

# ***Are motivational biases adaptive? An agent-based model of human judgement under uncertainty***

Dylan Evans<sup>1</sup>, Annerieke Heuvelink<sup>1 and 2</sup> and Daniel Nettle<sup>3</sup>

<sup>1</sup> Corresponding author:  
Biomimetics Group  
Department of Mechanical Engineering  
University of Bath  
Bath BA2 7AY  
UK  
Email: d.evans@bath.ac.uk

<sup>2</sup> Department of Philosophy  
University of Utrecht  
Heidelberglaan 8  
3584 CS Utrecht  
Netherlands

<sup>3</sup> Departments of Psychology and Biological Sciences  
The Open University  
Walton Hall  
Milton Keynes MK7 6AA  
UK

## **Abstract**

Human judgement has been shown to involve a number of biases, under which people's assessment of the likelihood of a state of affairs is affected by their judgement of how desirable that state of affairs is. It has been suggested that such biases are adaptive, since in an uncertain world they lead people to make the least costly and most beneficial decisions. However, it has never been shown formally that making biased decision-making could be an evolutionarily adaptive strategy. We approach this question by constructing an agent-based computer model, in which agents with different decision rules are allowed to compete in a variety of environments. We find that under certain environmental conditions unbiased agents using classical decision theory are outperformed by agents with biases similar to those observed in humans in empirical studies. Our model supports the view that the biases people display in judgement under uncertainty may be design features rather than design flaws of the human mind.

## Introduction

Human judgement under uncertainty has been shown to involve consistent departures from normative rationality (Kahneman and Tversky, 1973; Kahneman, Slovic and Tversky 1982). Amongst these departures are 'motivational biases', in which people's judgements about the probability of the occurrence a future event are influenced by their estimates of the costs and benefits of such an event. The classic manifestation of motivational bias is over-optimism about one's own future prospects or ability to achieve a desired goal. Such effects have been observed empirically across a wide range of tasks (Miller and Ross 1975, Larwood and Whitaker 1977, Alloy and Abramson 1979, Zuckerman 1979, Lewinsohn et al. 1980, Mirels 1980, Weinstein 1980, Alicke 1985, Campbell 1986, Vazquez 1987, Rudski 2000; for a review see Taylor and Brown 1988) The biases disappear when people are depressed ('depressive realism': Alloy and Abramson 1979) and when people are asked to estimate the probability of the same events happening to *others* (Mirels 1980, Lewinsohn et al. 1980), so unbiased judgements are possible. When non-depressed participants make judgements about future events involving the self, though, motivational biases are remarkably resilient.

Taylor and Brown (1988) argue that motivational biases are important and characteristic of normal, healthy human cognition. However, from the standpoint of probability theory, they appear to be irrational and thus are most obviously viewed as defects in human reason compared to a rational norm. Recently some psychologists have argued that they could nonetheless be adaptive. Hasleton and Buss (2000) tackle cognitive biases from the standpoint of what they call 'error management theory'. Every judgement has two possible errors – in the case of a judgement about likelihood, for example, it is possible to under or over estimate. The costs and benefits of making the two errors are not necessarily symmetrical. For example, if an action has a large potential benefit to the self, and only a small cost if it goes wrong, then an overestimate of one's potential for success will do little harm, and just might bring in a large benefit. An underestimate of one's potential for success could lead to apathy and missing out on opportunities. If there is irreducible uncertainty about true probabilities, then, an overestimate may be better than an underestimate for the likelihood of events with a large potential return to the self. Thus biases, rather than being design flaws, may be better seen as adaptive reasoning strategies in an uncertain world (Hasleton and Buss in press, Nettle in press).

This argument appears plausible. However, adaptive explanations should not rest on mere assertion, but should be grounded in an independent demonstration that the proposed mechanism could indeed lead to increased fitness, and an exploration of the conditions under which it could evolve. In evolutionary biology, such explorations are carried out using various kinds of formal and computer modelling, in which the returns to possible behaviours or structures are simulated under a wide variety of conditions.

We have investigated the adaptive hypothesis for motivational biases by constructing an agent-based computer model. In the model, agents with different decision rules are allowed to compete in a variety of environments. We find that under certain environmental conditions the unbiased agent, who uses a simple expected value calculation as a decision rule, is outperformed by biased agents who behave in ways

similar to those observed in humans. It can be plausibly argued that the conditions under which these biased agents outperform the unbiased agent are similar to those which human beings most frequently encounter. Our findings therefore add support to the view that motivational biases are in fact adaptive.

