MTR190016

MITRE TECHNICAL REPORT

MITRE

Is Something Wrong? A Review of the Failure-Detection Research

Dr. Valerie Gawron

January 2019

The author's affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by the author.

Approved for Public Release; Distribution Unlimited. Case Number 16-3426.

McLean, VA

Center for Advanced Aviation System Development

Table of Contents

List of Tables

Is Something Wrong? A Review of the Failure-Detection Research

The research to be reviewed in this article deals with observers passively monitoring an automated system and searching for malfunctions in that system. The prime example is the pilot monitoring the autopilot. In this case the operator (i.e., the pilot) who has learned to rely on the automated system for crucial decision-making, must detect a failure in that system. Research examining this problem has been sparse.

An early study in this area was performed by Miller and Elkind (1967). They introduced failures (changes in the magnitude or polarity of gain) during performance of a compensatory tracking task. The authors reported that the detection latencies for these failures were highly variable, especially for gain increases and polarity reversals. In an earlier study, Miller (1965), cited in Young, 1969) reported that gain increases and reversals were quickly detected. He suggested that the short detection times associated with these failures were due to the large errors in the system that resulted when gain increases or polarity reversals went undetected. This suggests, not surprisingly, that the salience of the failure is a critical element in failure-detection performance. Rolfe (1972) that such performance is likely to occur:

- "When an unexpected event is very probable.
- When the operator is anxious.
- When the operator's attention is being distracted.
- During a period of reaction following a time of high stress.
- When a set pattern of interpretation and action has been held for a long time" (p. 76).

Curry and Gai (1976) took a different approach and state that responding to a system failure can be decomposed into three functions: 1) failure detection, 2) failure identification, and 3) corrective action. Errors occur at any of these three functions when the operator's mental model of how the system works, based on his or her training and experience, does not match reality. Errors also occur whenever there is deficient information - either in quantity or accuracy. When the operator has deficient information, workload induced stress can cause deficiencies in cognitive functions. This is especially true for continuous failures rather than discrete failures and for decreases in frequency and variance in failures rather than for increases. The authors also hypothesized that fatigue and low motivation would result in even further deficiencies in cognitive functioning. Gai and Curry (1977) directly tested this hypothesis. Two pilots flew fully automated approaches and landings in an aircraft simulator. In several of the approaches, errors in the glide slope or airspeed indicators were introduced. The authors reported that failure-detection time decreased as failure magnitude increased. Further, the pilots tended to detect errors in the glide slope indicator faster than those in the airspeed indicator.

Vreuls, Barneby, Nichols, and Dent (1968) also examined the effect of failure salience on detection times. These researchers introduced three types of autopilot failure during final approaches flown in an aircraft simulator:

1. Passive control failure (dead fail),
2. Soft over control failure (0.25 degree/second drift), or
3. Flare mode engagement failure.

They reported that 39 of the 216 passive autopilot failures went undetected while only 5 of the 216 soft over autopilot failures were missed. The number of misses decreased when the pilots were provided with a caution light. Without a caution light, however, the detection latencies for passive failures (mean = 15.68 seconds) were reliably longer than those for soft over failure (mean = 7.15 seconds). The most difficult failure to detect (i.e., the one associated with the longest latencies) was a passive failure in the pitch axis. Vreuls, Barneby, Nichols, and Dent (1968) also reported the occurrence of three inappropriate responses to system failures; pilots:

1. Misidentified the axis of the failure,
2. Disengaged the wrong axis, and
3. Disengaged the whole autopilot.

In another elaborate study, Ephrath and Curry (1977) measured the workload and failure-detection performance of 15 professional pilots landing an aircraft simulator in zero-zero conditions. The authors varied the parameters in Table 1. They found that workload (as measured by reaction time to a secondary, non-relevant stimulus onset) increased dramatically as the amount of automated control increased. Also, failure-detection times were longer when the failed axis was under manual rather than automatic control. Further, system failures went undetected only when the failed axis was being controlled manually and never when it was automatically controlled.

