Decision-making and Confidence Given Uncertain Advice: Part One

Matthew Dry
Department of Psychology, University of Adelaide
Adelaide, SA 5005 Australia.

Michael Lee
Department of Psychology, University of Adelaide
Adelaide, SA 5005 Australia.

Abstract

Providing advice to decision makers in real world environments is difficult because data can be both inaccurate and uncertain, leading to an erosion of trust. This paper explores how decision maker advice acceptance can be maximized by manipulating the way in which uncertain information is expressed. An experiment was performed comparing the use of 'Avatars' (computer-generated animated faces) and plain text displays to convey information to decision makers. Decision-makers were required to make a series of binary choices about the location of a target stimulus on the basis of advice conveyed by an automated system. Advice accuracy and uncertainty was manipulated across six conditions, ranging from always providing advice (even if the correct answer was not known with any certainty) to only rarely providing advice (when the correct answer was known with near-certainty). We examine decision-making behavior and subjective confidence to assess which of these conditions best maintains trust in the advice.

Introduction

Real world decision making often takes place in environments that are both dynamic and unbounded (Gigerenzer & Todd, 1999). Providing advice to decision makers in these environments is difficult because data can be both inaccurate and uncertain, leading to an erosion of trust. This paper explores the relationship between trust and advice ambiguity in an environment in which accuracy of advice and degree of uncertainty are positively correlated (i.e., in an environment where an advisor who rarely expresses an opinion is very likely to be correct on those occasions when they do, whereas an advisor who always gives an opinion is less likely to be correct on average).

Accuracy, Uncertainty and Trust

The experimental task required decision-makers to collect treasure located behind one of two closed doors. They were asked to make a series of binary choices (to open the left or right door) on the basis of advice conveyed by an automated system. The advice was derived from a known probability relating to the location of the treasure. On the basis of this probability the system could advise the decision-maker to choose either the left or right door. Alternatively, if (according to a set criterion level) the probability was too ambiguous (i.e., too close to 0.5) to advise one course of action over the other, the system could make a default response in which no advice was given.

The system criterion levels for determining the form of the system response will necessarily have a direct bearing upon the relative accuracy and uncertainty of the response. A strict criterion will produce a high number of default responses in which no advice will be given. However, when advice is given it is likely to be accurate. Alternatively, a lax criterion produces few default responses, but the overall accuracy of the advice is much lower.

If trust is determined primarily by accuracy of advice, then trust levels should be high when the system is set to a strict criterion and low when the system employs a lax criterion. Similarly, it follows that the proportion of trials in which the decision-makers follow the system advice should be higher under a strict criterion than a lax criterion.

However, research has suggested that advisor caution plays a significant role in influencing decision-maker expectations and understandings. For example, it has been demonstrated that decision-makers are less likely either to trust or follow advice from conservative advisors (Sniezek & Van Swol, 2001), and that cautious advisors are rated as less accurate than over-confident advisors, regardless of the actual accuracy of the advice (Price & Stone, 2004). This research suggests that decision-makers will have higher trust and be more likely to follow advice when the system employs a lax criterion.

Display Modality and Trust

There has been recent interest in the development and use of realistic anthropomorphized interface agents or 'avatars' to convey information to decision-makers (e.g., Dehn & van Mulken, 2000). One way that avatars may differ from conventional text displays is in the ability to evoke affective responses such as confidence or trust. The tendency for people to treat computers as if they were social actors has been well documented (Nass & Moon, 2000; Picard, 1997; Waern & Hook, 2000), and a common finding in the literature is that the more computer interfaces present characteristics associated with humans, the more likely they are to elicit social responses.

Given this, it was expected that display modality would have an effect upon decision-maker trust levels and responses. In other words, it was expected that decision-makers would express a higher level of trust for, and be more likely to follow advice conveyed by avatar displays compared to conventional text displays.

Method

Participants

A total of 60 participants (28 males, 32 females), with a mean age of 25 years completed the experiment.

