HUMAN REDUNDANCY AS SAFETY MEASURE IN AUTOMATION MONITORING

Dietrich Manzey[1], Karl Boehme[1] & Markus Schoebel[2]
[1]Technische Universitaet Berlin, Germany
[2]Universitaet Basel, Switzerland

The present study addresses effects of human redundancy on automation monitoring performance. Forty-six participants performed a multi-task, consisting of three sub-tasks which simulate basic demands of operators in a chemical plant. One of the tasks involved the monitoring of an automated process. Participants were randomly assigned to three groups: (1) "Non-Redundant": participants worked on all tasks alone. (2) "Redundant": participants were informed that a second crewmate would work in parallel on the monitoring task. (3) "Redundant-Feedback": like the "redundant" condition with the additional information that crewmembers' individual monitoring performance would be tracked and fed back. Results provide evidence of social loafing effects in monitoring performance. Participants in the "redundant" condition cross-checked the automation significantly less than participants in the other groups. Moreover they were more prone to miss automation failures which occurred surprisingly. The anticipation that individual performance will be tracked and fed back after the task in the "redundant-feedback" group reduced this effect. The results suggest that human redundancy does not necessarily constitute an effective measure for enhancing reliability of automation monitoring and that expected positive effects can at least partially be off-set by a sort of social-loafing effect.

INTRODUCTION

The use of redundancy is a common strategy to improve the fault-tolerance of technological systems. Its main purpose is to optimize systems' overall reliability and safety through, e.g., having different components working in parallel or implementing multiple independent component backups in order to substitute safety critical system functions in case of failures (e.g. provision of emergency generators in nuclear power plants, Felsenthal & Fuchs, 1976).

It has been suggested that the same basic principles may also be applied to enhance reliability of human performance in safety-critical or other risky tasks (e.g. Landau, 1969). More specifically, human redundancy for performing critical tasks has specifically been assumed to be one of the key elements of success of high-reliability organizations such as nuclear power plants, air traffic control networks, commercial aviation systems or chemical plants (Hofmann, Jacobs, & Landy, 1995; La Porte & Consolini, 1991; Roberts, 1990). With respect to such organizations, human redundancy particularly is conceived as a recovery factor to control human error (Swain & Guttmann, 1983). For example, doubling the number of people involved in monitoring safety critical systems is expected to enhance the probability that at least one of them will detect an error.

However, transferring the idea of technological redundancy to human work is not as straightforward as it seems (Conte & Jacobs, 1997; Felsenthal & Fuchs, 1976). In order to reduce the risk of system failures, redundant units need to work parallel to and completely independent of each other. Yet, individuals in redundant work systems are normally aware of each other and, as a

consequence, independence is not given (Sagan, 2004). One can assume that the knowledge that someone else is carrying out the same task can affect individual performance in ways that might counteract the benefits expected from human redundancy. This is suggested by the observation that individuals often tend to exert less effort if they work on tasks in group settings as compared to working alone. Such an effect has been well documented and is referred to as social loafing (Latané, Williams & Harkins, 1979; Karau & Williams, 1993). Accordingly, Sagan (2004) suggests that introducing human redundancy in complex technological facilities like nuclear power plants might even increase than lower safety issues.

These considerations are of particular significance for domains where human redundancy usually represents the safety measure of choice, e.g. supervisory control of automated systems. First evidence that human redundancy might not enhance or even impair reliability in automation monitoring has been provided by Mosier and colleagues (Mosier, Skitka, Dunbar & McDonnell, 2001; Skitka, Mosier & Rosenblatt , 2000) and Domeinski, Wagner, Schoebel & Manzey (2008). Mosier et al. (2001) investigated whether human redundancy in monitoring cockpit automation would lead to performance improvements in terms of less automation bias. Comparing two-pilot crews versus single pilot crews they did not find any difference in the number of omission and commission errors committed by the crews. Yet, it is not clear from their data whether this finding was related to effects of social loafing in the two-pilot crews or just the fact that both crew members were affected individually by the same automation induced bias in decision-making. More direct evidence for social loafing effects was provided by a study of

Domeinski et al. (2007). In this study participants were required to perform an automation monitoring task which was one of three tasks that had to be performed concurrently simulating basic demands of operators of a chemical plant. Their data showed that participants who worked redundantly reduced their individual monitoring effort significantly compared to conditions where single operators were alone responsible for the task or where operators worked redundantly but got the information that their crew mate would be a low-performer. However, the participants of this study had only less task practice and worked for only 12 minutes on the task which might have impacted the results. Furthermore, the data did not reveal whether the social loafing effects observed in the redundant work setting would have increased the risk to miss any automation failures.