Table 1. ***Variables Used by Ephrath and Curry (1977) in their Study of 15 Professional Pilots Landing on Aircraft Simulator***

Type and Amount of Automated Control
• Completely Automated Flight Control
• Manual Control of Lateral Axis
• Manual Control of Longitudinal Axis
• Completely Manual
Degree of Wind Disturbance
• No Wind
• 45-Degree Tailwind of 5 Knots with Gusts to 15 Knots
• 45-Degree Tailwind of 10 Knots with Gusts to 30 Knots
Type of System Failures
• Lateral Axis Failure
• Longitudinal Axis Failure
• No Failure

Other findings included that detection times for lateral-axis failures were longer than those for longitudinal failures at comparable workloads. Also, as expected, higher wind turbulence resulted in higher workloads and longer failure-detection times. However, failure-detection time was not a monotonic function of workload. Ephrath and Curry (1977) concluded from this that the participation mode (i.e., type and flight control) influenced failure detection independent of the associated differences in workload among modes.

From the findings described above, it would seem that automatic control is superior to manual control. However, other studies have not been as supportive. This is especially true of research performed by Kessel and Wickens (1978) at the University of Illinois. They have conducted experiments over a five-year period that have consistently shown that operators generally are slower to detect a failure and recover from it when they are passively monitoring, rather than actively controlling, the system. Their research is described in detail below.

Kessel and Wickens (1978) reported two experiments that compared failure-detection performance in manual and automated modes. In the first experiment, step changes in system order were introduced into a 2-dimensional, pursuit-tracking task. Subjects were instructed to press a trigger on the joystick when they detected a failure. Kessel and Wickens found consistent superiority in failure-detection latency and accuracy when the operator was manually controlling the system rather than passively monitoring it. They attributed this difference to the proprioceptive cues available only in the manual control mode. In their second study, they directly tested this hypothesis by having subjects perform the tracking task using a joystick with 0 or 520 grams of resistance. Contrary to expectation, they found that the isotonic joystick did not degrade failure-detection performance.

Wickens and Kessel (1979) again compared failure-detection performance in manual and automatic control modes and again found longer failure-detection latencies in the automatic than the manual mode. Wickens and Kessel (1980) added a secondary task (either tracking or mental arithmetic) to their experimental paradigm and found that failure detection in the automatic mode was adversely affected by the mental-arithmetic task but not by the secondary tracking task. Conversely, failure detection in the manual mode was worse when the secondary task was tracking. Again, overall, failure-detection performance in the manual mode was characterized by shorter detection latencies and greater accuracy than that in the automatic mode.

Wickens and Kessel presented their conclusions in a general review paper published the next year.

> *Our conclusions asserted that the impact of the loading tasks was upon the processing channels used to monitor the system, visual for the AU (automatic mode) detection, proprioceptive for MA (manual mode). The former utilizes resources associated with perception and central processing, the same resources as those demanded by the mental arithmetic task; the latter utilizes response-related resources, coincident with the subcritical loading task. ... A point of more general relevance here is that automation ... does not necessarily eliminate or even reduce the workload of the human operator, but merely shifts the locus of processing demands (p. 146).*

In a similar vein, Sheridan (1976) proposed the following general guideline for system control:

> *In-the-loop participation (manual control) is best when error alone is insufficient or the input of motor 'identification signals' permits quick adaptation; but if control is advertently noisy or requires full attention to steering displays and leaves little time for other displays which offer important failure cues, the man should monitor and let the machine control (p. 176).*

Rouse (1981) tried to resolve the automatic/manual control superiority controversy. He suggests that:

> *It seems reasonable to conjecture that having to control while monitoring for failures is beneficial if performing the control task provides cues that directly help to detect failures and if the workload is low enough to allow the human to utilize*

these cues. Otherwise, controlling simply increases workload and decreases the amount of attention that can be devoted to failure detection (p. 83).

In comparing Wickens and Kessel's work (based on performing a two-dimensional tracking task) with that of Curry, Ephrath, and Vreuls (landing aircraft), Rouse's suggestion seems eminently plausible.

Parasuraman, Molloy, Mouloua, and Hilburn (1996) also considered the effect of workload. These authors summarized a study in which subjects performed tracking and fuel-management tasks manually and either monitored an automated engine status task or performed this task manually. In the manual mode, subjects detected 70% of the engine malfunctions. In the automated mode, they detected less than 40%. For pilots, the detection rate was less than 60%. In another experiment, they reported that the detection rate of automation failures varied inversely with automation reliability.

