UNDER-RELIANCE ON THE DECISION AID: A DIFFERENCE IN CALIBRATION AND ATTRIBUTION BETWEEN SELF AND AID

Kees van Dongen[1] and Peter-Paul van Maanen[1,2]
[1] TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
Email: {kees.vandongen, peter-paul.vanmaanen}@tno.nl
[2] Department of Artificial Intelligence, Vrije Universiteit Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

It is often assumed that two heads are better than one, but reliance on decision aids is often inappropriate. Decisions to rely on an aid are thought to be based on a comparison between the perceived reliability of own performance and that of the decision aid. Unfortunately, perceived reliabilities are unlikely to be perfectly calibrated. This may result in inappropriate decisions to rely on advice. In a laboratory experiment with 40 participants, we studied whether calibration improves after practice, whether calibration of own reliability differs from calibration of the aid's reliability and whether unreliability of the aid is attributed differently. Under-trust in own reliability disappears after practice but under-trust in the aid's reliability persists. Unreliability of the decision aid is less likely to be attributed to temporary, external and uncontrollable factors. This asymmetry in attribution and calibration may explain under-reliance on decision aids.

INTRODUCTION

We take advice of others to improve our judgment and to share responsibility (Harvey & Fischer, 1997). We rely more on advice when we believe that the adviser is highly accurate, is credible, has more experience (Harvey & Fischer, 1997), or is confident (Sniezek & Van Swol, 2001). We rely more on advice when we are not confident or have less experience or knowledge. We also rely more on advice when it corresponds to our own opinion (Yaniv, 2004). These findings are consistent with theories on advice taking and control allocation that include the concept of relative trust (Moray, Inagaki, & Itoh, 2000; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Van Dongen & Van Maanen, 2005). Trust is defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability (Lee & See, 2004). Trust can refer to the advice of another agent or to one's own judgment. Trust, like the feelings and perceptions on which it is based, is a covert or psychological state that can be assessed through subjective ratings. To assess trust some studies have used scales of trust (e.g., Lee & Moray, 1992), some scales of perceived reliability (Wiegmann, Rich, & Zhang, 2001) or both (Madhavan, Wiegmann, & Lacson, 2003). We distinguish trust from the decision to depend on advice, the act of relying on advice, and the appropriateness of relying on advice (Van Maanen & Van Dongen, 2005).

It is often assumed that two heads are better than one, but reliance on decision aids is often inappropriate. A rational decision-maker would depend on the advice of a decision aid when this would increase the probability of goal achievement. At the same time he would depend on his own judgment when relying on advice would decrease goal achievement. When the decision aid's advice is actually more reliable than our own judgments, one should follow the aid's advice. However, the decision to depend on advice is not based on a comparison between actual reliabilities, but on how these are perceived and interpreted. Unfortunately, perceptions of the reliability of our own judgment or that of an aid may be imperfectly calibrated. Perceptions of reliability may be prone to systematic and random error. When the direction or magnitude of this error differs between self and aid this could lead to inappropriate reliance decisions: Under-reliance or over-reliance on the aid may be the result.

Calibration

Calibration refers to the correspondence between a person's perception of the reliability of an agent and the true reliability of that agent. Studies of judgment under uncertainty have indicated that humans are often overconfident about their own performance (Alba & Wesley, 2000). In most of these studies participants were required to answer general knowledge questions, often didn't receive direct feedback about their performance, and were asked to estimate confidence in their answers after each question. Although pervasive in the literature, overconfidence is not universal (Brenner, Koehler, Liberman, & Tversky, 1996). May's (1987, 1988) results for instance yielded 9% overconfidence when confidence was rated after each answer and 9% under-confidence when percentage correct was estimated after each block. A

confidence judgment for a particular question seems to depend on the balance of arguments for and against that specific judgment. Overconfidence may be reduced by helping humans to attend to contradictory evidence or alternative hypotheses (Van Dongen et al., 2005). Estimated percentage correct, on the other hand, is likely to be based on a general evaluation of the difficulty of the task or based on feedback about performance (Brenner et al., 1996). Imperfect calibration can also be caused by random error. If random error is the main problem, the focus might be on better feedback and training to reduce noise (Klayman et al., 1999). By substituting general knowledge questions with questions that require a prediction, less overconfidence and even under-confidence may be expected because of the inherent uncertainty that accompanies any future event (Vreugdenhil & Koele, 1988).

