Operationalization of Learned Carelessness: An Experimental Approach

Frederik Aust
University of Duesseldorf, Duesseldorf, Germany

Christoph Moehlenbrink and Meike Jipp
German Aerospace Center, Braunschweig, Germany

The theory of learned carelessness offers an explanation why humans take unnecessary risks by omitting safety precautions against better judgment, but empirical research on learned carelessness is scarce. To test the theory 16 commercial aircraft pilots inspected flight plans on a multi-function display and the occurrence of flight plan errors was manipulated. Pilots rated effort of check performance, risk resulting from check omission and we measured the rate of falsely accepted erroneous flight plans. Participants who repeatedly encountered erroneous flight plans detected more errors during the test phase than participants who previously received only error-free flight plans ($p < .01$). The results provide evidence that subjective risk resulting from check omission affected the development of learned carelessness ($p = .053$), while effort of check performance displayed no effect ($p = .80$) due to invariance in ratings.

INTRODUCTION

Human factors research has a long tradition in civil and military aviation. Over the years an extensive effort has been made to ensure safe operations and reduce the number of accidents. It is a pilot's primary task during flight and on the ground to ensure passenger safety, but as humans by nature make mistakes, pilot error remains the most common cause of general aviation accidents (Hollnagel, 1993). As air traffic continually increases it is of great importance to understand which circumstances lead to a failure of perceptive and cognitive processes to warrant safe operations.

A common measure to prevent human error in aviation is the frequent use of checklists, which accurately guide check procedures, to ensure pilots take all necessary steps in the correct order. Checklists are worked off repeatedly but incidents are rarely encountered. According to Frey and Schulz-Hardt (1997) procedures of this type can lead to the development of a psychological state called learned carelessness.

Theory of Learned Carelessness

Frey and Schulz-Hardt (1997) assume humans to develop safety critical shortcuts in routine procedures, if after many repetitions they encounter no critical incidents. Through operant conditioning (Skinner, 1953), model learning (Bandura, 1977) and social norms the *monopoly hypothesis* "everything is fine and will of its own volition remain to be fine" is acquired. Behavioral or cognitive steps are increasingly abbreviated or omitted to reduce effort. The human tendency to follow the path of least cognitive resistance is also referred to as "cognitive miser" (Wickens & Hollands, 2000). A reduction in effort is positively reinforcing and therefore increases the likelihood of future shortcuts in the absence of negative consequences. The underlying motivation is postulated to be hedonism: maximizing pleasure while minimizing discomfort. Once learned carelessness has developed it will distort a person's perception, selection and interpretation of subsequent information in favor of the monopoly hypothesis. This top-down information processing impairs motivation and capability to detect incidents. The result is unreasonably risky behavior. However, developing learned carelessness is not universally disadvantageous. Without mechanisms to conserve cognitive resources, operators would be unable to perform safety critical actions in complex environments (Lüdtke, 2004).

Frey and Schulz-Hardt (1997) imply several testable factors moderating the development of learned carelessness: (1) An increasing number of repetitions of a procedure should facilitate the development of learned carelessness. (2) Effortful steps of a procedure should be especially prone to be cut short and omitted. (3) The higher the safety risk resulting from careless behavior the less likely it should be. Studies accordingly have shown people to take higher risks when negative consequences are improbable (e.g. Musahl, 1997). (4) Encountering critical incidents should reduce learned carelessness. High effort combined with low safety risk resulting from omission should therefore best predict the development of learned carelessness.

Closely related to the concept of learned carelessness are those of automation complacency (Wiener, 1981; Parasuraman, Molloy & Singh, 1993) and automation bias (Mosier & Skitka, 1996). To-date, there is no widely accepted definition for automation complacency. Many definitions agree it describes a psychological state, in which an automated system is monitored at a suboptimal frequency resulting in observable effects on system performance (Parasuraman & Manzey, 2010). Automation bias is at hand when an operator overly relies on outcomes of decision aids instead of adequately seeking and processing information (Mosier & Skitka, 1996). This trust in automation can lead to omission errors, when users miss a critical signal without system alerts, or commission errors, when they follow a decision aid's incorrect advice.