## Methods

We designed an agent-based simulation using NetLogo, an agent-based parallel modelling and simulation environment produced by the Center for Connected Learning and Computer-Based Modelling at Northwestern University (Wilensky, 1999). In the model, three types of agent encounter successive opportunities, each of which is characterised by three properties: a probability of success ( $p$ , ranging from 0-1), a benefit of success ( $b$ , ranging from 0.0001 to 10 energy points) and a cost of failure ( $c$ , ranging from 0.0001 to 10 energy points).

Agents have only one goal - to maximise their energy points. In other words, their utility function is a linear function of their energy level. Agents have some knowledge of the cost of failure ( $c$ ), the benefit for success ( $b$ ), and the probability of success ( $p$ ), for each opportunity they face. The level of error affecting the agents' knowledge of these values can be set by the user to integer values between 0 (perfect information) to 10 (great uncertainty) inclusive. This error level determines the standard deviation used for a normal distribution of which the mean is the true value of  $c$ ,  $b$  or  $p$  of the opportunity. A random number drawn from this distribution determines the agents' guesses about the values of  $c$ ,  $b$  and  $p$ . When the error is high, the standard deviation can be large enough to make the agents' guesses about the values of  $c$ ,  $b$ , and  $p$  exceed the ranges of the possible values of these properties; in such cases, a new guess is repeatedly made until the agent's guess lies inside the range of possible values (problems with this method of generating noise are discussed below). The error affecting the agents' knowledge of  $p$  varies independently of the error affecting the agents' knowledge of  $c$  and  $b$ .

At each time step, every agent is presented with a new opportunity, and must decide whether or not to 'play' that opportunity or not. This decision is made according to the agent's decision rule. There are three types of agent, each with a different decision rule:

1. The RATIONAL agent uses an unbiased calculation of expected utility to decide whether to play; i.e. it only plays when the expected utility of playing is greater than that of not playing, or when:  $pb > (1-p)c$ . Since the agents do not have perfect information about  $p$ ,  $b$  and  $c$  (except when error levels are set to zero), a more accurate description of their decision rule would be:

$$p_e b_e > (1-p_e)c_e \rightarrow \text{play}$$

where  $p_e$  = the agent's estimate of  $p$ .

2. The OPTIMISTIC agent also uses the principle of maximum expected utility, but uses a biased estimate of  $p$  (its estimate of  $p$  is multiplied by its estimate of  $b$  divided by its estimate of  $c$ ). This means that the agent has an

unrealistically high estimate of  $p$  whenever the benefits exceed the costs, and an unrealistically low estimate of  $p$  where the potential costs outweigh the benefits. More formally:

$$b (p_e (b_e/c_e)) > (1 - p_e (b_e/c_e)) c_e \rightarrow \text{play}$$

3. The EMOTIONAL agent always plays if it estimates  $b$  to be more than twice  $c$ , and never plays when it estimates  $b$  to be less than half  $c$ . When it estimates that  $b$  and  $c$  are between these limits, its chance of playing is proportional to its estimate of  $p$ . More formally:

$$c_e/b_e > 2 \vee (0.5 \leq c_e/b_e \leq 2 \ \& \ \text{random} \{0 - 1\} < p_e) \rightarrow \text{play}$$

If an agent decides to play an opportunity, its chance of success is determined by the probability of success associated with that opportunity; that is, a new random number between 0 and 1 is drawn, and the agent succeeds if this random number is lower than or equal to the true value of  $p$  for the opportunity the agent is currently playing. If the agent plays and succeeds, its energy level is increased by the benefit of success associated with the opportunity. If it plays and fails, its energy level is decreased by the cost of failure associated with that opportunity. If an agent does not play, its energy level remains the same for that turn. Agents start with zero energy. Agents never die, and may have negative energy levels. There is no reproduction and hence no evolution.

## Results

We let the program run 10 times for every possible combination of the error of  $p$  and the error of  $c$  and  $b$ . One run lasts for 500 time steps. For each run we recorded the final average energy of each type of agent. We used this data to calculate the means and 95% confidence interval for all those runs for each kind of agent.