Stimuli

Information was presented to the participants via an avatar display (avatar condition), or a plain text display (text condition). For the avatar condition a "neutral trust" face, based upon the outcomes of previous research, was used. Figure 1 is an example of the avatar display. Advice in the avatar condition was conveyed to the decision-maker via headphones. The audio message and the movement of the avatar face were synchronized in order to simulate naturalistic, human-like behavior.

The three messages that an avatar (or its textual equivalent) could convey were: "Go Left", "Go Right", or "I Have No Idea". The criterion bounds on the underlying probability that led to each of these three alternatives were manipulated across six conditions: 0.5, 0.4 to 0.6, 0.3 to 0.7, 0.2 to 0.8, 0.1 to 0.9, and 0.0 to 1.0.

For each trial two values (N_1, N_2) were randomly drawn from a uniform distribution ranging on [0,1]. N_1 determined the advice that was given to the decision-maker, and N2 determined the position of the target door. If N_1 was less than or equal to the lower bounds of the criterion the system advised "Go Left", if N_1 was greater than or equal to the upper bounds of the criterion the system advised "Go Right". Alternatively, if N_1 fell between the criterion bounds the system returned the default response "I Have No Idea". For the "Go Right" trials if $N_1 > N_2$ the advice was correct. For the "Go Left" trials if $N_1 < N_2$ the advice was correct. Table 1 summarizes the probability of advice being correct, proportion of no advice trials and potential proportion of correct responses across the six criterion levels.

Procedure

The subjects attended a single test session lasting around 45 minutes. The subjects were required to make a binary decision (to go through one of two doors) based upon the advice presented to them. Once the subject had made their decision, they were asked to indicate how confident they were in their decision on a 5-point scale ranging from Guess (1) to Sure (5). The subjects took part in all six advice criterion conditions in a randomly chosen order. Each criterion condition was comprised of 50 separate trials.

Table 1. Probability of advice being correct, proportion of advice trials and potential proportion of correct responses across the six criterion conditions.

	Criterion Condition					
	1	2	3	4	5	6
Lower bound	0.5	0.4	0.3	0.2	0.1	0.0
Upper bound	0.5	0.6	0.7	0.8	0.9	1.0
p. advice trials	1.00	0.80	0.60	0.40	0.20	0.00
p. advice correct	0.75	0.84	0.91	0.96	0.99	-
Potential p. correct	0.75	0.77	0.75	0.68	0.60	0.50

Half of the subjects took part in the avatar condition, with the remaining half completing the text condition.

Results

The empirical confidence data are summarized in Figure 2. It can be seen that regardless of criterion condition the mean confidence for advice trials is well above that for the no advice trials. Furthermore, contrary to expectation, for both the advice and no advice trials mean confidence for subjects in the text condition is greater than that for subjects in the face condition.

Interestingly, although subjects appeared to be more trusting of advice that was displayed textually, Figure 3 indicates that the subjects were more likely to follow advice from an avatar. Although the error-bars indicate a high degree of overlap, subjects in the face condition consistently follow advice more frequently than subjects in the text condition, regardless of criterion condition.

Bayesian hypothesis testing

Largely following the approach adopted by Vickers, Lee, Dry and Hughes (2003) Bayesian statistical inference (Kass & Raftery, 1995) was employed to determine the nature of the relationship between criterion condition, display modality and both decision-maker trust and advice-

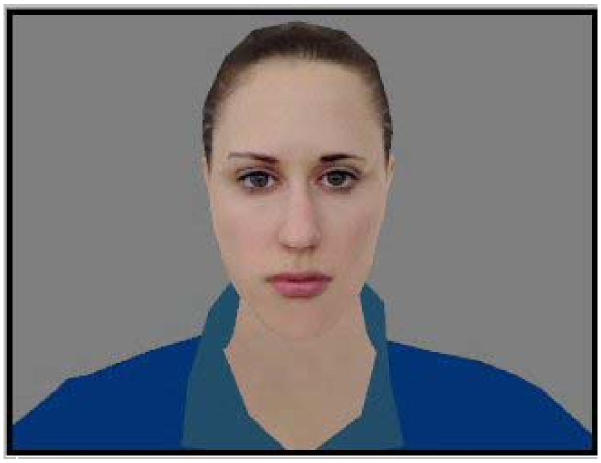

Figure 1: Example of avatar display employed in the experimental task.

following behaviour. The analyses compared twelve competing models that made different assumptions about the effects of the conditions and their interactions.