In addition, Kessel (1986) performed four experiments using a simple collision avoidance game to assess the effects of automation on performance. She concluded that automation degraded operator performance because the operator had to predict and adapt to the automation system's actions. When the roles were reversed and the automated system had to adapt to operator performance, operator performance was enhanced.

Detection rates also seem to vary with pilot strategy. Beringer (1996) analyzed the recovery from automation malfunctions of 29 pilots none of whom had less than 300 hours of flight time. The data were collected in a fixed-base aircraft simulator. There were four types of failures: command over role (rate = 6 deg/sec), soft roll (rate = 1 deg/sec), soft pitch (rate = 0.2 deg/sec), and runaway pitch up trim. Response times were longer for covert changes (e.g., soft pitch ranged from 21.2 to 85.1 seconds) than for overt changes (e.g., command over roll ranged from 11.8 to 53.8 seconds). Two distinct response strategies were observed: immediate disconnect and manual override. In a follow up study, Beringer and Harris (1997) used the same simulator to observe malfunction recoveries of 24 pilots ranging in flight hours from 290 to 10,000 hours. Malfunction types were selected to cover the entire range from largely covert to largely overt: soft pitch failure rate = 0.2 deg/sec), attitude indicator failure (slow drift), runaway pitch down, and runaway roll servo (roll rate 12 to 15 degrees per second). First response and/or recognition times were commensurate with the overtness of the malfunction:

1. Soft pitch failure rate – 330 milliseconds to 73.7 seconds
2. Attitude indicator failure – 12.7 to 263 seconds
3. Runaway pitch down – 3.6 – 15.8
4. Runaway roll servo – 1.09 to 4.88 seconds.

The problem is not restricted to aviation but to any automated system. Kerstholt and Passenier (2000) reported the results of two experiments in maritime supervisory control. In the first experiment, 39 maritime-studies students diagnosed disturbances that could be real or false alarms. The presence of false alarms increased the rate at which the automated systems were sampled but also increased the problem-solving time. The problem-solving time was longer in part because the students solved the problems sequentially not concurrently. This tunnel vision has been noted in many automation related accidents (e.g., Eastern L-1011, Miami Florida, December 29, 1972). In the second experiment, the authors provided the student with either interactive support in which the student had to enter values related to the subsystems that were evidencing disturbances or a completely automated system that diagnosed the problem. The subjects were 30 maritime-studies students. Ironically even though both systems gave the same advice, more incorrect actions were taken with the completely automated system. When the support systems were removed, students who had used the interactive system showed a more structured approach to problem solving that student who had used the completely automated system.

Even with the controversial findings described above, there have even been attempts at developing a mathematical model of failure-detection performance. Specifically, Repperger, Haas, Schley, and Koivo (1998) have developed mathematical models to predict loss of control due to failures in the human interface device.

Summary

Automation has been used to enhance safety, increase productivity, reduce operator workload, and minimize error. But along with its many benefits have come problems. One of the most serious is a system failure that forces the operator (or user) to suddenly and unexpectedly enter the control loop. The operator, who has learned to rely on the automated system for crucial decision-making, must first detect the system failure and then take the steps necessary to recover. In some systems (e.g., aircraft, air traffic control, nuclear power plants) the operator must not only act, but also act quickly and confidently since recovery time may be critical.

To date, research comparing failure-detection performance in automated and manual control modes has been both sparse and contradictory. Work performed in the late seventies by Ephrath and Curry (1977) appeared to confirm the superiority of automatic over manual control in several areas including failure-detection performance. A series of studies by Kessel and Wickens (1978, 1982), however has repeatedly found longer failure-detection latencies in the automatic than in the manual mode. Other researchers (Ephrath, 1980; Stewart, 1978; Wiener & Curry, 1980b) have also reported problems associated with automated systems that range from no reduction in workload over manual systems to automation-induced errors. Clearly, additional research is necessary to discover the reasons for the contradictory findings and to establish the framework for making optimal use of the benefits inherent in automated systems.

Findings in the vigilance and failure-detection literature suggest the importance of operator arousal to performance. In the vigilance literature, stimulants, loud ambient noise, and the threat of electric shock improved detection performance. Lengthening signal durations, maintaining a high signal rate, increasing signal intensity, and increasing signal complexity also improved vigilance. It would seem that making the task more stimulating, improved performance.