The different classes of agent performed significantly differently from each other, and responded differently to changes in error levels. Figure 1 shows final energy levels for the three types of agents where the error in the assessment of costs and benefits was kept low at 2, whilst the error in assessment of probabilities is allowed to vary. Where error in the assessment of probabilities is low, the rational agents outperform the other two types. As error in the assessment of probabilities increases, all types of agents have lower energy, since, as this is a more uncertain world, they all make more bad decisions. However, the optimistic and emotional agents are less severely affected than the rational ones, and there is a cross-over level of uncertainty where the rational agents cease to be the best performing ones. The optimistic agents perform best until another cross-over point where uncertainty is still higher, where the emotional agents become the best performers.

If error in the assessment of costs and benefits is pegged a little higher, at 4, the crossover zone where rational agents are outperformed by biased ones is smaller (figure 2). Where the error in the assessment of costs and benefits is 6 or any higher value, the cross-over does not occur at all, and the rational agents are as good as or

better than the other types over the whole range of error in the assessment of probabilities (figure 3).

Graph of the data with ErrorCB = 2

Shown are the means and error bars using 95% Confidence Interval.

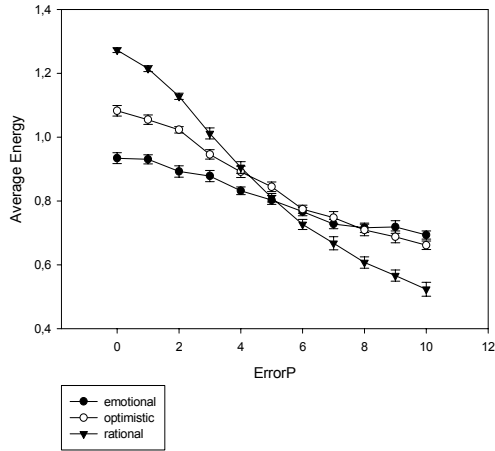


Figure 1. Average final energy levels (mean and 95% confidence intervals) for the three types of agents over ten runs, with errorcb constant at 2, and errorp allowed to vary (horizontal axis).

Graph of the data with ErrorCB = 4

Shown are the means and error bars using 95% Confidence Interval.

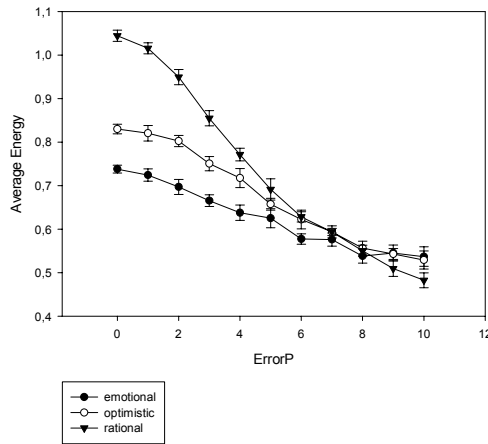


Figure 2. Average final energy levels (mean and 95% confidence intervals) for the three types of agents over ten runs, with errorcb constant at 4, and errorp allowed to vary (horizontal axis).

Graph of the data with ErrorCB = 6

Shown are the means and error bars using 95% Confidence Interval.

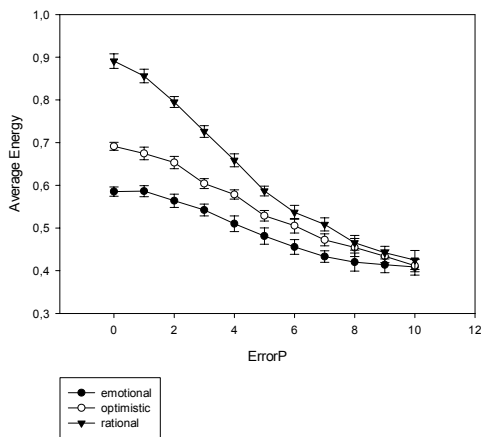


Figure 3. Average final energy levels (mean and 95% confidence intervals) for the three types of agents over ten runs, with errorcb constant at 6, and errorp allowed to vary (horizontal axis).