Concerning the perception of the aid's performance, Wiegmann et al. (2001) found that the reliability of decision aids is often underestimated. A reason for this may be that decision aids do not perform as expected. Dzindolet, Pierce, Beck, Dawe, and Anderson (2001) argue that the perception of the reliability of an automated aid is filtered through the operator's 'perfect automation schema', or the expectation that automation will perform at near perfect rates. This expectation may lead operators to pay too much attention to information that is in conflict with the schema; the errors made by automation, triggering a rapid decline in trust when decision aids make errors (Dzindolet, Pierce, Beck, & Dawe, 2002). With decision aids that are not perfectly reliable often a bias toward under-reliance is found (Parasuraman & Riley, 1997; Dzindolet et al., 1999).

Attribution

Lee and See (2004) argue that the dimensions purpose, performance and process provide a concise set of distinctions that describe the basis of trust in support systems across a wide range of application domains. Purpose describes *why* a decision aid was developed. Performance refers to *what* behaviors the aid shows and includes characteristics such as reliability and predictability. Process information describes *how* the aid operates. When the underlying reasoning process of a decision aid or the factors that affect that process are not observable, process concerns the qualities and characteristics that are attributed to an agent. This means that trust in one's own performance or that of an aid does not only depend on the reliability one observes, but also on how for instance error or unreliability is causally attributed (Falcone & Castelfranchi, 2004). Following 'causal attribution theory' success or failure can be ascribed to factors internal to the agent, such as knowledge, or to external factors, such as unreliable information, and either to temporary factors, or to permanent factors of the agent or of the environment. Causes of unreliability can be attributed to factors that are controllable or uncontrollable, like motivation or task difficulty, respectively. One common problem in assigning causes is called the correspondence bias. The correspondence bias or fundamental attribution error is the tendency of people to over-emphasize dispositional causes, i.e., permanent factors internal to an agent, for the errors of others, while under-emphasizing situational causes, i.e., temporary factors external to the agent. In contrast, our own errors are more likely to be attributed to temporary, external, and uncontrollable factors, like bad luck. Gilbert and Malone (1995) point out that if an observer is to have any hope of making a correct attributional analysis that takes into account the role of situational causes for the behavior of others, he or she must first perceive and recognize the situation in which the agent is functioning. The problem is that the situational factors that negatively affect the performance of a decision aid, for instance factors that affect the reliability of the data that is used for a recommendation, are often obscured. When theory on causal attribution holds in the context of human-machine interaction, we expect participants to attribute errors of the decision aid less to temporary, external and uncontrollable factors.

Hypotheses

In the present study participants were required to perform a prediction task with advice of a decision aid and were provided with feedback about their performance and that of the aid. We investigated 1) whether calibration of own perceived reliability differs from calibration of that of the aid, 2) whether calibration of own perceived reliability and that of the aid improves after practice and 3) whether own unreliability is attributed to different causes than unreliability of the decision aid. Based on the literature we expect underestimation of the reliability of the aid and, to a lesser degree, underestimation of own reliability. Perception of own reliability is expected to improve after practice. In light of the above research questions we tested three hypotheses:

H1: Underestimation of the decision aid is more prevalent than underestimation of the self.

H2: Underestimation of own reliability decreases after practice; that of the aid persists.

H3: Unreliability of the decision aid is less attributed to temporary, external and uncontrollable factors compared to own attribution of unreliability.

METHOD

Participants

Forty college students (16 female) participated in the experiment. Their ages ranged from 18 to 38 years (M = 23 yrs). Participants were paid 35 euro for their participation.