The present experiment capitalizes on this earlier research. Using the same experimental paradigm as Domeinski et al. (2007) it was investigated (1) whether effects of social loafing could be replicated with participants who are better trained and work over a longer period on the task, (2) whether these effects are sufficiently strong to eventually raise safety risks by impairing the detection of single automation failures, and (3) whether implementing some sort of tracking of individual performance also in the redundant work setting would represent a countermeasure for negative consequences of human redundancy.

METHOD

Participants

A sample of 46 students (29 male, 17 female, age: 21-45 years) participated in the experiment. None of them had prior experiences with the task. Participation was compensated with Euro 10,- plus an option to win one of four additional prizes of Euro 50,- each, which were drawn among all participants.

Apparatus and Tasks

A PC-based laboratory multi-task environment was used for the experiment (Multi-Task Operator Performance Simulation, M-TOPS). It consists of three different tasks which were developed to simulate basic work demands of control room operators in a chemical plant. The user-interface is shown in figure 1.

Resource Ordering Task (ROT). This task is shown in the upper left quadrant of the interface. Basically it represents a mental arithmetic task. Participants are instructed that they always have to assure the availability of required chemicals in order to keep the chemical process running. For this purpose the actual and the set

value of an ascertained chemical is presented. Participants then have a maximal time-window of 15 seconds to calculate the difference, to type the result in the designed ordering field, and to initiate the order by clicking on an order icon. After an order has been sent a new task is presented after a fixed interval of three seconds. However, participants can speed up the sequence by actively initiating a new trial via mouse-click on the "arrow" icon (upper right).
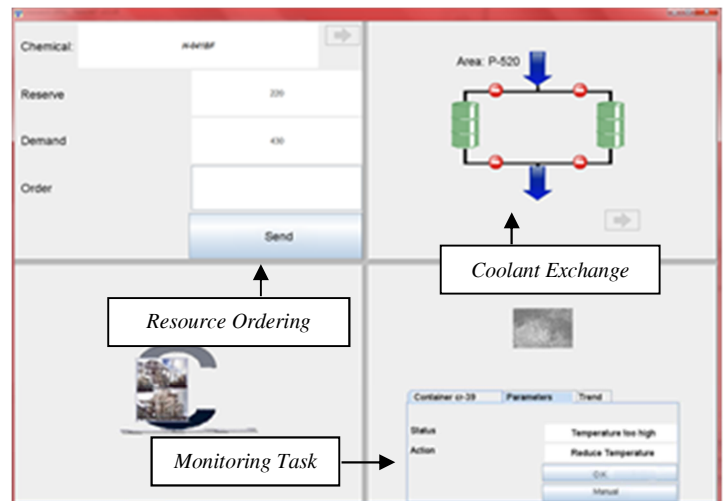

Figure 1. User interface of M-TOPS.

Coolant Exchange Task (CET). This task is displayed in the upper right quarter of the interface and involves an exchange of used coolant in two coolant tanks. For this purpose, different valves have to be opened and closed in a defined sequence. A complete exchange-cycle needs a minimum of 40s. After each completed exchange of coolant, indicated by a change of the color of the tank, participants have to re-activate the sub-system by a mouse-click on the "arrow" icon and a new set of tanks representing another sub-system is displayed.

Monitoring Task (MOT). This task is displayed in the lower right quadrant of the interface. Participants have to supervise a system that autonomously analyses and controls the processes in different "reaction chambers". Messages generated by the system always include an automatically derived description of the current state in a given compartment (e.g. "temperature high"; "process running") and an information about the action that will be implemented by the automation (e.g. "reduction of temperature"; "no action needed"). The participants are instructed that the automation would work highly reliable but not perfect. They can verify the automation by double-checking different raw data of the simulated process (e.g. temperature, pressure) and inspecting a graph of "heat distribution" in a given chamber. Complete verification always includes

inspection of two different parameters which can be made visible by single mouse-clicks. Whenever an automation failure is identified, participants have to transfer the process control for the given chamber to manual control (mouse-click on "manual" icon). Participants have between 15-25 seconds to decide either to cross-check or to accept the automation. Varying time-windows are chosen in order to avoid highly routinized workflows. The specific importance of this task for overall safety of the production process is stressed in the instruction.

