constrained to unity.

and past periods. By contrast, habit in the model is a single-value, weighted summation of behaviors in an unbounded set of present and past periods. Habits are like a crude, summarized memory. Habits are built up steadily once a repeated behavior emerges. Once acquired, they are more difficult (but not impossible in our model) to reverse.<sup>18</sup>

It would be possible to extend the number of preceding periods in the calculation of the inertia values from two to a much higher number. A (weighted) summation of a larger number of these values could then approximate to habit, but with inertia (unlike habit), any informational significance in the values of the individual elements in the inertia vector is retained. However, the cost is an increase in memory and computational capacity.

Crucially, in a more complex world, the number of scalar values represented by multiple inertia vectors relating to multiple past behaviors and behavioral variables would be vastly increased. The storage limits of any finite memory could be readily challenged. By contrast, habit is a cruder summary of past behavior but requires much less memory. This issue of memory and computational limitations is raised again below.

## 6. Discussion—the nature of habit

The most important result of the simulations described in section four concerns the effect of introducing processes of habituation into the modeling of agent behavior. In a substantial region of parameter space, strength of habit can increase the systemic rate of convergence towards a left/right convention. In some circumstances it can also enhance systemic resistance to error.

In the above model, each car is programmed by three parameters ( $SSensitivity_n$ ,  $OSensitivity_n$ ,  $Avoidance_n$ ) governing its sensitivity to traffic patterns ahead and its propensity to make an avoidance maneuver. A fourth parameter ( $Habitgene_n$ ) governs the tendency that a driver has to take account of its acquired habituation. The values of these four exogenously given parameters are akin to instincts: they are fixed for the lifetime of each car. By contrast, a fifth parameter ( $Habituation_{n,t}$ ) governing the particular habitual disposition to go left or right is an outcome of the actual behavior of the car. The value of this parameter is not given and is literally path dependent.

The conception and role of habit in this model contrasts greatly with a definition of habit elsewhere. Becker (1992, p. 328) writes: “I define *habitual* behavior as displaying a positive relation between past and current consumption.” Becker here defines habit not as a behavioral propensity but as sequentially correlated behavior. A car may maneuver to the left to avoid oncoming traffic, but its propensity may still be to drive-to-the-right. If there is an observed succession of left-driving behavior, this is not necessarily the underlying disposition of the agent. Becker’s definition conflates propensity with actuality. However, if past behavior were taken to mean a potentially infinite sequence of past events, then a propensity acquired through habituation could approximate to mean past behavior. In this extreme case, propensity with actuality could coincide, but in general, and in contrast to

<sup>18</sup> As well as comparing the effects of habit (alone) with inertia (alone), we considered their interaction in models where both habit and inertia were present. They jointly influenced convergence outcomes, but in different ways. For low levels of habit, increases in inertia further improve convergence. For high levels of habit, increases in inertia decrease convergence. This further statistical analysis is also available from the authors.

Becker, we distinguish between habit and behavior by defining habit as a disposition or propensity, rather than correlated behavior.

Becker (1992, p. 331) is on stronger ground when he writes: “Habit helps economize on the cost of searching for information, and of applying the information to a new situation.” It is true that habit removes some actions from conscious deliberation and helps the agent to focus on other, more strategic or immediate decisions. However, the model here suggests that there is something more to habit than economizing on decision-making. After all, each car in the model makes only one simple binary decision at each point of time. Habit is doing much more in our model than simply economizing on the time taken to search for and process information.

The model suggests that a crucial role played by habit is to build up and reinforce an enduring disposition in each agent concerning the appropriate side of the road on which to drive, especially in a situation where information concerning the traffic ahead is limited. A sequence of similar and repeated behaviors creates in each agent a habitual predilection, which can stimulate a “belief” or “conviction” that a particular behavior is appropriate.

Again this is reminiscent of the arguments of the pragmatists, who saw acquired habits as the basis of firmly held beliefs. For Peirce (1878, p. 294) the “essence of belief is the establishment of habit.” Similarly, in our model, habit creates stubborn “beliefs” in the appropriateness of an action that weigh heavily in the decision-making process of each agent. The evolution of an equilibrium convention depends largely on one set of stubborn “beliefs” triumphing over the other. Once a stable convention forms, it is encoded in the dispositions of the majority, and it can resist the intrusion of a substantial amount of erratic behavior. As James (1981, p. 125) wrote in 1893: “Habit is thus the enormous fly-wheel of society, its most precious conservative agent.” Accordingly, habit is more than a means of economizing on decision-making for individuals; it is a means by which social conventions and institutions are formed and preserved.