Models 1 to 6 made the assumption that there would be no difference between the two display conditions, in which case the data can best be described by a single line. Additionally, models 1 to 6 made varying assumptions concerning the relationship between the dependant variable (either mean confidence or mean proportion of trials in which advice is taken) and criterion condition. Model 1 assumed that the dependant variable would increase linearly across the 5 criterion conditions in which advice was given. Models 2 to 4 assumed that the dependant variable would peak at criterion conditions 4, 3 and 2 respectively. Model 5 assumed that the dependant variable would decrease linearly across the 5 criterion conditions, and Model 6 assumed that the dependant variable would remain constant across criterion conditions.

Each of the models 1 to 6 had a corresponding model (models 7 to 12 respectively) that made the same assumptions relating to criterion condition, but also had the additional assumption that there would be a difference between the two display conditions. In this case the models were fit by two separate lines, one corresponding to subjects in the avatar condition and one to subjects in the text condition.

Confidence

Figure 4 shows the maximum likelihood fits to the empirical confidence data for the advice trials under each of the twelve models, assuming a Gaussian likelihood function. Given these data fits and the known parametric complexity of the models, it is possible to calculate the Bayesian Information Criterion (BIC) for each model (Schwarz, 1978). The relative likelihood of each model can then be determined by calculating Bayes Factors (Kass & Raftery, 1995). Table 2 summarizes the results of the analyses, showing the maximum likelihood fit of the series predicted by each model, the number of model parameters, and the BIC and Bayes factors.

Before the results of the analyses are discussed it is necessary to explain the graphical similarity between models 1, 5 and 6 in terms of the line of best fit. As described above, Model 1 forces a positive linear function through the data points, and Model 5 a negative linear function. However, Figure 4 demonstrates that the optimal fit in both of these cases has a slope of close to zero, and is virtually indistinguishable from the best fit of Model 6.

The Bayesian analyses indicate that the data provide the most evidence for Model 12. This suggests that mean confidence ratings are higher in the text condition than in the avatar condition, but there is no difference between the mean confidence ratings across of criterion conditions 1 to 5. However, Models 6 and 2 are only 1.31 and 1.91 times less likely Model 1, and according to Jeffreys' (1961) guidelines for interpreting Bayes factors a difference of less than 3.2 is 'not worth more than a bare mention.' In other words, the data are only slightly less likely to be best fit by a model that makes the assumption that there is no difference between the display conditions, and that confidence is either stable across the five criterion conditions (Model 6), or increases linearly across conditions 1 to 4 then falls away (Model 2).

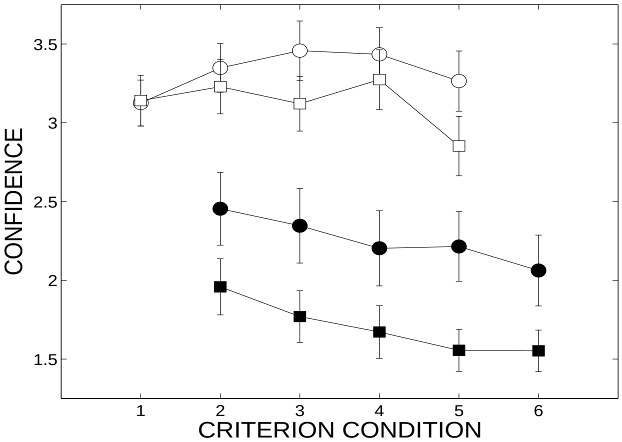

Figure 2: Summary of mean participant confidence for advice trials (white markers) and no advice trials (black markers). Circles denote text condition, squares denote avatar condition. Error bars represent one standard error.