Design

The participants were randomly assigned to one of three experimental groups: (1) Non-redundant (Non-R): Participants of this group were told that they would work in a control room of a chemical plant and that they were the only responsible operator for all three tasks. All tasks were declared to represent tasks with specific safety relevance (2) Redundant (R): Participants of this group got the same basic instruction as the Non-R group but were additionally told that they would work as member of a two-person crew. The ROT and CET were declared as tasks they were alone responsible for. But they were informed that the MOT would be performed in parallel by their crew mate and that they both were responsible for this task. That is, whenever one of them would detect an automation error and indicate it, the corresponding reaction chamber would be transferred to an independent work-station which than would take over manual control. This would not be notified to the partner. Every other detail but this part of the instruction corresponded directly to the non-redundant group. (3) Redundant-Feedback (R-FB): In this condition participants were instructed like in the "R" condition with the complementary information that in addition to the team performance also the individual's monitoring performance would be logged and fed back at the end of the experimental session.

A second experimental factor was represented by a time-on-task factor (within-subjects), i.e. performance in all three tasks was assessed separately for twelve successive blocks lasting four minutes each. During the first nine blocks and blocks #11 and 12 all diagnoses and actions proposed by the automation were correct. Yet, in block #10 a total of four automation failures occurred which were reflected in wrong diagnoses of the automation. This variation was included in order to assess performance of the participants in detecting rare automation failures which occurred surprisingly after a prolonged period of reliable performance.

Dependent variables

Performance measures for all three tasks were derived from a log-file listing all actions done by the participant.

Three measures were taken to assess monitoring performance in the MOT: (1) *Monitoring Quality* operationally defined as the number of reaction chambers for which the automation was double-checked by accessing the two relevant system parameters needed to verify it, (2) *Monitoring Quantity* defined as number of reaction chambers for which participants initiated any action to double-check the automation independent on whether or not this check was performed correctly, (3) *Automation Failure Detection Performance* defined by the percentage of participants in each group who correctly detected the different automation failures in block #10.

Performance in the ROT was assessed by the number of correctly sent orders.

Performance in the CET was operationally defined as the number of completed coolant exchange cycles per block.

In addition to the performance measures also some subjective measures were collected which were used for purposes of manipulation check as well as for some insight in the impact of the different conditions on perceptions of task responsibility and effort. The participants had to indicate for each task on 7-point Likert scales (1) to what extent they were alone in charge of the task (1: "not at all"; 7: "fully"), (2) to what extent they felt responsible for task performance (1: "very little"; 7: "very much"), and (3) how much effort they had invested in conducting the task (1: "very little"; 7: "very much").

Procedure

Two participants performed the experiment at the same time. They were sitting side to side at two workstations, separated by a movable wall. In the non-R condition the two participants were informed that they would work individually at two independent work stations. In the two redundancy conditions participants were formally assigned to a specific crew and informed that their work stations would be linked and that they would work together as a team.

An introduction to M-TOPS and the specific working condition (non-R, R-FB, R) was provided by means of a written task description including an in-depth description of the automation verification procedure as part of the MOT and some practice trials for each

subtask separately. After this introduction the participants performed a six minute practice trial according to their working conditions, with all three tasks to be performed simultaneously. The practice trials were followed by the data collection trials lasting 48 minutes (12 blocks á 4 minutes without any noticeable breaks in between). In the two redundancy conditions, the two participants were allowed to start the practice and data collection trials only when they had got a feedback on their screens that their partner had also logged into the system. This was done to support the illusion that both participants would really work on connected systems.

At the end of the session all participants completed a short survey containing some questions on demographic background and the subjective ratings described above. The session concluded with a debriefing of participants.

RESULTS

Manipulation Check and Subjective Data

As manipulation check, participants were asked to what extent they were alone in charge of the three sub-tasks. No differences emerged in responses to this question with respect to the ROT and CET where participants of all groups provided mean scores > 5.88 on the 7-point scale. However, this pattern was different for the MOT. As expected, participants of the Non-R group provided much higher ratings to this question (mean: 6.00) than the participants of the two redundant working groups (mean R-FB: 2.81; mean R: 2.75), $F(2,43)=29.55$, p<. 01.