Tasks and Procedures

Participants read a story about a software company interested in evaluating the performance of their adaptive software before applying it to more complex tasks on naval ships. The story pointed out that the level of reliability between software and human performance was comparable and around 70%. Participants were asked to perform a prediction task with advice of a decision aid and were instructed to maximize the number of correct answers by relying on their own predictions as well as the advice of the decision aid.

In 40 training trials participants had to discover a pattern in the temporal order of correct answers (e.g., 1,1,2,3,2, etc.). Participants could use the discovered pattern for their predictions. The participants made their own independent prediction by clicking on one of the three buttons in the first row (Figure 1). Then the decision aid communicated its advice by highlighting one of the three buttons on the second row. Neither the data nor the algorithm on which this advice were based was made transparent to the participant. On the third row participants were asked to indicate the correct button based on either their own initial prediction or the advice of the decision aid. On the forth row feedback about the correct answer (highlighted button) and the success of the reliance decision (red or green color) was provided. The feedback allowed participants to adjust their mental model of the pattern and by comparing the correct answer with the responses on the first three rows participants were able to calibrate their perceptions of the reliability of their own predictions, the predictions of the decision aid, and their reliance decisions.

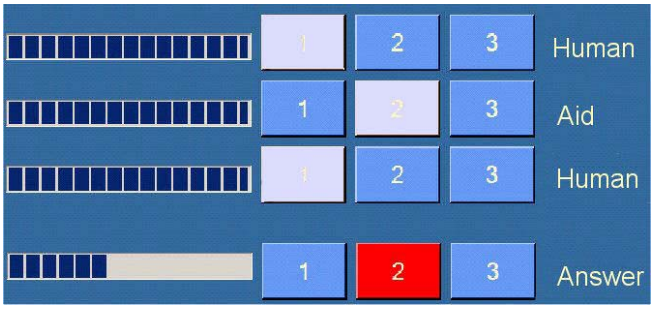

Figure 1. Prediction task

Design

Reliability of prediction performance of the aid was set to vary between 60 and 80% with an average of 70%. The source of error was not transparent. Average reliability of human performance was manipulated by setting a variation in the feedback such that the average of own prediction reliability would also be around 70%. Ten percent of the answers differed randomly from what would be expected by extrapolating the pattern in the history of feedback (e.g., 1,1,2,3,2, etc.). The timing of erroneous advice and unexpected feedback were both random, making the decision whom to rely on difficult. To be able to observe learning effects every participant performed two experimental blocks, each consisting of 100 trials.

During the two blocks the prediction reliability of the human and the prediction reliability of the aid were measured. Reliability was defined as the percentage correct. After each experimental block participants rated the perceived prediction reliability of themselves and that of the aid on a scale between 0 and 100%. Participants were asked to indicate the causes of unreliability of own prediction performance and that of the aid. On a scale between -3 and 3 participants rated the degree to which they disagreed or agreed with three statements about the causes of unreliability. Scores lower than zero on the “temporary factors”, “external factors” and “uncontrollable factors” question indicate that participants do not attribute unreliability to those factors.

RESULTS

Calibration of human reliability

Averaged over both blocks the mean *perceived* human prediction reliability was 4% lower than the mean *actual* human prediction reliability, $t(79) = -2.53$, $p < .01$. However, we observed both participants that underestimated their reliability (under the diagonal, Figure 2) and participants that overestimated their reliability (above the diagonal, Figure 2). The slope of the calibration curve was higher than ideal and perceived reliability was too close to either 0 or 100.

Learning effects. For the first block participants underestimated their performance with 5 %, $t(39) = -2.16$, $p < .05$. Underestimation was not statistically significant in the second block, $t(39) = -1.46$, $p > .05$. The correlations between perceived human reliability and actual human reliability increased from $r = .42$, $p < .05$ to $r = .51$, $p < .05$.