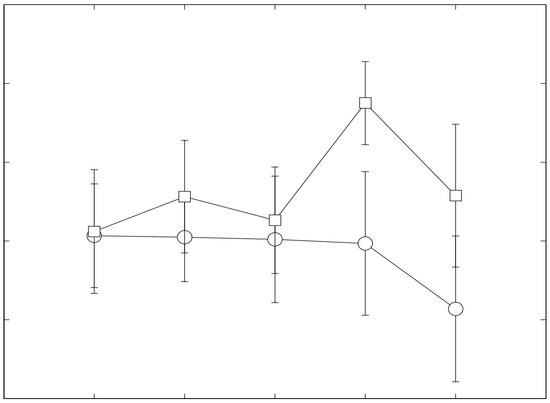

Figure 3: Summary of empirical data for mean proportion of advice taken. Circles denote text condition, squares denote avatar condition. The x-axis indicates the five criterion conditions in which advice ("Go Left/Right") was given to the subjects. Error bars represent one standard error.

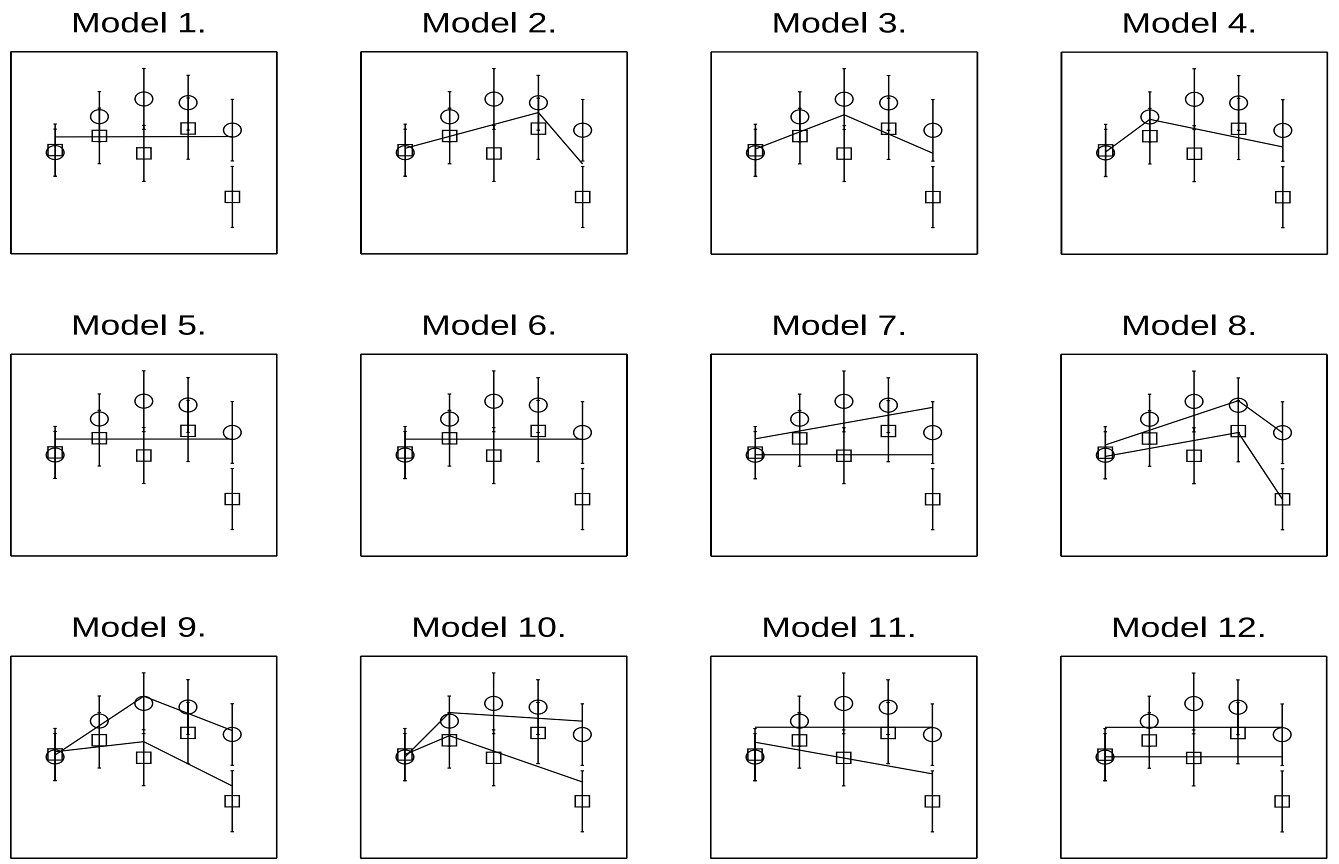

Figure 4: Summary of fit to empirical mean confidence data for the twelve models. Circles denote text condition, squares denote avatar condition. The solid line indicates the best fit to the empirical data under the constraints of the model. The x-axis indicates the five criterion conditions in which advice ("Go Left/Right") was given to the subjects. Error bars represent one standard error.

Proportion of advice taken

Figure 4 shows the maximum likelihood fits to the empirical proportion of advice taken data under each of the twelve models. Table 2 summarizes the results of the analyses, showing the maximum likelihood fit of the series predicted by each model, the number of model parameters, and the BIC and Bayes factors.