Our model raises questions concerning the distinction between preference exogeneity and endogeneity. By introducing the concept of meta-preferences, Becker and others have argued habit formation is not an example of meta-preference endogeneity. Becker’s (1992, p. 340) argument is that habits and addictions can be placed within a meta-preference function in which data concerning “different variables and experiences,” pertaining to different time periods, enter as arguments. These

*meta-preferences are stable. . . . The message is not that preferences at time  $t$  for different people depend in the same way on their consumption at time  $t$ . Rather, it is that common rules determine the way different variables and experiences enter the meta-preferences that motivate most people at most times (p. 340).*

It is instructive to consider the scale of the mental operation that is implied here. Note that as the number of time periods increases, the number of arguments in Becker’s meta-preference function must increase proportionately. Essentially, Becker argues that utility is a function of the following type:

$$U = f(x_1, x_2 \cdots x_i \cdots x_t)$$

where  $U$  is utility and each  $x_i$  is a vector of “variables and experiences” at time  $i$ . In each complete standard run of our model, each surviving agent moves 500 times, meaning that

its preference function would have to have 500 arguments for each of the five variables involved. However, ours is an extremely simple model, running overall for only 500 iterations per surviving agent. Assume that a nearer-to-human individual lives for 30,000 days, and makes  $10^4$  decisions each day, governing  $10^4$  variables. If so, the Beckerian meta-preference function must have  $3 \times 10^{12}$  arguments. It is likely that the demands of the Beckerian meta-preference function significantly exceed the computational capacities of the human brain. To use the words of Roy Radner (1970, p. 457), the unboundedly rational agent requires “capabilities of imagination and calculation that exceed reality by many orders of magnitude.”

Habit, in the sense that we are using the term, makes computation manageable by vastly reducing the computational and memory requirements of the agent. Habit works not simply or principally by reducing the “cost of searching for information” but also by reducing the memory and computational capacity required to make any decision to act. In formal terms,

$$U = f(\mathbf{h}, \mathbf{m}_{t-s}, \mathbf{m}_{t-s+1}, \dots, \mathbf{m}_{t-1}, \mathbf{x}_t)$$

where  $\mathbf{h}$  is the vector of habits and  $\mathbf{m}_{t-s} \dots \mathbf{m}_{t-1}$  constitute the selective memories of past events, where  $s$  is less than  $t$ . The number of elements in the vector  $\mathbf{h}$  and in any vector  $\mathbf{m}_i$  are each less than the number of “variables and experiences” in  $\mathbf{x}_t$ .

The attribution of a Beckerian preference function to each driver in our model would mean that each driver would have to remember simultaneously, and *for every one of its moves*, at least three computed variables ( $S_{L,n,t}$ ,  $O_{L,n,t}$ ,  $C_{R,n,t} - C_{L,n,t}$ ) plus all of its past left/right positions. If the maximum number of moves were 500, then each agent would require a mental storage for at least 2000 scalar variables. With a greater number of moves the memory requirement increases in proportion. Instead, in our model, only *two* scalar variables (current habituation, plus the number of past moves) have to be stored in the memory of each car at any point in time, *for any length of run*.

Another aspect of Becker’s treatment of individual preferences is also questionable. Becker continues, in the same passage as above,

forward-looking rational actors maximize the utility from their meta-preferences, not from current preferences alone, because they recognize that choices today affect their utilities in the future (p. 340).

In contrast, it could be argued that many actions of agents in the real world, in this respect like the actions of drivers in our model, are not forward-looking in the sense that they consider the full consequences of present actions in the future. Habit is a disposition, sometimes even reinforced by ethical conviction; it does not typically involve a detailed or extensive consideration of future outcomes. No agent in the model considers whether the future emergent convention will be to the left or to the right. It just acts, in part to survive the traffic maelstrom and in part according to its acquired propensity or “belief” that one type of behavior is more appropriate. Of course, things are much more complicated in the real world. People do make decisions based on forward-looking considerations. However, the suggestion here is that forward-looking decisions cannot account for all of behavior, including behavior that is habit-driven. Habit is a past-driven propensity, and not necessarily the outcome of a forward-looking calculation.

For these reasons it is preferable to regard habit formation as an endogenous change of preferences rather than an outcome of decisions governed by a meta-preference function that deals with a number of variables over a series of time periods.<sup>19</sup>

## 7. Discussion—downward causation

Another heuristic use of our model is that it provides a framework to consider the nuanced interpretations and meanings of the concept of “downward causation,” a concept largely unfamiliar to economists but quite well known in the philosophy of psychology and the philosophy of biology (Campbell, 1974; Sperry, 1969, 1991; Popper and Eccles, 1977; Mayr, 1985).

The concept of downward causation depends upon the ontological notion that any complex system has “higher-level” systemic properties as well as “lower-level” components. At the systemic level there may exist “emergent properties” that are, by definition, additional properties that depend upon but are not explicable or predictable from an analysis of the components at the lower level. The concept of emergent properties has recently become prominent in discussions of the complex simulations, pioneered in Santa Fe and elsewhere (Lane, 1993).<sup>20</sup>

Downward causation refers to possible effects of higher-level properties on lower-level components. The term “downward causation” originates in psychology in the work of Sperry (1969). In the literature, the notion of “downward causation” has weak and strong forms. In a relatively weaker case, Donald Campbell sees it in terms of evolutionary laws acting on populations, arguing that all processes at the lower levels of an ontological hierarchy are restrained by and act in conformity to the laws of the higher levels. In other words, if there are systemic properties and tendencies, then individual components of the system act in conformity with them. For example, a population of individual organisms is constrained by processes of natural selection. Here evolutionary processes help to reconstitute populations but not necessarily individuals.