Figure 1. Setup of the testing environment: The left display shows the Advanced Human Machine Interface, the right one the primary flight display. The computer monitor on the far left displays the secondary color discrimination task.

In their literature review Parasuraman and Manzey (2010) suggest that automation complacency and automation bias may be attributed to common attentional processes. They put forward an integrated model of complacency, in which a "complacency potential" can cause attentional biases in information processing and lead to a loss of situation awareness. If automation functions correctly in this state, no performance consequences arise and over time processes of learned carelessness are induced, increasing the complacency potential. A failure of automation on the other hand leads to omission or commission errors, thereby decreasing the complacency potential.

Research Question

Learned carelessness is thought to be a general principle based on operant mechanisms and not limited to interactions with automation. The theory's explanations and predictions may be applicable to a wide range of tasks and situations (Frey & Schulz-Hardt, 1997). Schulz-Hardt, Frey and Lüthgens (1996) have suggested behavioral indicators of learned carelessness including reduced vigilance for unexpected events and development of behavioral or cognitive shortcuts, but empirical research on the subject is scarce. We found only one published study on the validity of the theory of learned carelessness. Lüdtke and Möbus (2004) report a case study on cognitive modeling of commercial aircraft pilots' behavior. To validate their cognitive model, behavioral data of one pilot were compared to data simulated either with or without a learned carelessness component. Taking the additional component into account significantly improved the simulation's fit to the observed behavior.

Experimental studies and paradigms are needed to produce more reliable evidence on the theory of learned carelessness. In a first attempt, we developed a multi-step inspection task in a flight simulation to elicit learned carelessness in pilots and observe the development of shortcuts within a procedure. We furthermore tested the predictive quality of subjective effort and risk resulting from omission, which are postulated to moderate the development of shortcuts.

METHOD

Sample

Sixteen German commercial aircraft pilots from different airlines participated in the experiment. Due to technical problems three pilots had to be excluded from the analyses. The remaining participants' mean age was 34 years ($SD = 11.31$) with a range of 23-66 years and the mean self-reported flight time was 5927 hours ($SD = 5930$). One pilot was female and except for two captains all participants were first officers.

Material

The study was conducted in a low-fidelity flight simulator (Figure 1). No forward field of view was simulated because this study was concerned with behavior during instrument flight. The bottom right of the flight simulator's four computer monitors displayed a primary flight display. The primary flight display served pilots as a means to monitor the aircraft's flight and increased the ecological validity of the simulation. In addition participants wore a helmet carrying an eye- and head-tracking device.

Primary task. The pilots' primary task was to check flight plans on the Advanced Human Machine Interface, a multi-function display and graphical user interface for the Advanced Flight Management System, for errors (Korn & Kuenz, 2006). The system allows for flight plan creation, editing, trajectory generation and negotiation with air traffic control via data link communication. The system is able to generate 4D-trajectories in space and time and has two views available: a horizontal view displaying maps, waypoints and weather, and a vertical profile view. Pilots performed this task using a computer mouse. The Advanced Human Machine Interface serves as an ecologically valid testing environment for the study of learned carelessness in pilots because multi-function displays are being increasingly incorporated into modern cockpit layouts. It was displayed on the simulator's bottom left monitor.

All flight plans contained approaches to Frankfurt airport (Germany) from one of four orientations. Participants had to check for six possible errors: (1) The flight plan did not end on the runway, (2) the aircraft had already passed the first waypoint of the flight plan, (3) the generated trajectory contained an undesired circle back to its first waypoint the aircraft had already passed, (4) the cruise flight level was set to 0 feet, (5) the glide slope intercept altitude was incorrect, and (6) the runway altitude was incorrect. Checks for Errors 1-3 were conducted in the horizontal view of the Advanced Human Machine Interface, Checks 4-6 in the vertical profile view.

Participants answered 5 questions on a 6-point likert scale to evaluate the perceived effort of check performance and the perceived risk resulting from check omissions. Because of the high similarity and interdependence of Check 2 and Check 3, subjects rated them together with one item. The ratings were used to test for differences between checks and to predict the likelihood of development of learned carelessness.