Once again the Bayesian analyses indicate the most likely model is Model 12, with Model 6 being only 2.52 times less likely. This indicates that there is little evidence to suggest any change in the proportion of advice followed across criterion conditions 1 to 5, but it is unequivocal as to whether there is any difference between the two display conditions.

Response pattern sub-groups

It is possible that the apparent lack of a relationship between criterion condition and the dependant variables is due to the process of averaging across distinct response patterns of sub-groups within the empirical data set. For

Table 2. Weighted sum-squared error data fit, parametric complexity, Bayes Information Criterion (BIC), and Bayes Factor for each of the twelve models to the empirical mean confidence data.

Model	WSSE	Parametric	BIC	Bayes
	Data Fit	Complexity	Value	Factor
1	8.80	2	13.41	4.14
2	4.95	3	11.86	1.91
3	6.52	3	13.43	4.18
4	6.79	3	13.70	4.80
5	8.80	2	13.41	4.14
6	8.80	1	11.10	1.31
7	5.10	4	14.31	6.50
8	1.23	6	15.04	9.39
9	1.82	6	15.63	12.61
10	2.31	6	16.12	16.12
11	5.18	4	14.39	6.79
12	5.96	2	10.56	1.00

Note- The Bayes factors are taken in relation to the most likely model which in this case is Model 12.