In our model, this weaker form of downward causation is clearly present. As a left/right driving convention begins to be formed, more and more cars drive in conformity with that emerging convention. If a convention begins to emerge, then those that survive tend to be

<sup>19</sup> There is closer relation between our concept of habit and the idea of reinforcement in the works of Ido Erev and Alvin Roth (Erev and Roth, 1998; Roth and Erev, 1995). In their reinforcement model and our habit model, past behavior influences current behavior through more than expectations. Our habit function is thus closer in spirit to the function typically used to model simple reinforcement, but it is not identical. Crucially, the force of the past in our habit function does not decay due to forgetfulness as in Roth and Erev (1995). By contrast, in our formulation of habituation the force of the past accumulates as time unfolds. Erev and Roth (1998) later generalized their reinforcement model to unify reinforcement learning and probabilistic fictitious play. In this generalization, they defined a “subjective reinforcement” of a player’s initial beliefs as the sum of initial expectations and accumulated experience. According to the function used to model this idea, the initial expectations were modified by a time-dependent term defined as the average return of action  $k$  at time  $t$  divided by the number of times that the strategy associated with  $k$  has been played up to  $t$ . This formulation is quite close to the habit function used in our model where habits are built up steadily, and once acquired they are more difficult to reverse. Consequently, our concept of habituation is close to that of reinforcement in their generalized model.

<sup>20</sup> On the concept of emergent properties and its history, see Blitz (1992) and Humphreys (1997).

those that conform. Evolutionary selection acts on the population of agents, causing a shift in the characteristics of the population as a whole. This is an outcome of “natural selection” and amounts to weak downward causation.

In the population as a whole, this evolutionary selection works on both the four fixed parameters and the single variable expressing habit. The set of values in the population as a whole changes by means of the death of the unsuccessful and the birth of the new agents. However, for any individual agent, evolutionary selection does not cause a change in the values of the four fixed parameters.

A stronger notion, which can be described as “reconstitutive downward causation,” involves changes acting on individuals as well as populations as a result of causal powers associated with higher levels (Hodgson, 2003, *in press*). Sperry (1991, p. 230-1) also suggests a strong interpretation of downward causation. He recognizes, for example, that “higher cultural and other acquired values have power to downwardly control the more immediate, inherent humanitarian traits.” Sperry also recognizes that explanations based on downward causation should be carefully focused on real causal mechanisms. This is the problem: while it is tempting to explain the behavior of units in terms of collectives or wholes, the precise causal mechanism is difficult to determine.

If there is some mechanism whereby an actual or emerging convention can affect or “reconstitute” the characteristics of the individual units, then this would amount to reconstitutive downward causation. System-wide outcomes (at a “higher” level) would affect the characteristics of individual units (at a lower level).

In our model, this stronger form of downward causation is also present and is associated with a discernible causal mechanism because as the left/right convention begins to emerge, more and more surviving cars develop the habit to drive on the left or the right, according to that convention. Strength of habit is based on two of the five variables that form the “preference function” of each agent. For each individual, one of these preference elements (Habituation<sub>*n,t*</sub>) can change. In this way, emerging and enduring systemic properties reconstitute “downwards” the preferences of the agent. Part of the achievement here is to show that both forms of downward causation can be represented in an agent-based model. In particular, we can identify a specific causal mechanism of reconstitutive downward causation.

Another crucial point to recognize is the specific mechanism by which reconstitutive downward causation operates. It is on *habits* rather than merely on behavior, intentions or other preferences. Clearly, the definitional distinction between habit (as a propensity or disposition) and behavior (or action) is essential to make sense of this statement. Long ago Veblen (1899, p. 190) similarly identified habit as the psychological mechanism by which circumstances change preferences or dispositions: “The situation of today shapes the institutions of tomorrow through a selective, coercive process, *by acting upon men’s habitual view of things*” (emphasis added).

The existence of a viable mechanism of reconstitutive downward causation contrasts with other, untenable “top-down” or “methodologically collectivist” explanations in the social sciences where there are unspecified “structural,” “cultural,” or “economic” forces controlling individuals. Crucially, the mechanism of reconstitutive downward causation that is outlined here affects the dispositions, thoughts, and actions of human actors. People do not develop new preferences, wants or purposes because mysterious “social forces” control them. What does happen is that the framing, shifting and constraining capacities of social

institutions give rise to new perceptions and dispositions within individuals. Upon new habits of thought and behavior, new preferences and intentions emerge.

Hence, the concept of reconstitutive downward causation does not rely on new or mysterious types of cause or causality. As Sperry (1991, p. 230) rightly insists, “the higher-level phenomena in exerting downward control do *not disrupt* or *intervene* in the causal relations of the downward-level component activity.” Sperry’s maxim excludes any version of methodological collectivism or holism where an attempt is made to explain individual dispositions or behavior entirely in terms of institutions or other system-level characteristics. Instead, we are obliged to explain particular human behavior in terms of causal processes operating at the individual level, such as individual aspirations, dispositions or constraints.