3D Graph of the three different types of agents

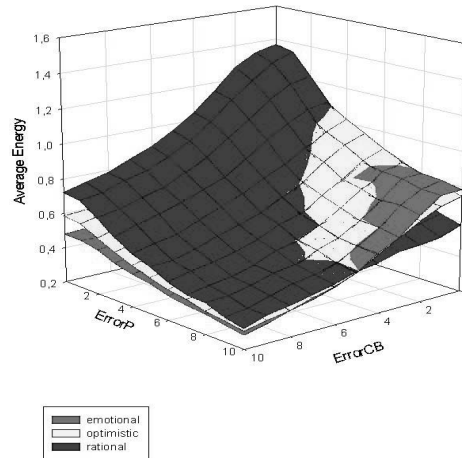


Figure 4. Three dimensional representation of the average final energy levels of the three types of agents over ten runs, for all possible combinations of errobc and errorp.

The results are summed up in figure 4. The overall slope of all the surfaces is downward; as uncertainty of any type increases, the average energy levels of all agents are reduced as they make more bad decisions. When both types of uncertainty are very high, overall energy levels are low, and there is no significant difference between the different decision rules.

The rational agents perform better than the other two types where uncertainty is low and over most of the rest of the surface. However, where there is fairly high uncertainty in the assessment of probabilities, but fairly low uncertainty in the assessment of costs and benefits, then the two biased decision rules significantly outperform the unbiased one.

## **Discussion**

Our results show that an agent using a biased decision rule can outperform an unbiased one under a specific set of circumstances: where the degree of uncertainty in estimating the probability of obtaining a favourable outcome is moderately high, but it is possible to get a fairly accurate picture of the relative benefits of success and costs of failure.

It seems plausible that these conditions – good information about costs and benefits, poor information about likelihood – do characterise the situation in which many human judgements are made. The costs and benefits arising from outcomes are relatively recurrent; that is, much the same values obtain each time a similar situation arises. Thus, individuals can learn by experience or observation of others roughly what the costs and benefits of a given scenario will be. Costs and benefits that have recurred very frequently through human evolution may have even become innate motivators, such as the intrinsic motivational response to an attractive potential mate or the opportunity to gain status. The probability of being successful in a particular endeavour is, however, not recurrent. Other people may or may not be competing for the same resource, and the prospects are highly contingent on specific features of the time and context.

The optimistic agent in our model is based on empirical observations of people's actual judgements; it overestimates its chances of success when the action would be relatively beneficial to the self, and underestimates its chances of success where the action would be potentially costly to the self (Larwood and Whitaker 1977, Miller and Ross 1975, Alicke 1985, Campbell 1986, Lewinsohn et al. 1980, Mirels 1980, Weinstein 1980, Taylor and Brown 1988). Our results suggest that this type of bias is indeed adaptive under the conditions described above, and so support the contention that motivational biases are design features rather than design flaws of the human mind (Nettle in press, Hasleton and Buss in press).

The optimistic agents in the model are biased in their assessment of the probability of success, whereas the emotional agents pay no attention to probabilities at all when they estimate that the cost-benefit ratio is smaller than 0.5 or greater than 2. There is some evidence that people often ignore probabilities, or at least weight changes in probability less heavily than they should (Kahneman and Tversky 1984).

Rottenstreich and Hsee (2001) suggest that people are generally less sensitive to changes in probability when the prize is affect-rich than affect-poor. In our terms, an affect-rich prize would be one with a high value of  $b$  relative to  $c$  (or vice versa, for an affect-rich punishment). Lotteries are a case in point. Many people will buy lottery tickets despite freely available evidence that the probability of winning is tiny, as long as the headline prize is huge relative to the cost of a ticket. The emotional agents mimic this behaviour, in ignoring probabilities altogether as long as the potential rewards or harms are estimated to be sufficiently large. (The cut off points of  $c/b < 0.5$  or  $> 2$  are arbitrary, but the objective is to capture the logic of a naturalistic decision rule, not provide an exact psychological model). The results suggest that this can also be an adaptive strategy where the level of uncertainty in judgement of probabilities is high.

Our 'rational' agents are directly based on the standard procedures in classical decision theory for maximising expected utility (Von Neuman and Morgenstern, 1944). Since such principles are generally regarded as the 'gold standard' of rational decision-making, it is interesting that our two biased agents can outperform the unbiased 'rational' agent under certain circumstances. The results appear to refute the claim that classical decision theory provides a model of optimal behaviour under all circumstances. In particular, any decision rule that gives more weight to  $b$  and  $c$ , and less weight to  $p$ , than the classical formulation of the principle of maximum expected utility, will be less affected by conditions of asymmetrical noise – more precisely, by conditions in which there are high levels of uncertainty associated with  $p$  but low levels of uncertainty associated with  $b$  and  $c$ . This is why both the optimistic and emotional agents do better than the 'rational' agent under such circumstances; they both give more weight to  $b$  and  $c$ , and less weight to  $p$ , than the rational agent (which implements a simple form of the principle of maximum expected utility).