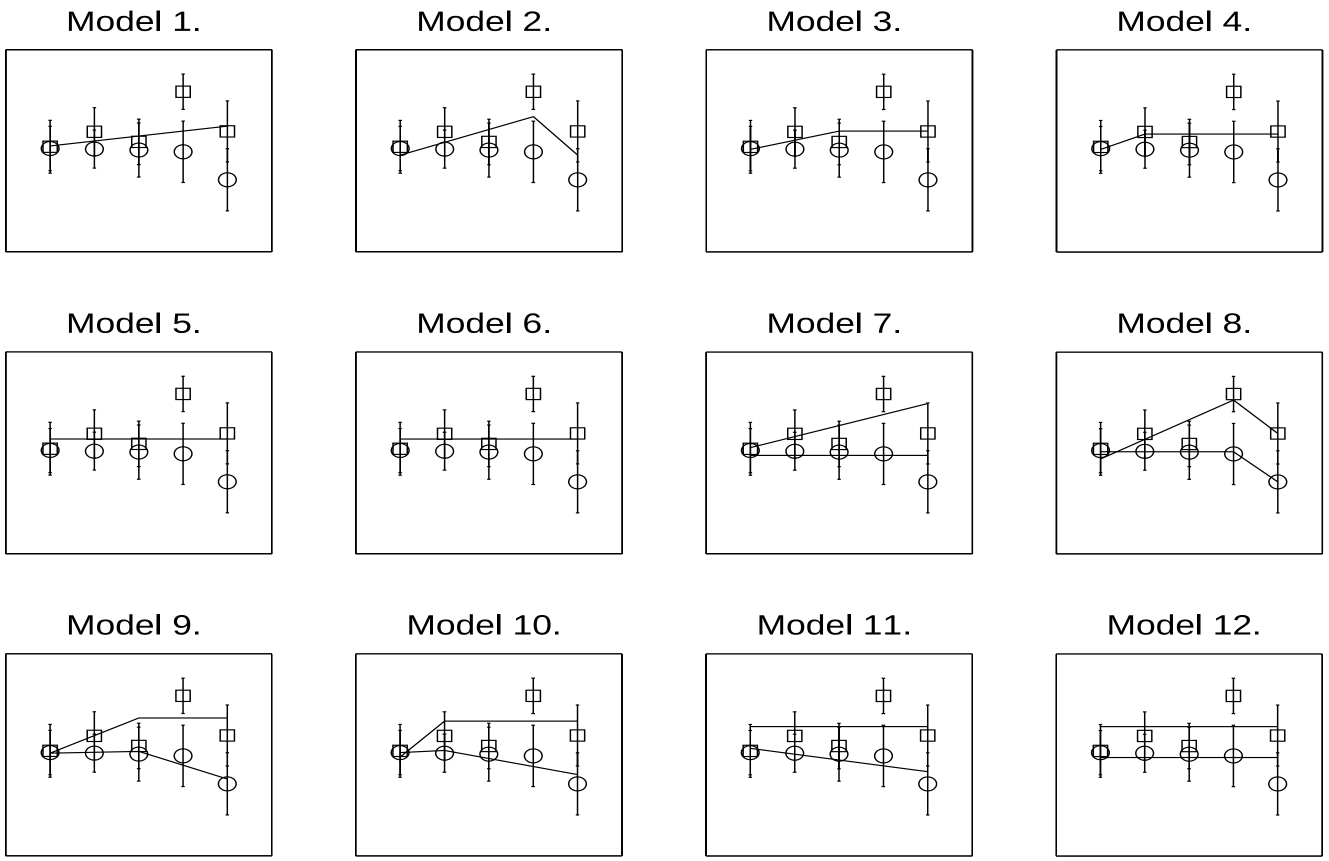

Figure 4: Summary of fit to empirical mean proportion of advice taken data for the twelve models. Circles denote text condition, squares denote avatar condition. The x-axis indicates the five criterion conditions in which advice ("Go Left/Right") was given to the subjects. Error bars represent one standard error.

example, if half of the decision-makers were following a response pattern analogous to Model 1, and half were following a response pattern analogous to Model 5, the resulting aggregated data would resemble a flat line.

In order to determine if the data set was comprised of two or more distinctive sub-groups Models 1 to 6 were fit to each participant's individual mean confidence and advice acceptance data. Using the data fits and parametric complexities of the models, the Bayesian Information Criterion (BIC) was calculated for each model. For each participant, the model with the minimum BIC value was selected as the most likely, and the results tabulated.

For a majority of the participants Model 6 provides the most likely account of both the confidence data (73% of participants) and response behaviour data (100% of participants). These results appear to rule out the possibility of sub-groups within the data set, and add further weight to the previous analyses finding no relationship between criterion condition and either confidence or proportion of advice trials followed.

Table 3. Weighted sum-squared error data fit, parametric complexity, Bayes Information Criterion (BIC), and Bayes Factor for each of the twelve models for empirical mean proportion of advice taken.