It is a central tenet of the pragmatist philosophical and psychological perspective to regard habit and instinct as foundational to the human personality. Reason, deliberation and calculation emerge only after specific habits have been laid down; their operation depends upon such habits. In turn, the development of habits depends upon prior instincts. Instincts, by definition, are inherited. Accordingly, *reconstitutive* downward causation upon instincts is not possible. However, as noted above, the weaker form of downward causation does operate on whole populations and on its pool of habits and instincts.

The ongoing acquisition and modification of habits is central to human existence. All action and deliberation depend on prior habits that we acquire during our individual development. For example, much deliberative thought is dependent on, as well as being colored by, acquired habits of language. In addition, to make sense of the world, we have to acquire habits of classification and habitually associated meanings. To act in and adapt to the world, our complex nervous system has to be developed and rehearsed. Habit is a crucial and neglected element in cognition, deliberation and reason.

As long as we can explain how institutional structures give rise to new or changed habits, then we have a possible and acceptable mechanism of reconstitutive downward causation. Of course, institutions may directly affect our intentions by providing incentives, sanctions or constraints. In contrast, a reconstitutive causal mechanism involves factors that are foundational to purposes, preferences and deliberation as a whole (Margolis, 1987). This is where habits come in. By affecting habits, institutions can indirectly influence our intentions (Hodgson, 2003, *in press*).

## 8. Conclusion

The model discussed in this article shows how a left/right traffic convention may emerge in an agent-based model. The main factor inhibiting this convergence is error. Also, in limited circumstances, agile avoidance behavior can lead to recurrent, cycling patterns of behavior with no emergent left/right convention. The simulation results show that increases in the “strength of habit” of agents in the model when combined with evolutionary selection pressure can help to suppress both of these disturbing factors.

This simulation points to some of the deeper conceptual issues involved in the evolution of conventions, particularly the nature of rational decision-making and its reliance upon habit. Overall, the simulations show that the systemic convergence to a left/right convention is often improved and sustained by strength of habit. Accordingly, habit plays an important

part alongside the “intelligent” and calculative aspects of agent behavior, particularly in cases where information is limited.

In contrast the analyses of Jones (1984) and Schlicht (1998) maintain that conventions and customs emerge principally because individuals have a preference for them. In our simulations, this is not generally the case where information is limited. In these circumstances, habit is additionally and vitally important because it can often enhance stable behavior and help to create stable outcomes.

The specification of habit in the model is redolent of the concept in the works of pragmatist philosophers such as Peirce and James. Habit acts in the model as if it were the foundation of a “conviction” or firmly held “belief.” This suggests that the evolution of conventions may depend not only on the rational calculations of actors but also on the widespread development of convictions or norms concerning appropriate behavior.

This model also has implications for an understanding of the nature and role of habit. In the specification here, the conception of habit is clearly distinguished from serially correlated behavior. This definition contrasts significantly with that in the work of Becker and others.

We also identify a mechanism of “reconstitutive downward causation” among agents. Although each car has four inert “instincts,” the fifth variable concerning habituation changes as agent behavior changes. As a left/right convention emerges among the population as a whole, this provides a channel of movement for every individual. Accordingly, individual habits reflect the emergent convention among the whole population. As a result, the formation of individual habits is guided by systemic conventions. This is tantamount to a change of preferences, and it results from a “downwards” causal process from the emergent institution to the individual.

A possible criticism of this thesis could stem from a Beckerian approach where each agent has a meta-preference function with arguments representing all relevant temporal and other variables. We have shown that this approach comes up against the problem of computational limitations of agents required to deal with large and increasing amounts of information concerning their past. It makes more sense to treat preferences as partially endogenous and malleable. Furthermore, in contrast to the idea of a meta-preference function, the conception of habit defined here greatly reduces the number of variables that each agent has to take into account.

Given the powerful effect of habituation in our model, reconstitutive downward causation may provide a degree of durability and stability in institutional structure that is not explained adequately in standard models. The circular, positive feedback from institution to individuals and from individuals to institutions can help to enhance the durability of the institutional unit. There may be stable emergent properties that exist *not despite*, but *because of*, endogenous preference formation.

With the theoretical framework proposed here, it may also be possible to overcome the dilemma between methodological individualism and methodological collectivism. By acting not directly on individual decisions, but on habitual dispositions, institutions exert reconstitutive downward causation without reducing the role of individual agency. Upward causation, from individuals to institutions, is still possible, without assuming that the individual is given or immanently conceived. Explanations of socio-economic phenomena are reduced neither to individuals nor to institutions alone.

## Acknowledgements

The authors wish to thank Brian Arthur, Kenneth Binmore, Nathalie Lazaric, Axel Leijonhufvud, Paul Ormerod, J. Barkley Rosser Jr., Koye Somefun, Robert Sugden, Margherita Turvani, Viktor Vanberg, Kumaraswamy Velupillai, Ulrich Witt, three anonymous referees and others for their comments on earlier drafts of this paper.

## Appendix A. Further results from simulations with the standard model

In this appendix, the results of the standard model simulations outlined in the text and shown in Fig. 2 are reported in more detail. All tables show results for 25 combinations of error ( $\varepsilon$ ) and habit weight ( $w_{\text{Habit}}$ ): five levels of error and five levels of habit weight. The value shown for each of these twenty five levels is based on the average of 200 standard runs, and each run was based on 20,000 iterations.