In the limit, where the noise affecting  $p$  is so great as to make estimates of  $p$  uncorrelated with the true value, the sensible thing to do is to ignore  $p$  altogether and just assume that  $p$  is 0.5. In that case an agent would play if and only if  $b_e > c_e$ . What the optimistic and emotional agents do somewhat resembles that – at least, it resembles that closely enough to allow them to do better than the 'rational' agent under conditions of high noise affecting  $p$ , while in these conditions the rational agents are just blindly believing a worthless estimate of  $p$ . The classical decision theorist may point out that this is not, in fact, very rational at all. Indeed, he might add that, even with limited noise, the 'rational' agents are not really rational, because they do not take the noise into account. In effect, they operate under the false assumption that their estimate of  $p$  is always perfectly accurate (i.e. that  $p_e = p$ ). A more accurate assessment of our findings, then, would emphasise that our results apply *only when the agents are unaware of noise*.

This brings us to the important question of how noise is generated in our model. The level of error affecting the agents' knowledge of  $p$ ,  $b$  and  $c$  can range from 0 (perfect information) to 10 (great uncertainty). This error level determines the standard deviation used for a normal distribution of which the mean is the true value of  $c$ ,  $b$  or  $p$  of the opportunity. A random number drawn from this distribution determines the agents' guesses about the values of  $c$ ,  $b$  and  $p$ . Problems arise, however, when the noise is high enough to lead the agents to produce, at least on some occasions, estimates that lie outside the range of possible true values (such as estimates of  $p$

that are less than zero or greater than one). In such cases, a new guess is repeatedly made until the agent's guess lies inside the range of possible values. This, however, means that in such situations the agents' estimates are not normally distributed around the true value. The actual distribution is a normal distribution with one or two pieces missing – the parts that extend outside the range of true values. But unless the true value is exactly in the middle of the range of possible values (eg.  $p = 0.5$ ), the removed pieces will not be equal in size, which means the distribution is skewed. For example, if the true value of  $p$  is greater than 0.5, and there is enough noise to generate estimates of  $p$  that are greater than 1, then the agents will tend to underestimate  $p$  (so that  $p_e < p$ ). Similarly, if the true value of  $p$  is less than 0.5, the agents will tend to overestimate  $p$  (so that  $p_e > p$ ).

The rational course of action for an agent that knows about this skew would be to account for it by adjusting its estimate of  $p$  a little bit in the direction away from 0.5 before deciding whether to play. As it happens, this is more or less how the optimistic agents behave. They do not *always* skew their estimates of  $p$  in the direction away from 0.5, but they have a tendency to do so in the cases when it matters, and this provides them with enough of an advantage to enable them to beat the rational agents when the noise affecting  $p$  is high.

In order to make sure that our results were not an artefact of the way we implemented noise and uncertainty, we tried out a variety of different ways of ensuring that the agents' estimates of  $p$ ,  $b$  and  $c$  lie inside the range of possible values when noise is high. For example, instead of making a new guess repeatedly (re-sampling), we used a simple cutoff procedure in which estimates that lay outside the range of possible true values were re-set to the maximum or minimum possible value. This produces a rather odd distribution, but when we ran the model with this error function, we replicated our original results. We also tried using different error distributions (such as a uniform distribution of error rather than a normal distribution) and this too yielded essentially the same results. The fact that we were able to replicate our results with a variety of different error procedures makes it more likely that our results are not merely an artefact of programming.

In this program we did not investigate the possibility of giving different computational costs to different decision rules. This would be interesting, since then we could then introduce an evaluation function that favoured cheaper rules. The reason we did not build such an evaluation function into our program is that we do not have any empirical data about the computational costs associated with different decision rules, at least insofar as they are implemented in the human brain, and all of the algorithms used here are relatively simple. It would, however, be possible to calculate the computational cost of implementing the various decision rules in a robot with a given level of computational resources. This could be done by measuring how long it takes the robot to arrive at a decision on the basis of one rule, and comparing that to the time taken to decide on the basis of other rules. It is plausible to think that such measurements would allow a cheaper rule, that performs a little worse than the more expensive rule, to score better overall (Gigerenzer and Todd, 1999).