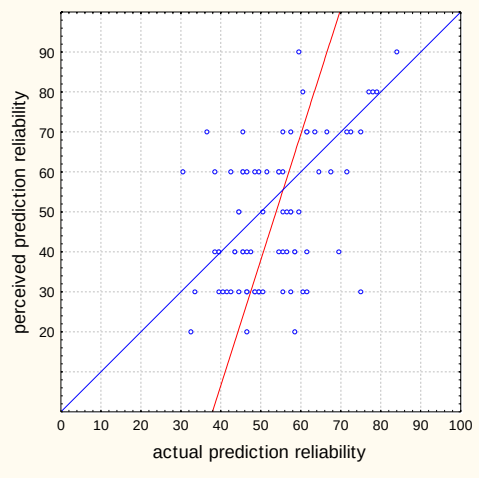

Figure 2. Calibration human reliability

Calibration of reliability of the aid

Averaged over both blocks the mean *perceived* computer reliability was 7% lower than the mean *actual* computer reliability, $t(79) = -4.41$, $p < .01$. Underestimation was more prevalent, but we observed both pessimists that over-weighed errors as well as optimists that under-weighed errors (Figure 3).

Learning effects. Participants underestimated the performance of the decision aid in the first block (7 %), $t(39) = -2.47$, $p < .05$, as well as in the second block (8 %), $t(39) = -3.79$, $p < .01$. The standard deviation of perceived reliability reduced from the first ($SD = 14.30$) to the second block ($SD = 13.53$).

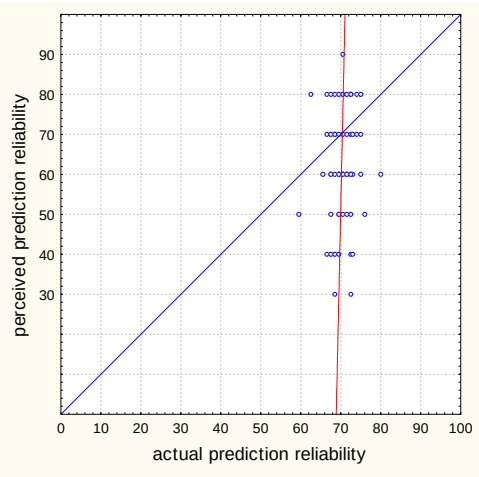

Figure 3. Calibration of reliability of the aid

Attribution

Compared to human unreliability ($M = .41$), unreliability of the decision aid is less attributed to temporary factors ($M = .05$), $t(79) = 2.02$, $p < .05$. The degree to which human unreliability ($M = -0.79$) and unreliability of the decision aid were attributed to external factors ($M = -1.09$) did not differ significantly, $t(77) = 1.66$, $p > .05$. Compared to human unreliability ($M = -0.26$), unreliability of the decision aid is less attributed to uncontrollable factors ($M = -0.85$), $t(79) = 2.92$, $p < .05$.

DISCUSSION

The above results confirm our first hypothesis (H1), namely that underestimation of the decision aid is more prevalent than underestimation of the self. Although under-trust was more prevalent, we observed both pessimists that seem to over-weigh errors, as well as optimists that under-weigh errors. This suggests that trust and reliance decisions are affected by personality traits.

Calibration is imperfect and overall people think reliability is worse than it actually is. Our results indicate that the direction of error in calibration between self and aid does not differ, but that the magnitude of underestimation does, especially after practice. As expected, under-trust in own performance seems to disappear, but under-trust in the performance of the decision aid persists (H2). When reliance decisions are based on a difference in perceived reliability, an asymmetry in the magnitude of error could lead to inappropriate reliance decisions (Dzindolet et al., 2003). In this case under-reliance on the aid is expected.

The persistent under-trust in the aid raises the question why the aid's errors receive more weight than the own errors. Our results show that, compared to the own unreliability, unreliability of the decision aid is interpreted as a more permanent attribute of the decision aid itself, rather than a temporary attribute of the situation in which it functions. Errors were also less attributed to uncontrollable factors. This confirms our third hypothesis (H3) and explains why people are less forgiving for errors of the aid. This also consistent with Gilbert and Malone's (1995) theory that people are less likely to attribute errors of another to situational factors when these are obscured. From this perspective the problem of under-trust and under-reliance on decision aids seems to be determined by a failure to understand or evaluate the situations that affect good and bad aid performance (Cohen, 2000).