Table A.1 shows the mean values and standard deviations for the degrees of convergence to a left/right convention. The mean values from this table are presented in Fig. 2 above. As noted already, with these mean values, 1.0 would indicate that all cars were on the left or right for the entire run, and 0.5 would indicate that there was no mean inclination towards either side of the road. According to a *t*-test, the convergence values for  $w_{\text{Habit}} = 1.0$  are significantly higher than the convergence values for  $w_{\text{Habit}} = 0.5$ , and the convergence values for  $w_{\text{Habit}} = 0.5$  are significantly higher than the convergence values for  $w_{\text{Habit}} = 0$ . Higher values of  $w_{\text{Habit}}$  (1.5 and 2.0) do not lead to higher convergence values than  $w_{\text{Habit}} = 1.0$ .

Table A.2 shows data on the average proportion of time in each case where all cars were on the same side of the road. Generally, this proportion of time at unanimity increases as the habit weight increases from zero to 1.5. According to a *t*-test, the unanimity values for  $w_{\text{Habit}} = 1.5$  are significantly higher than the unanimity values for  $w_{\text{Habit}} = 1.0$ , the unanimity values for  $w_{\text{Habit}} = 1.0$  are significantly higher than for  $w_{\text{Habit}} = 0.5$ , and the unanimity values for  $w_{\text{Habit}} = 0.5$  are significantly higher than for  $w_{\text{Habit}} = 0$ . There are no significant differences in the unanimity values for  $w_{\text{Habit}} = 1.5$  and  $w_{\text{Habit}} = 2.0$ .

Table A.3 shows the expected number of drivers to die when all 40 cars move once, averaged over the whole run. However, as noted below, these death rates are much higher in the earlier phase of each run. Mean death rates generally increase with error. In addition, for any level of error, an increase in the weight of habit, up to 1 or more, significantly reduces

Table A.1  
Degrees of convergence to a left/right convention

Mean values	$w_{\text{Habit}}$					S.D.	$w_{\text{Habit}}$				
	0	0.5	1	1.5	2		$\varepsilon$	0	0.5	1	1.5
0.000	0.638	0.966	0.975	0.974	0.971	0.000	0.123	0.056	0.032	0.032	0.036
0.005	0.599	0.899	0.962	0.968	0.967	0.005	0.075	0.109	0.032	0.034	0.034
0.010	0.592	0.843	0.946	0.959	0.962	0.010	0.064	0.135	0.055	0.033	0.028
0.015	0.572	0.827	0.931	0.941	0.951	0.015	0.058	0.125	0.069	0.049	0.032
0.020	0.577	0.778	0.902	0.932	0.944	0.020	0.056	0.128	0.082	0.053	0.032

Table A.2  
Proportion of time at unanimity

Mean values		$w_{\text{Habit}}$					S.D.	$w_{\text{Habit}}$				
$\varepsilon$		0	0.5	1	1.5	2	$\varepsilon$	0	0.5	1	1.5	2
0.000		0.363	0.908	0.921	0.920	0.909	0.000	0.198	0.115	0.077	0.075	0.084
0.005		0.233	0.630	0.713	0.730	0.728	0.005	0.064	0.113	0.061	0.074	0.072
0.010		0.156	0.449	0.550	0.572	0.579	0.010	0.042	0.092	0.066	0.057	0.053
0.015		0.108	0.325	0.416	0.439	0.450	0.015	0.032	0.064	0.052	0.054	0.049
0.020		0.078	0.222	0.306	0.342	0.354	0.020	0.025	0.056	0.051	0.043	0.043

Table A.3  
Death rates for whole run

Mean values		$w_{\text{Habit}}$					S.D.	$w_{\text{Habit}}$				
$\varepsilon$		0	0.5	1	1.5	2	$\varepsilon$	0	0.5	1	1.5	2
0.000		0.831	0.125	0.107	0.112	0.130	0.000	0.279	0.106	0.074	0.079	0.093
0.005		1.343	0.571	0.445	0.414	0.430	0.005	0.216	0.167	0.104	0.111	0.120
0.010		1.788	0.959	0.769	0.738	0.737	0.010	0.220	0.200	0.134	0.141	0.118
0.015		2.175	1.354	1.129	1.068	1.065	0.015	0.229	0.208	0.158	0.142	0.151
0.020		2.534	1.730	1.454	1.381	1.380	0.020	0.231	0.235	0.192	0.171	0.169

the death rate. According to a *t*-test, all differences in death rates between successive levels of  $w_{\text{Habit}}$  are significant.

However, death rates fall dramatically once convergence is established. Table A.4 shows the expected number of drivers to die when all 40 cars move once, for the second half of every simulation. Note that the death rates are much lower than in Table A.3. Generally, the majority of deaths occur in the early, transition phase of the simulation. With low levels of error, death rates are very low for all levels of habit. Once again, death rates increase with error. According to a *t*-test, death rates for the second half of each run differ significantly for  $w_{\text{Habit}} = 0$  and  $w_{\text{Habit}} = 0.5$ , and for  $w_{\text{Habit}} = 1.0$  and  $w_{\text{Habit}} = 1.5$ . There are no significant differences in the second half death rates for the other successive values of  $w_{\text{Habit}}$ .