Secondary task. A fifth monitor was placed on the far left of the flight simulator setup in the participants' peripheral field of view (Figure 1). The monitor displayed red and green rectangles as a secondary color discrimination task to increase pilots' workload. Workload affects learned carelessness because conserving mental resources is especially important when the cognitive load during task performance is high. Workload has accordingly been shown to facilitate complacency (Parasuraman, Molloy & Singh, 1993). Participants responded to the stimuli using the left and right buttons of a second computer mouse in their non-dominant hand.

Procedure

Each participant was invited for two consecutive days. First pilots were introduced to the Advanced Human Machine Interface. Upon arrival of a new flight plan from air traffic control participants had to perform checks for the six possible errors. The checks could be performed in any preferred order. Error-free flight plans were accepted as new routes of travel and activated. Erroneous flight plans had to be rejected. In addition to the six errors, pilots were instructed to reject flight plans leading through areas of bad weather, to increase the procedure's complexity and participants' workload. Because no flight plan had to be rejected due to bad weather, this check was omitted from the analyses. Pilots practiced this procedure until they successfully performed and verbalized all seven checks on three consecutive flight plans.

The flight plans were presented in a fixed order in five *build up* blocks of 30 flight plans each. The first three blocks were completed on the first day, the fourth and fifth block on the second day. Pilots worked self-paced and each build up block lasted approximately 35 minutes. The second and fifth block were followed by a 15-minute break to counter effects of fatigue, monotony, and vigilance decrement. During build up blocks an observer watched the pilots' performance and logged whether each check was performed or not with the help of the eye tracking system's live-view. If the observer was uncertain about check performance, it was marked as performed.

Following the fifth block, pilots unknowingly completed a final test block consisting of 20 flight plans. During this block each of the six possible errors occurred once. The number of falsely accepted erroneous flight plans was measured. After the test phase participants answered the questionnaire evaluating the perceived effort of check performance and the perceived risk resulting from check omissions.

During the entire experiment, participants performed the secondary color discrimination task while inspecting flight plans. Pilots were instructed to respond as quickly and accurately as possible.

Variable Description

Participants were randomly assigned to one of two groups. The presence of erroneous flight plans during the build up blocks was manipulated as independent variable. To induce learned carelessness the experimental group ($n = 9$) checked only error-free flight plans during build up blocks. The control group ($n = 4$) received 12 erroneous flight plans during each build up block, each possible error occurring twice. We measured the number of observably performed checks during build up blocks, error rates and the checks' effort and omission risk ratings as dependent variables. The pilots' error rate was measured as the relative number of falsely accepted erroneous flight plans during the test phase. For these dependent variables we propose the following hypotheses:

- The experimental group will perform more observable checks during the first than during the fifth build up block. There will be no difference in the control group.
- During the test phase the experimental group will falsely accept more erroneous flight plans than the control group.
- The higher a check's effort rating the more flight plans containing the corresponding error will be falsely accepted during the test phase.
- The lower a check's omission risk rating the more flight plans containing the corresponding error will be falsely accepted during the test phase.

Data Analyses and Results

The data were analyzed using R 2.11.1 (R Development Core Team, 2010) and G*Power 3.1.2 (Faul, Erdfelder, Lang & Buchner, 2007). If not reported otherwise the α-error probability was .05.

Effort and Omission Risk Ratings

To test the hypothesis that learned carelessness can be predicted through the effort of a procedure's step and the subjective risk resulting from its omission steps necessary to check flight plans on the Advanced Human Machine Interface had to differ in respect to these variables. Check 2 and Check 3 entered the analyses as one because they were not rated separately they.

Checks 1-4 ($Mdn = 3$) were rated more effortful than Check 5 and Check 6 ($Mdn = 2$). However the Friedman-Test for repeated measures found the differences to be statistically insignificant ($\chi^2 = 5.98$, $df = 4$, $p = .20$). The subjective risk resulting from check omission also exhibited little variance between checks (Figure 2). Checks 1-4 and Check 6 received median ratings of 4. Only Check 5's omission was rated as riskier ($Mdn = 5$) but the differences between checks again were statistically insignificant ($\chi^2 = 4.43$, $df = 4$, $p = .35$).