Reliance behavior, however, is not only affected by differences in perceived reliability, but also by the difference in access to the underlying justification for one's own judgment and that of the aid (Yaniv, 2004). The human decision maker has access to the reasons supporting his or

her own prediction as well as to the strength of those reasons, but has no access to the reasons underlying the advice of a decision aid. An assumption of the cognitive approach is that the weight placed on a certain judgment depends on the evidence that could be recruited to support that judgment. Whether reliance behavior is affected by both differences in perceived reliability and differences in access to underlying justifications will be subject to further research. We expect under-trust and under-reliance to be reduced when both the sources of error and the justifications for advice can be made observable in decision aids.

ACKNOWLEDGMENTS

This research was funded by the Royal Dutch Navy under program number V206. We would like to thank José Kerstholt and Rick van der Kleij for their comments and suggestions.

REFERENCES

Alba, J. W., & Hutchinson, J. W. (2000). Knowledge calibration: What consumers know and what they think they know. *Journal of Consumer Research, 27*, 123–156.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes, 65*, 212–219.

Cohen, M. S. (2000). A situation-specific model of trust in decision aids. In *Proceedings of Human Performance, Situation Awareness & Automation: User-Centered Design for the New Millennium*, Savannah, GA, 143–148.

Dongen, K. van, & Maanen, P.-P. van (2005). Designing for Dynamic Task Allocation. In *Proceedings of the Seventh International Naturalistic Decision Making Conference (NDM7)*, Amsterdam, The Netherlands.

Dongen, K. van, Schraagen, J. M., Eikelboom, A., & Brake, G. te (2005). Supporting Decision Making by a Critical Thinking Tool. In *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica: CA. Human Factors and Ergonomics Society, 517–521.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (1999). Misuse and disuse of automated aids. In *Proceedings of the Human Factors Society 43rd Annual Meeting* Santa Monica, CA: Human Factors and Ergonomics Society, 339–343.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer* Studies, 58, 697–718.

Falcone, R., Castelfranchi, C. (2004). Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004)*, New York, USA, 740–747.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21–38.

Harvey, N. & Fischer, I. (1997). Taking advice: Accepting help, Improving judgment, and Sharing responsibility. *Organizational Behavior and Human Decision Processes, 70*, 117–133.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It Depends on How, What, and Who You Ask. *Organizational Behavior and Human Decision Processes, 79*, 216–247.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human machine systems. *Ergonomics, 22*, 671–691.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*, 50–80.

Maanen, P.-P. van, & Dongen, K. van (2005). Towards Task Allocation Decision Support by means of Cognitive Modeling of Trust. In *Proceedings of the Eighth International Workshop on Trust in Agent Societies (Trust 2005)*, Utrecht, The Netherlands, 168–77.

Madhavan, P., & Wiegmann, D. A. (2005). Effects of information source, pedigree, and reliability on operators' utilization of diagnostic advice. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, 487–491.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2003). Automation failures on tasks easily performed by operators undermines trust in automated aids. In *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica: CA. Human Factors and Ergonomics Society.

May, R. S. (1987). *Calibration of subjective probabilities: A cognitive analysis of inference processes in overconfidence* (in German). Frankfurt: Peter Lang.

May, R. S. (1988). Overconfidence in overconfidence. In A. Chaikan, J. Kindler, & I. Kiss (Eds*.), Proceedings of the 4th FUR Conference.* Dordrecht: Kluwer.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied, 6*, 44–58.

Parasuraman, R., & Riley, V.A. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*, 230–253.

Sniezek, J. A, & van Swol, L. M. (2001). Trust, confidence, and expertise in a judge–advisor system. *Organizational Behavior and Human Decision Processes, 84*, 288–307.

Vreugdenhil, H., & Koele, P. (1988). Underconfidence in predicting future events. *Bulletin of the Psychonomic Society, 26*, 236–237.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: the effects of aid reliability on user' trust and reliance. *Theoretical Issues in Ergonomics Science, 2*, 352–367.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93*, 1–13.