Table A.4  
Death rates for the second half of each run

Mean values		$w_{\text{Habit}}$					S.D.	$w_{\text{Habit}}$				
$\varepsilon$		0	0.5	1	1.5	2	$\varepsilon$	0	0.5	1	1.5	2
0.000		0.177	0.001	0.000	0.000	0.000	0.000	0.087	0.009	0.000	0.000	0.002
0.005		0.319	0.112	0.084	0.076	0.079	0.005	0.078	0.054	0.028	0.024	0.024
0.010		0.430	0.218	0.166	0.157	0.154	0.010	0.072	0.062	0.041	0.038	0.037
0.015		0.533	0.323	0.258	0.238	0.235	0.015	0.079	0.075	0.051	0.041	0.042
0.020		0.624	0.408	0.337	0.315	0.317	0.020	0.084	0.080	0.060	0.048	0.051

## Appendix B. Different decision horizons and longer runs

The decision horizon affects the calculations concerning the pattern of traffic ahead. The standard results reported above were obtained in simulations in which each driver had a decision horizon of 10 and each car made 500 moves. With a horizon of 20 rather than 10, but no increase in the number of iterations, the importance of  $S_{L,n}$  increases substantially and the effect of habituation on improving convergence becomes insignificant. Then typically the degree of convergence is significantly related to  $w_{S\text{direction}}$  and error, where increases in  $w_{S\text{direction}}$  improve convergence. With these shorter runs, habituation is significant where the decision-making horizon is smaller and information concerning the general pattern of traffic is limited. Habit becomes more significant as rationality is bounded.

However, supplementary analyses were conducted to determine whether the effect of habit was significant when the drivers' horizons and the length of the runs were *both* increased. Long runs of 100,000 iterations were used, noting that the selection dynamics became stable for at least half of the runs of this length. It was examined whether habit had a significant convergence improving effect when the horizon was increased in steps up to 100. A horizon of 100 is the point of omniscience, where the whole ring is in view.

We again sampled random weights within the parameter space, as described previously above. In every instance, the  $w_{S\text{direction}}$ ,  $w_{O\text{direction}}$ , and  $w_{\text{Avoidance}}$  coefficients were assigned randomly generated and then normalised values. A number of samples of 30 at each point was sufficient, since the standard deviations of the convergence outcomes are lower when a longer decision horizon is used.

At each randomly generated point,  $w_{\text{Habit}}$  took the successive values 0.0, 0.5, 1.0 and 1.5, and  $\varepsilon$  took the successive values of 0.00, 0.01 and 0.02. This provided 12 observations at each point and a total of 360 observations overall. We continued to use the levels of error ( $\varepsilon$ ) between 0.00 and 0.02.

As defined in the main text, the *habit effect* is the degree of convergence with  $w_{\text{Habit}}$  as unity, minus the degree of convergence with  $w_{\text{Habit}}$  as zero. In statistical terms, a *t*-test of the difference between the convergence means for  $w_{\text{Habit}} = 0$  and  $w_{\text{Habit}} = 1$  (error ranging uniformly across the sample from 0 to 0.02 for both values of  $w_{\text{Habit}}$ ) was used to determine whether the habit effect was significant.

The key result shown in Table A.5 is that habit significantly improves the mean degree of convergence for horizons from zero up to and including 25 zones of the ring. The maximum habit outcome is at a horizon of 10, which is the value used in the standard runs reported in the main text.

However, with horizons above 25, in the absence of habit ( $w_{\text{Habit}} = 0$ ) the degree of convergence exceeded 0.98. In this 30–100 range, the habit effect was insignificant. Nevertheless, even in cases where habit did not significantly increase convergence, there was not a significant decrease of convergence.

The results for each value of the decision horizon are mean values of convergence based on 360 observations, each based on the mean of 30 samples. The *P*-value of the habit effect is the *P*-value of the *t*-test, comparing the mean for  $w_{\text{Habit}} = 0$  and  $w_{\text{Habit}} = 1$ .

In the analyses reported so far, all drivers were endowed with fixed identical values of horizon (0, 5, 10, 15, 20, 25, 30, 50 and 100). Further analyses were conducted to assess the effect of variations in horizon among drivers. The results were based on 360 observations