Check Performance

Behavioral observation. By comparing two independent behavioral observations of two participants totaling 1,800 decisions the inter-rater reliability was computed to be $\kappa = .79$. Observations were not comparable between groups because the control group encountered erroneous flight plans during build up blocks. After detection of an error, pilots immediately rejected the flight plans without performing the remaining checks. The following two tests were Bonferroni-corrected yielding a α-error probability of .025. The experimental group exhibited no difference in observable checks between the first

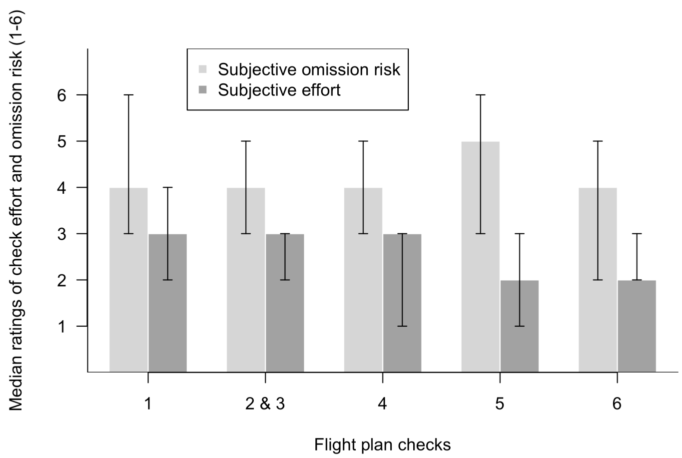

Figure 2. Ratings of effort necessary to perform flight plan checks on the Advanced Human Machine Interface and risk resulting from check omission. The error bars represent the first and third quartile.

and the fifth build up block ($t = -.56$, $df = 53$, $p = .71$, $1\text{-}\beta = .30$) and neither did the control group ($t = 1.37$, $df = 23$, $p = .09$, $1\text{-}\beta = .17$).

Error rates. Error 2 and Error 3 were correctable by regenerating the trajectory. The new trajectories could then rightly be accepted. Thus error rates for Check 2 and Check 3 cannot be interpreted and were not analyzed. Due to the small number of participants we used Fisher's exact test to compare group differences. The following five tests were Bonferroni-corrected yielding a α-error probability of .01.

The analysis of error rates shows that the control group (.00) accepted significantly fewer erroneous flight plans during the test phase than the experimental group (.53, $p < .01$, $1\text{-}\beta = .99$). Analyzing checks individually we found a marginally significant difference in error rates between the groups for Check 5 ($p = .02$, $1\text{-}\beta = .37$, Table 1). Note that Check 5 received the highest omission risk rating but the lowest effort rating. However, these results should be interpreted with care due to the analyses' lack of statistical power, which were far from the conventionally aspired minimum of .80 ($.01 \leq 1\text{-}\beta \leq .37$; Cohen, 1992). Differences in error rates between experimental and control group could be ascribed to a response bias, if the control group displayed an increased rate of falsely rejected flight plans during the test phase. As the analysis shows, the control group exhibited no false rejections, whereas the experimental group falsely rejected two flight plans (1.59%). Using Fisher's exact test the difference did not reach significance ($p > .99$, $1\text{-}\beta = .01$, one-tailed). The difference in error rates between the groups therefore is not the result of a response bias.

To test the hypothesis that ratings of effort and risk resulting from omission predict the development of learned carelessness, we compared ratings of checks that were not performed during the test phase with ratings of checks that were performed using Wilcoxon's rank-sum test. Performed checks' risk resulting from omission was rated higher than that of checks not performed. This difference was close to significance ($W = 223$, $p = .053$, $1\text{-}\beta = .52$, one-tailed). However the analysis yielded no significant difference between effort ratings of omitted and performed checks ($W = 263.5$, $p = .80$, $1\text{-}\beta = .37$, one-tailed). To quantify the ratings' ability to predict rejections or acceptances of erroneous flight plans, respectively, we computed rank biserial correlations (Glass, 1966). The correlation between omission risk ratings and error detection was $r_{rb} = -.27$ and according to Willson's test for large samples close to significance ($z = -1.59$, $p = .055$, one-tailed, Willson, 1976), while effort ratings exhibited an insignificant correlation of $r_{rb} = -.14$ ($z = -.82$, $p = .79$, one-tailed).