A similar effect emerged with respect to the subjectively perceived responsibility for the MOT. Compared to participants of the Non-R group who felt considerably responsible for this task (mean: 6.64), the participants of the R-FB group (mean: 5.88) and even more those of the R group (mean: 5.06) expressed lesser levels of felt responsibility for the MOT, $F(2,43)=3,82$, p<. 05.

No differences between the experimental groups emerged in the subjective effort ratings for either task.

Performance Measures

Monitoring Task. Measures of monitoring quantity and quality were highly correlated, $r(46)= 0.84$, p < .001. Therefore statistical analyses were only performed for the measure of monitoring quality. Means of this measure for all groups across the twelve successive blocks are shown in figure 2. As becomes evident, participants in the Non-R control group performed considerably more double-checks of the automation than

participants in the R-group whereas performance in the R-FB group differed only slightly from the Non-R condition. In addition, the experience of automation failures in block #10 only exerted a visible impact on the participants in the R-FB group. While these participants increased their monitoring performance considerably in the following two blocks, no effects on monitoring performance was observed in the two other groups. Whereas performance in the non-redundant working group remained on an anyway high level, performance of the participants of the R-group remained comparatively low even after exposure to automation failures.
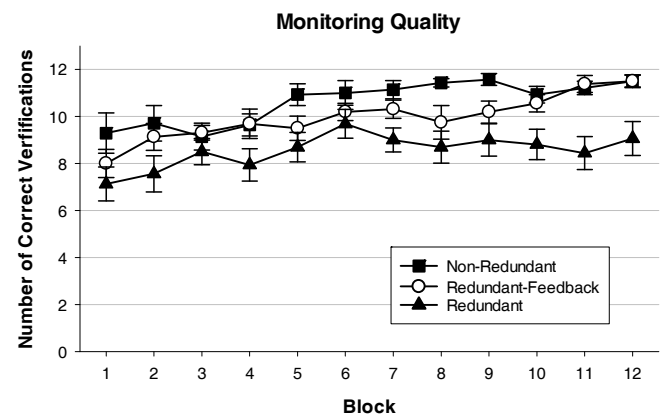

Figure 2. Mean number and standard errors of complete cross-checks in the Monitoring Task for the three groups across blocks.

Data were analyzed by a 3 (Group) x 12 (Blocks) ANOVA. This analysis revealed two main effects for group, $F(2,43) = 5.63$, p < 0.01 and block, $F (5.76, 247.53) = 14.83$; p < 0.01, as well as a significant group x block interaction, $F(22,473) = 1.70$, p<0.05.

In addition, a separate 3(Group) x 4 (Blocks) ANOVA was conducted only including blocks #8, 9, 11 and 12 (i.e. the blocks immediately before and after the block with automation failures). This analysis also revealed two significant main effects for group, $F(2,43)=8.2$, p<.01, and block, $F(1.85, 79.39)=3.8$, p<.05, as well as a significant group x block interaction, $F(3.7, 79.39)= 4.2$, p<.01.

The percentages of participants in each group who correctly identified the automation failures in block #10 are shown in table 1 for each of the four failures separately. In addition, also the percentages of participants are listed who identified ALL four failures in this block correctly.

Table 1: Automation failure detection performance

| Group | Automation Failures in Block #10 | | | | |
	1[stv]	2nd	3rd	4th	ALL
non-R	85.7	100	92.9	78.6	71.4
R-FB	75.0	81.3	93.8	93.8	56.3
R	50.0	56.3	68.8	87.5	31.3

As expected, clear performance differences emerged between the non-R and R group with respect to detection of the first two failures, as well as to the percentage of participants who detected ALL four failures correctly. Contrasts of the two groups by means of chi^2-tests (Fisher-Yates) revealed all these differences significant (one-tailed testing): $\chi^2(1)= 4.29$, p<.05 (1st), $\chi^2(1)= 7.99$, p<.01 (2nd). $\chi^2(1)= 4.28$, p<.05 (ALL).

Resource Ordering Task. Participants in all experimental groups showed a comparable performance in this task. Only the main effect of blocks became significant, $F(4.57, 196.62)= 16.46$, p< .01. Performance improved successively across blocks in all groups.

Coolant Exchange Task. Performance in the CET differed between groups, $F(2,43) = 3.27$, p<.05.. Participants of the non-R group on average completed slightly more exchange cycles (52.2; s=8.2) than participants in the R-FB (48.34; s=8.47) and R group (46.28, s=10.65). In addition also the block effect became significant, $F(6.69, 287.85)=16.42$, p<.01, indicating to a performance improvement across blocks in all groups.