Model	WSSE Data Fit	Parametric Complexity	BIC Value	Bayes Factor
1	9.02	2	13.63	5.48
2	5.85	3	12.76	3.55
3	8.87	3	15.77	16.02
4	9.28	3	16.19	19.70
5	9.77	2	14.38	7.97
6	9.77	1	12.08	2.52
7	3.81	4	13.03	4.05
8	0.66	6	14.48	8.37
9	3.53	6	17.34	35.10
10	4.12	6	17.94	47.25
11	5.13	4	14.34	7.83
12	5.62	2	10.23	1.00

Note- The Bayes factors are taken in relation to the most likely model which in this case is Model 12.

Discussion

Display condition

Contrary to expectation the subjects were less trusting of the avatar display than the text display. However, the results also indicate that advice was followed more often in the avatar condition than in the text condition. The results of the Bayesian hypothesis testing are somewhat ambiguous in relation to the existence of any meaningful differences between the two conditions. For both the confidence data and the proportion of advice taken data a model assuming no differences between the display conditions was only slightly less likely than a model assuming two separate distributions for the avatar and text groups. Furthermore, if we accept that there is indeed a difference between the two groups the size of the effect is small for both mean confidence (d = 0.21) and mean proportion of advice taken (d = 0.20).

Despite this, it could be argued that if one of the primary aims of providing advice to decision makers is to ensure that the advice is actually followed then even a small effect should be considered advantageous. Furthermore, the lower mean confidence levels for the avatar display relative to the text display do not appear to have a negative impact upon proportion of advice followed. All things considered, in terms of relative utility, the avatar display is the superior modality for providing advice.

Criterion condition

The results of the Bayesian hypothesis testing indicate that there is no relationship between the criterion bounds set for system advice and either decision-maker trust or the likelihood of decision-makers following advice. This runs contrary to the hypothesis that decision-maker trust is based purely upon the accuracy of the advice that is being presented to them. Rather, as suggested in previous research (Price & Stone, 2004; Sniezek & Van Swol, 2001), it appears that uncertainty plays a role in determining decision-maker confidence and behavior.

The results of the present study are directly applicable to environments in which there is a correlation between accuracy and uncertainty. In particular they indicate that the optimal criterion for determining advice in a two-alternative forced choice situation is criterion condition 2, corresponding to lower and upper bounds of 0.4 and 0.6 respectively. Using this criterion the proportion of potentially correct responses is maximized (0.772), with no apparent negative consequences in relation to trust or likelihood of following advice. Furthermore, the results indicate that there is little or no variation in response behavior across subjects.

Unfortunately, due to the correlation between the accuracy of advice and the proportion of "I Have No Idea" trials it is difficult to determine the exact nature of the relationship between confidence, accuracy and uncertainty. To this end a future experiment is planned in which accuracy and uncertainty will be manipulated in a fully balanced design.

Acknowledgements

We acknowledge the research contributions of Glen Smith and Michael Paradowski, and financial support from the Defence Science Technology Organisation.

References

Dehn, D., & van Mulken, S. (2000). The impact of animated interface agents: a review of empirical research. *International Journal of Human-Computer Studies, 52*, 1-22.

Gigerenzer, G. & Todd, P. M. (Eds.). (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Jefferys, H. (1961). *Theory of Probability*. Oxford, U.K.: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773-796.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81-103.

Picard, R. W. (1997). *Affective Computing*. Massachusetts: MIT Press.

Price, P. C., & Stone, E. R. (2004). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making, 17*(1), 39-57.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes, 84*(2), 288-307.

Vickers, D., Lee, M. D., Dry, M. J., & Hughes, P. (2003). The roles of the convex hull and the number of potential intersections in performance on visually presented traveling salesperson problems. *Memory & Cognition, 31*(7), 1094-1104.

Waern, J. H., & Hook, K. (Eds.). (2000). *Interface agents: A new metaphor for human-computer interaction and its application to Universal Access*. Mahwah, NJ: Lawrence Erlbaum.