Discussion

Frey and Schulz-Hardt (1997) postulate, that a growing number of incident-free repetitions of a procedure should increase carelessness on a cognitive and behavioral level. Observing pilots' performance, we found no difference in performance between their first and last 30 flight plan inspections in the build up phase. However, as predicted, after processing 150 error-free flight plans, pilots overlooked significantly more errors during the test phase than those who had repeatedly encountered erroneous flight plans. The difference was not attributable to a response bias. It appears we induced carelessness on a cognitive level only. Eye movement has been proposed as a direct measure of attention allocation in the research of automation complacency and related concepts (Parasuraman & Manzey, 2010). Observers viewed participants eye movements during flight plan inspection to determine whether checks were performed. Although we did not perform a detailed analysis the results provide evidence that in our inspection paradigm careless pilots exhibited a dissociation of fixation and attention allocation called "looking-but-not-seeing" (e.g. Thomas & Wickens, 2006). A possible explanation for this result is impression management (Goffman, 1956). Because the task was self-worth-relevant, participants may have felt the need to present themselves as responsible pilots by performing checks but nonetheless doing so carelessly.

In this study, the subjective risk resulting from a check's omission appeared to affect the acquisition of learned carelessness, while the effort of check performance did not. Omitted checks were descriptively rated as less safety critical but not as more effortful. The difference in checks' effort ratings, however, was small and may have been either a genuine effect or the result of social desirability. In any case, future researchers should consider independent expert evaluations and the use of objective measures to evaluate task components (e.g. processing time for the assessment of effort). The difficulties predicting error rates through the ratings according to the theory hamper the interpretation of the results in terms of learned carelessness.

An alternative explanation for the pattern of results is a decrement in vigilance as an effect of time and difference in signal probability (Parasuraman, 1986). However, the inspection task was self-paced, there was no spatial uncertainty about signals, pilots had a 15-minute break before the test block and a descriptive ROC-analysis of responses during the test block revealed a difference between groups in sensitivity and a more conservative criterion in the control group in spite

Table 1
Rate of falsely accepted erroneous flight plans during the test phase for each check

Groups	Flight plan checks					
	1	2	3	4	5	6
Control (n = 4)	.00	-	-	.00	.00	.00
Experimental (n = 9)	.66	-	-	.22	.78	.44
p-value	.09	-	-	.46	.02	.18
Power (1-β)	.04	-	-	<.01	.37	.01

of a higher signal rate during build up blocks ($d'_{exp} = 1.74$, $c_{exp} = 1.19$, $d'_{con} = 3.01$, $c_{con} = .86$). Furthermore the predictive validity of the omission risk ratings suggests that the pattern of results cannot be entirely explained by a vigilance-decrement over time alone. We introduced frequent breaks during the build up phase also to control for effects of fatigue but possible confoundings with experimental conditions cannot be precluded. Studies controlling fatigue using objective measures like electroencephalography are necessary.

In this study, we used error rates to operationalize learned carelessness as suggested by Schulz-Hardt, Frey, and Lüthgens (1996). To further test the validity of the theory, the development of methodologies measuring effects on cognition before manifestation of performance effects is desirable (Frey & Schulz-Hardt, 1997; Parasuraman & Manzey, 2010). As learned carelessness is assumed to be a general phenomenon, research using procedures and tasks from other areas of aviation (e.g. maintenance) is also of interest. Furthermore additional studies replicating the results with larger samples are needed because interpretations of some of our analyses' were difficult due to a lack of statistical power.

In summary, the study at hand appears to have successfully induced processes of learned carelessness in commercial aircraft pilots. Although not without methodological challenges the results provide evidence that some of the predictions made by the theory of learned carelessness appear to be valid. This being the first experimental study on the subject, however, additional research into the mechanisms of its development and the demarcation from other concepts is necessary.