Table A.5

Degrees of convergence and death rates in long runs with different horizons

Horizon	Degrees of convergence				Death rates			
	$w_{\text{Habit}} = 0$	$w_{\text{Habit}} = 1$	Habit effect	<i>P</i> -value effect	$w_{\text{Habit}} = 0$	$w_{\text{Habit}} = 1$	Habit effect	<i>P</i> -value effect
0	0.50	0.53	0.03	0.00	0.27	3.56	−3.29	0.00
5	0.51	0.80	0.30	0.00	1.15	0.69	0.46	0.00
10	0.56	0.88	0.32	0.00	1.68	0.86	0.82	0.00
15	0.69	0.95	0.26	0.00	0.72	0.63	0.09	0.33
20	0.85	0.98	0.13	0.00	0.59	0.58	0.01	0.95
25	0.94	0.98	0.04	0.00	0.43	0.53	−0.10	0.12
30	0.98	0.98	0.00	0.15	0.51	0.56	−0.05	0.43
50	0.99	0.99	0.00	0.77	0.53	0.56	−0.03	0.65
100	0.98	0.98	0.00	0.36	0.51	0.59	−0.07	0.29
Random	0.97	0.98	0.01	0.00	0.55	0.61	−0.06	0.45
Mutating	0.50	0.84	0.33	0.00	1.35	1.08	−0.27	0.05

of 30 samples each, and the number of iterations was 100,000. Horizon values in the range from zero to 100 were randomly assigned to each driver with an equal probability. Runs of this model with random assignment and replacement of values for horizon, the selection dynamics only led to a slight change in the value of drivers' horizon. See the results in the row marked 'Random' in Table A.5. After 100,000 iterations the mean horizon in the population was 51.67, only slightly above the initial expected mean value of 50. However, habit significantly improved the mean degree of convergence even at this relatively high average value of horizon.

As can be seen from Table A.5, habit does not generally increase death rates. When the horizon is at a value of 5 or 10, habit not only increases convergence, but also significantly decreases death rates. Death rates are not significantly influenced either by habit when the horizon is above 10 or by drivers being allocated a random horizon. Only at the extreme limit of drivers scanning zero zones does habit increase death rates.

Finally, we tested a model with a mutating random assignment in order to determine whether habit supported the evolution of higher values of the decision horizon. A value of horizon was initially randomly assigned to each driver. With probability 0.4 they received a horizon of 0, and with probability 0.6 they received a horizon between 1 and 100. The values above 0 followed a Poisson distribution with mean approximately 20. The expected horizon value was approximately 12 zones for this model. When drivers crashed, the decision horizon was replaced according to the same random assignment procedure. The results show that habit significantly increased the mean horizon of the population of drivers that had evolved after 100,000 iterations. For  $w_{\text{Habit}} = 0.0$ , the mean value of the drivers' horizon was 8.29. When  $w_{\text{Habit}}$  was increased to 0.5, there was a significant increase ( $P = 0.00$ ) of the mean value of the drivers' horizon to 10.70. Further increases in  $w_{\text{Habit}}$  to 1.0 (and 1.5) resulted in further significant increases ( $P = 0.00$ ) of the mean value of the drivers' horizon to 12.41 (and 12.65). Habit thus supported the evolution of a higher decision horizon, which in turn significantly improved convergence. See the results in the row marked 'Mutating'

in Table A.5. Other test runs show that this result is general for a number of distributions of horizon.

Since drivers have more information as the value of horizon increases, the wider implication is that habit must be viewed as a complement to, rather than a detractor from, deliberative rationality. This is because habit does not decrease convergence even in the case where drivers scan all 100 zones of the entire ring. As can be seen from Table A.5, neither does habit come at a cost of significantly increased death rates, unless the drivers' are endowed with a zero horizon.

## References

- Alessie, R., Kapteyn, A., 1991. Habit formation, interdependent preferences and demographic effects in the almost ideal demand system. *Economic Journal* 101, 404–419.
- Aoki, M., 2001. *Toward a Comparative Institutional Analysis*. MIT Press, Cambridge, MA.
- Arthur, W.B., 1994. *Increasing Returns and Path Dependence in the Economy*. University of Michigan Press, Ann Arbor, MI.
- Becker, G.S., 1992. Habits, addictions and traditions. *Kyklos* 45, 327–346.
- Becker, G.S., Murphy, K.M., 1988. A theory of rational addiction. *Journal of Political Economy* 96, 675–700.
- Blanciforti, L., Green, R., 1983. An almost ideal system incorporating habits: an analysis of expenditures on food and aggregate commodity groups. *Review of Economics and Statistics* 65, 511–515.
- Blitz, D., 1992. *Emergent Evolution: Qualitative Novelty and the Levels of Reality*. Kluwer, Dordrecht.
- Campbell, D.T., 1974. "Downward causation" in hierarchically organized biological systems'. In: Ayala, F.J., Dobzhansky, T. (Eds.), *Studies in the Philosophy of Biology*. Macmillan, London, pp. 179–186.
- Duffy, J., Ochs, J., 1999. Emergence of money as a medium of exchange: an experimental study. *American Economic Review* 89, 847–877.
- Eiser, J.R., Pahl, S., Prins, Y.R.A., 2001. Optimism, pessimism, and the direction of self-other comparisons. *Journal of Experimental Social Psychology* 37, 77–84.
- Erev, I., Roth, A.E., 1998. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* 88, 848–881.
- Field, A.J., 1979. On the explanation of rules using rational choice models. *Journal of Economic Issues* 13, 49–72.
- Field, A.J., 1984. Microeconomics, norms and rationality. *Economic Development and Cultural Change* 32, 683–711.
- Hodgson, Geoffrey M., 1993. *Economics and Evolution: Bringing Life Back Into Economics*. Polity Press and University of Michigan Press, Cambridge, UK and Ann Arbor, MI.
- Hodgson, G.M., 1998. The approach of institutional economics. *Journal of Economic Literature* 36, 166–192.
- Hodgson, G.M., 2003. The hidden persuaders: institutions and individuals in economic theory. *Cambridge Journal of Economics* 27, 159–175.
- Hodgson, G.M. *The Evolution of Institutional Economics: Agency, Structure and Darwinism in American Institutionalism*. Routledge, London, in press.
- Howitt, P., Clower, R.W., 2000. The emergence of economic organization. *Journal of Economic Behavior and Organization* 41, 55–84.
- Hull, C., 1943. *Principles of Behavior: An Introduction to Behavior Theory*. Appleton-Century, NY.
- Humphreys, P., 1997. How properties emerge. *Philosophy of Science* 64, 1–17.
- James, W., 1981. *The Principles of Psychology* (reprinted in two volumes from the 1893 edition). Harvard University Press, Cambridge, MA.
- Jones, R.A., 1976. The origin and development of media of exchange. *Journal of Political Economy* 84, 757–775.
- Jones, S.R.G., 1984. *The Economics of Conformism*. Basil Blackwell, Oxford.
- Kiyotaki, N., Wright, R., 1989. On money as a medium of exchange. *Journal of Political Economy* 97, 927–954.
- Knight, J., 1992. *Institutions and Social Conflict*. Cambridge University Press, Cambridge.
- Lane, D.A., 1993. Artificial worlds and economics (Parts I and II). *Journal of Evolutionary Economics* 3, 89–107, 177–97.