DISCUSSION

The results of the present study confirm findings of earlier research and also provide new insights in the possible performance consequences of human redundancy in automation monitoring. The main findings can be summarized as follows: (1) Participants working in a redundant work setting were found to monitor the automation less carefully than participants who perceived themselves as the only responsible person for the task. This replicates results from the pilot study of Domeinski et a. (2007) who found similar effects, albeit for less trained subjects and significantly shorter work periods. (2) Going beyond these earlier findings, the present study also provides evidence that the reduction of individual monitoring performance in a redundant work setting can raise the risk of missing an automation failure that occurs surprisingly after a period of reliable functioning. (3) Most of the unwanted performance consequences of human redundancy could be reduced, albeit not fully avoided, by informing the redundantly working individuals that their individual performance would be logged and assessed. It can be assumed that this intervention has raised the perceived accountability for own performance even in a redundant work setting. Most interestingly it also led participants of this group to increase their monitoring effort

considerably in response to the experience of automation failures. No such effect was observed in the R group.

Overall this pattern of results fits nicely to the effects of social loafing known from group research in social psychology (Karau & Williams, 1993). They are further in line with the studies of team vs. individual effects on automation bias from Skitka et al. (2000) and Mosier et al. (2001). A possible theoretical explanation is provided by the so called *Collective Effort Model* (CEM, Karau & Williams, 1993) which suggests that individuals will be willing to exert effort on a collective task only to the degree that they expect their efforts to be instrumental in obtaining valued outcomes. The differences in the perceived responsibility for the MOT between the three groups further suggests that also a sort of social shirking of responsibility (Sagan, 2004) might have contributed to the observed loafing effects.

In summary and with respect to application the findings of the present study suggest that human redundancy may not necessarily be an effective safety measure for performing supervisory control tasks because possible positive effects on reliability can be offset(at least partially) by a sort of social-loafing effect.

REFERENCES

Conte, J. M., & Jacobs, R. R. (1997). Redundant systems influences on performance. *Human Performance, 10*(4), 361-380.

Domeinski, J., Wagner, R., Schoebel, M. & Manzey, D. (2007). Human redundancy in automation monitoring: Effects of social loafing and social compensation. Proceedings of the 52nd Human Factors and Ergonomics Society Annual Meeting, San Antonio (pp.   ). Santa Monica: HFES.

Felsenthal, D. S., & Fuchs, E. (1976). Experimental Evaluation of 5 Designs of Redundant Organizational Systems. *Administrative Science Quarterly, 21*(3), 474-488.

Hofmann, D. A., Jacobs, R., & Landy, F. (1995). High-Reliability Process Industries - Individual, Micro, and Macro Organizational Influences on Safety Performance. *Journal of Safety Research, 26*(3), 131-149.

Karau, S. J. & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology. 65*, 681–706.

La Porte, T. R., & Consolini, P. M. (1991). Working in practice but not in theory: theoretical challenges of high-reliability organizations. *Journal of Public Administration Research and Theory, 1*, 19-47.

Landau, M. (1969). Redundancy, Rationality, and Problem of Duplication and Overlap. *Public Administration Review, 29*(4), 346-358.

Latane, B., Williams, K., & Harkins, S. (1979). Many Hands Make Light the Work - Causes and Consequences of Social Loafing. Journal of Personality and Social Psychology, 37(6), 822-832.

Mosier, K.L., Skitka, L.J., Dubar, M. & McDonnell, L. (2001). Aircrews and automation bias: The advantages of teamwork? *The International Journal of Aviation Psychology, 11*, 1-14.

Roberts, K. H. (1990). Some characteristics of one type of high reliability organization. *Organization Science, 1*(2), 160-176.

Sagan, S. D. (2004). The problem of redundancy problem: Why more nuclear security forces may produce less nuclear security. *Risk Analysis, 24*(4), 935-946.

Skitka, L. J., Mosier, K. L., Burdick, M., & Rosenblatt, B. (2000). Automation bias and errors: Are crews better than individuals? *International Journal of Aviation Psychology, 10*(1), 85-97.

Swain, A.D. & Guttman, H.E. (1983). *Handbook of human reliability analysis with emphasis on in nuclear power plant application*s (NUREG/CR-1278). Sandia National Laboratories.