AUTHOR NOTE

Frederik Aust, Institute of Experimental Psychology, Univeristy of Duesseldorf; Christoph Moehlenbrink and Meike Jipp, Institute of Flight Guidance, German Aerospace Center.

This study was conducted as part of the European Union project HUMAN (No. 211988). We would like to thank Helge Lenz, Matthias Wies and Tristan Schindler for assisting with the data collection and Birk Diedenhofen for helpful comments on this manuscript.

Correspondence concerning this article should be addressed to Christoph Moehlenbrink, Institute of Flight Guidance, German Aerospace Center, Lilienthalplatz 7, D-38108 Braunschweig, Germany. christoph.moehlenbrink@dlr.de

REFERENCES

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice Hall.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin,* 112 (1), 155-159.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods,* 39, 175-191.

Frey, D. & Schulz-Hardt, S. (1997). Eine Theorie der gelernten Sorglosigkeit. In H. Mandl (Ed.), *Bericht über den 40. Kongress der Deutschen Gesellschaft für Psychologie* (pp. 604-611). Göttingen, Germany: Hogrefe Verlag für Psychologie.

Glass, G. V. (1966). Note on rank biserial correlation. *Educational and Psychological Measurement*, 26, 623-631.

Goffman, E. (1956). Embarrassment and Social Interaction. *The American Journal of Sociology,* 62 (3), 264–71.

Hollnagel, E. (1993) *Human reliability analysis: Context and control*. London, England: Academic Press.

Korn, B. & Kuenz, A. (2006). *4D FMS for increasing efficiency of TMA operations*. 25th Digital Avionics System Conference, October 15th 2006, Portland, OR.

Lüdtke, A. (2004). *Kognitive Analyse Formaler Sicherheitskritischer Steuerungssysteme auf Basis eines integrierten Mensch-Maschine-Modells*. Berlin, Germany: Akademische Verlagsgesellschaft Aka GmbH.

Lüdtke, A. & Möbus, C. (2004). A Cognitive Pilot Model to Predict Learned Carelessness for System Design. In A. Pritchett and A. Jackson (Eds.), *Proceedings of the International Conference on Human-Computer Interaction in Aeronautics*. Menlo Park, CA: The AAAI Press.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and application* (pp. 201–220). Mahwah, NJ: Erlbaum.

Musahl, H.-P. (1997). *Gefahrenkognition*. Heidelberg, Germany: Asanger.

Parasuraman, R. (1986). Vigilance, monitoring and search. In K.R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of perception and human factors: Vol. II. Cognitive processes and performance* (pp. 43-1-43-39). New York, NY: Wiley.

Parasuraman, R. & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52, 381-410.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *International Journal of Aviation Psychology*, 3, 1–23.

R Development Core Team (2010). *R: A language and environment for statistical computing*. Vienna, Switzerland: R Foundation for Statistical Computing.

Schulz-Hardt, S., Frey, D. & Lüthgens, C. (1996). Sorglosigkeit und Risikoakzeptanz. In G. Wenninger & C. Graf Hoyos (Eds.), *Arbeits-, Gesundheits- und Umweltschutz - Handwörterbuch verhaltenswissenschaftlicher Grundbegriffe* (pp. 468-479). Heidelberg, Germany: Asanger.

Skinner, B. F. (1953). *Science and Human Behavior*. New York, NY: Macmillan.

Thomas, L. & Wickens, C. D. (2006). Effects of battlefield display frames of reference on navigation tasks, spatial judgments, and change detection. *Ergonomics*, 49, 1154–1173.

Wickens, C. D.,& Hollands, J. G. (2000). *Engineering psychology and human performance*. New York, NY: Prentice Hall.

Wiener, E. L. (1981). Complacency: Is the term useful for air safety? In *Proceedings of the 26th Corporate Aviation Safety Seminar* (pp. 116–125). Denver, CO: Flight Safety Foundation

Willson, V. L. (1976). Critical values of the rank-biserial correlation coefficient. *Educational and Psychological Measurement*, 36, 297-300.