## Conclusion

From the standpoint of classical decision theory, motivational biases appear irrational. The computer model presented in this paper shows that, under certain environmental conditions, unbiased agents are outperformed by agents who are biased in similar ways to humans. Moreover, it is plausible that these environmental conditions are similar to those that humans encounter. Our findings therefore add support to the view that motivational biases are design features rather than design flaws of the human mind.<sup>1</sup>

## Acknowledgements

Ken Binmore, Nick Gotts, Verena Hafner, and Seth Tisue provided useful feedback on earlier versions of this paper. Dylan Evans' work was supported by a Platform Grant from the Engineering and Physical Sciences Research Council, EPSRC GR/M97503.

[4233 words]

## References:

- Alicke, M.D. (1985). Global self-evaluation as defined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, **49**, 1621-30.
- Alloy, L. B. and Abramson, L. Y. (1979). Judgement of contingency in depressed and non-depressed subjects: Sadder but wiser? *Journal of Experimental Psychology: General*, **108**, 443-479.
- Campbell, J.D. (1986). Similarity and uniqueness: The effects of attribute type, relevance and individual differences in self-esteem and depression. *Journal of Personality and Social Psychology*, **50**, 281-94.
- Gigerenzer, G. and Todd, P. M. (1999). *Simple Heuristics that Make us Smart*. Oxford: Oxford University Press.
- Hasleton, M. G. and Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, **78**, 81-91.
- Hasleton, M. G. and Buss, D. M. (in press). Biases in social judgement: Design flaws or design features? In J. Forgas, K. Williams and B. Von Hippel (Eds.),

---

<sup>1</sup> The NetLogo code for our model can be downloaded by visiting <http://www.dylan.org.uk/OptimismAISB.html>, where the model can also be run as an applet in a web browser.

*Responding to the Social World: Implicit and Explicit Processes in Social Judgments and Decisions.*

- Kahneman, D., Slovic, P. and Tversky, A. (eds) (1982) *Judgement under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.
- Kahneman, D. and Tversky, A. (1973) On the psychology of prediction. *Psychological Review*, **80**, 237-251.
- Kahneman, D. and Tversky, A. (1984) Choices, values and frames. *American Psychologist*, **39**, 341-350.
- Larwood, L and Whitaker, W. (1977) Managerial Myopia: Self-serving biases in organisational planning. *Journal of Applied Psychology*, **62**, 194-198.
- Lewinsohn, P.M., Mischel, W., Chaplin, W. and Barton, R. (1980). Social competence and depression: The role of illusory self-perceptions. *Journal of Abnormal Psychology*, **89**, 203-12.
- Miller, D. T. and Ross, M. (1975) Self-serving biases in attribution of causality: Fact or fiction? *Psychological Bulletin*, **82**, 213-225.
- Mirels, H. L. (1980) The avowal of responsibility for good and bad outcomes: The effects of generalized self-serving biases. *Personality and Social Psychology Bulletin*, **6**, 299-306.
- Nettle, D. (in press) Adaptive illusions: optimism, control and human rationality. In *Emotion, Evolution and Rationality*, eds. Dylan Evans and Pierre Cruse, Oxford: Oxford University Press.
- Von Neuman, J. and Morgenstern, O. (1944) *Theory of Games and Economic Behaviour*. Princeton: Princeton University Press.
- Rottenstreich, Y. and Hsee, K. (2001) Money, kisses and electric shocks: On the affective psychology of risk. *Psychological Science*, **12**, 185-190.
- Rudski, J.M. (2000). Illusion of control relative to chance outcomes. *Psychological Reports*, **87**, 85-92.
- Taylor, S. E. and Brown, J. D. (1988) Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin*, **103**, 193-201.
- Vazquez, C. (1987). Judgement of contingency: Cognitive biases in depressed and non-depressed subjects. *Journal of Personality and Social Psychology*, **52**, 419-431.
- Weinstein, N. D. (1980) Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, **39**, 806-820.

Weinstein, N. D. (1982) Unrealistic optimism about susceptibility to health problems. *Journal of Behavioural Medicine*, **5**, 441-460.

Wilensky, U. (1999). NetLogo. <http://ccl.northwestern.edu/netlogo/>. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

Zuckerman, M. (1979) Attribution of success and failure revisited: The motivational bias is alive and well in attribution theory. *Journal of Personality*, **47**, 245-287