In the failure-detection literature, performance was better in the automatic than in the manual mode when the control task was complex (i.e., piloting an aircraft), (e.g., Ehprath & Curry, 1977). The converse was true, however, when the task was less complex (i.e., 2-D, pursuit-tracking) (e.g., Kessel &Wickens, 1978).

References

Beringer, D. B. & Harris, H. C. (1997). *Automation in General Aviation Part II: four ways to reach zero feet AGL unintentionally – autopilot and pitch trim malfunctions,* (pp. 75–79). Proceedings of the Human Factors and Ergonomics Society 41st Annual Meeting,

Beringer, D. B. (1996). *Automation in General Aviation: Responses of pilots to autopilot and pitch trim malfunctions,* (pp. 86–90). Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting.

Curry, R. E. & Gai, E. G. (1976). Detection of random process failure by human monitors. In T. B. Sheridan & G. Johannsen (Eds.), *Monitoring behavior and supervisory control.* New York, New York: Plenum.

Ephrath, A. R. (1980). Pilot workload during an instrument approach. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-10(10)* 676–678.

Ephrath, A. R. & Curry, R. E. (1977). Detection by pilots of system failures during instrument landings. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-7 (12)* 841–848.

Gai, E. G. & Curry, R. E. (1977). Failure detection by pilots during automatic landing: Models and experiments. *Journal of Aircraft, 14(2),* 135–141.

Kerstholt, J. H. & Passenier, P. O. (2000). Fault management in supervisory control: the effect of false alarms and support. *Ergonomics, 43(9),* 1371–1389.

Kessel, C. & Wickens, C. D. (December 1978). Development and utilization of internal models in dynamic systems. A comparison of monitors and operators as failure detectors. Technical Report (ELP-78-2/AFOSR-78-5). Urbana-Champaign, Illinois: Engineering-Psychology Research Laboratory.

Kessel, C. J. & Wickens, C. D. (1982). The transfer of failure-detection skills between-monitoring and controlling dynamic systems. *Human Factors, 24(1),* 49–60.

Kessel, K. L. (1986). *Advanced aviation (AA) in tactical aircraft may have negative effects on pilot performance,* (pp. 1248–1253). Proceedings of the 1986 IEEE International Conference on Systems, Man, and Cybernetics.

Miller, D. C. & Elkind, J. I. (1967). The adaptive response of the human controller to sudden changes in controlled process dynamics. *IEEE Transactions on Human Factors in Electronics, HFE-8 (3),* 218–223.

Parasuraman, R., Molloy, R., Mouloua, M., & Hilburn, B. (1996). Monitoring of automated systems. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: theory and applications. Mahwah, New Jersey: Lawrence Erlbaum Associates,* (pp. 91–115).

Repperger, D. W., Haas, M. W., Schley, P. C., & Koivo, A. J. (1998). *Failure detection methods to predict loss of control involving human-interface devices Part I: Theory,* (pp. 2862–2866). Proceedings of the American Control Conference.

Rolfe, J. M. (June 1972). Ergonomics & air safety. *Applied Ergonomics,* 75–81.

Rouse, W. B. (1981). Human-computer interaction in the control of dynamic systems. *Computing Surveys, 13 (1),* 71–99.

Sheridan, T. B. (1976). Preview of models of the human monitor/supervisor. In T. B. Sheridan & G. Johannsen (Eds.), *Monitoring behavior and supervisory control.* New York: Plenum.

Stewart, S. R. (September 1978). *Utility of automation of order of battle and target intelligence data for intelligence analysis* (ARI-RR-1194). Alexandria, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences.

Vreuls, D., Barneby, S. F., Nichols, D. E., & Dent, P. L. (February 1968). *Pilot failure detection performance with three levels of fault warning information,* (pp. 68–9). System Research on Development Service, FAA, SRDS RD.

Wickens, C. D. & Kessel, C. (1979). *The effects of participatory mode and task workload on the detection of dynamic system failures. IEEE Transactions on Systems, Man, and Cybernetics, SMC-9*, 24–34.

Wickens, C. D. & Kessel, C. (1980). Processing resource demands of failure detection in dynamic systems. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 564–577.

Wiener, E. L. & Curry, R. E. (1980b). Flight-deck automation: Promises and problems. *Ergonomics, 23,* 995–1011.

Young, L. R. (1969). An adaptive manual control. *IEEE Transactions on Man-Machine Systems, MMS-10 (4),* 292–331.