- Margolis, H., 1987. *Patterns, Thinking and Cognition: A Theory of Judgment*. University of Chicago Press, Chicago.
- Marimon, R.E., Ellen, M., Thomas, J.S., 1990. Money as a medium of exchange in an economy with artificially intelligent agents. *Journal of Economic Dynamics and Control* 14, 329–373.
- Mayr, E., 1985. How biology differs from the physical sciences'. In: Depew, D.J., Weber, B.H. (Eds.), *Evolution at a Crossroads: The New Biology and the New Philosophy of Science*. MIT Press, Cambridge, MA, pp. 43–63.
- Menger, C., 1981. In: Dingwall, J. (Eds.), *Principles of Economics* (translated by Hoselitz, B.F. from the German edition of 1871). New York University Press, NY.
- Mitchell, W.C., 1937. *The Backward Art of Spending Money and Other Essays*. McGraw-Hill, NY.
- Oh, S., 1989. A theory of a generally acceptable medium of exchange and barter. *Journal of Monetary Economics* 23, 101–119.
- Peirce, C.S., 1878. How to make our ideas clear. *Popular Science Monthly* 12, 286–302. Reprinted in Peirce, Charles Sanders, 1958. *Selected Writings (Values in a Universe of Chance)*, edited with an introduction by P.P. Wiener. Doubleday, NY.
- Petty, R.E., Cacioppo, J.T., 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude and Change*. Spinger-Verlag, NY.
- Philps, L., Spinnewyn, F., 1984. True indexes and rational habit formation. *European Economic Review* 24, 209–223.
- Popper, K.R., Eccles, J.C., 1977. *The Self and Its Brain*. Springer, Berlin.
- Pollak, R.A., 1970. Habit formation and dynamic demand functions. *Journal of Political Economy* 78, 745–763.
- Radner, R., 1970. New ideas in pure theory: problems in the theory of markets under uncertainty. *American Economic Review (Papers and Proceedings)* 60, 454–460.
- Roth, A.E., Erev, I., 1995. Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior, Special Issue: Nobel Symposium 8*, 164–212.
- Schlicht, E., 1998. *On Custom in the Economy*. Clarendon Press, Oxford.
- Searle, J.R., 1995. *The Construction of Social Reality*. Allen Lane, London.
- Sened, I., 1997. *The Political Institution of Private Property*. Cambridge University Press, Cambridge.
- Sperry, R.W., 1969. A modified concept of consciousness. *Psychological Review* 76, 532–536.
- Sperry, R.W., 1991. In defense of mentalism and emergent interaction. *Journal of Mind and Behavior* 12, 221–246.
- Sugden, R., 1986. *The Economics of Rights, Co-operation and Welfare*. Basil Blackwell, Oxford.
- Veblen, T.B., 1899. *The Theory of the Leisure Class: An Economic Study in the Evolution of Institutions*. Macmillan, NY.
- Wärneryd, K., 1989. Legal restrictions and the evolution of media of exchange. *Journal of Institutional and Theoretical Economics* 145, 613–626.
- Wärneryd, K., 1990a. Legal restrictions and monetary evolution. *Journal of Economic Behavior and Organization* 13, 117–124.
- Wärneryd, K., 1990b. *Economic Conventions: Essays in Institutional Economics*. Economic Research Institute, Stockholm.
- Winston, G.C., 1980. Addiction and blacksliding: a theory of compulsive consumption. *Journal of Economic Behavior and Organization* 1, 295–324.
- Young, H.P., 1993. The evolution of conventions. *Econometrica* 61, 57–84.
- Young, H.P., 1996. The economics of convention. *Journal of Economic Perspectives* 10 (2), 105–122.