UNDERSTANDING PILOTS' EXPLANATIONS OF AUTOMATION SURPRISES

Robert Mauro
Decision Research & University of Oregon
Eugene, Oregon
Loukia Loukopoulou
San Jose State University Foundation
Zurich, Switzerland
Julia Trippe
Decision Research & University of Oregon
Eugene, Oregon

Automation surprise may result from inadequate or mistaken "mental models" of the automation (Sarter and Woods, 1995). To study pilots' mental models of their automation, 202 airline pilots were asked to explain five different events involving unexpected behaviors of aircraft automation. Pilots' abilities to correctly explain the behavior of the automation differed systematically across the scenarios. The number of complete and correct responses varied from 19% to 86%, depending on the scenario. As the complexity of the automation increased, understanding decreased. Performance on the scenarios was not related to flight experience, automation experience, or source of automation training. But pilots' conceptions of automation were related to performance on the scenarios. Implications for training are discussed.

In normal operations, the automated flight system of the modern airliner can control nearly all functions required for flight. This automation has greatly reduced problems due to pilot fatigue and other human frailties and has allowed more consistent and precise navigation and performance. However, automation has given rise to new problems caused by faulty interactions between the pilot and the autoflight system (AFS). This class of problems has been variously termed lack of mode awareness (Javaux & De Keyser, 1998), mode confusion (Degani, Shafto, & Kirlik, 1999), and automation surprise (Woods, Sarter, & Billings, 1997, Woods & Sarter (2000), Burki-Cohen, 2010). In these cases, the flight crew expects the automation to command one behavior and is surprised when it commands another. Automation surprise may result from: undetected failures in aircraft sensors or other systems, problematic interfaces that do not provide adequate information about the status of the aircraft (Feary et al., 1998; Norman, 1990, Degani et al., 1999), and inadequate or mistaken pilot "mental models" of the AFS (Sarter & Woods, 1995).

Automation surprises are a nuisance and a source of inefficiency in normal operations. Pilots repeatedly complain that flight management systems misbehave. Although they rarely malfunction, these systems *appear* on occasion to unilaterally decide to "drop" fixes, void altitude restrictions, or change modes of operation. In turn, these events set the stage for other errors and create problems for other aircraft and air traffic as controllers. Indeed, automation surprises have contributed to a number of airline accidents and subsequent loss of life (e.g., Reveley, Briggs, Evans, Sandifer, & Jones, 2010; Sherry & Mauro, 2014). To prevent or mitigate the effects of these "automation surprises" one must first understand why they occur. One must ask why the behavior of the autoflight system was not expected. Based on their training and experience, pilots build an understanding -- a "mental model" -- of how the AFS functions. If pilots' mental models are not completely accurate, situations may arise in which their expectations of how the AFS will behave will depart from reality.

In the research reported here, we sought to examine discrepancies between pilots' mental models of the AFS and its actual functioning by asking pilots to explain events in which a properly functioning AFS surprised the pilots. This technique of asking individuals to explain events is used to assess knowledge in education and training (Lee, Liu, & Linn, 2011) and to assess expert knowledge and decision-making (Hoffman & Lintern, 2006). By comparing the situations which yield complete and accurate explanations to those that do not, one can obtain insight into what aspects of pilots' mental models may be inaccurate and develop interventions to correct those problems.

Method

Airline pilots completed a questionnaire that asked them to explain several different events involving unexpected behaviors of aircraft automation. These events were selected in part based on prior research on pilot reports of automation surprise events (Trippe & Mauro, 2015). Instructions for responding to the scenarios were as follows:

Each of the following scenarios may or may not apply to the type of aircraft that you are presently flying. Whether or not these scenarios apply to your present aircraft, please tell us what you think caused the event described… For each of these scenarios, list as many different possible explanations as you can.

The scenarios were:

- Takeoff Default: *Facing a short runway with a hill ahead, the pilot entered a heading 10 degrees off of the runway heading into the heading window on the Mode Control Panel (MCP). Immediately after takeoff, the pilot engaged the autopilot and then immediately pressed the "HDG SEL" (Heading Select) button on the MCP to turn away from the terrain. However, the aircraft did not turn but continued to climb straight ahead. Why?*

- Altitude Capture: *Air Traffic Control (ATC) issued a clearance to a new altitude. The pilot set the altitude in the MCP "ALTITUDE" window and the aircraft began to climb. As the FMA on the PFD began flashing ACQ (Acquire), ATC issued a new altitude below the altitude that the aircraft was climbing through at that moment. The pilot set the new clearance altitude in the "ALTITUDE" window, but the aircraft failed to level off at any altitude and continued the climb. Why?*

- VNAV: *The pilot properly programmed the descent as required in the STAR and engaged the FMS. However, the aircraft failed to make the crossing restrictions. Why?*

- Runway Change: *While on the transition from the en route segment to the approach, ATC changed the expected runway. The pilot acknowledged the change and entered the change correctly into the FMC. However, the aircraft continued on its heading and failed to join the approach to the new runway. Why?*

- Localizer Intercept: *ATC provided a sharp turn to intercept the localizer. The pilot entered the heading into the MCP and armed the localizer. However, the aircraft flew through the localizer and proceeded to attempt to recapture the course from the other side. Why?*

To make the questionnaire a manageable length, two forms of the questionnaire were used. Both included the "Takeoff Default" scenario. The first questionnaire also included the "VNAV" and "Localizer Intercept" scenarios whereas the second questionnaire included the "Altitude Capture" and "Runway Change" scenarios. The questionnaires also gathered background information on flight experience (e.g., flight time, years of experience, rank, and automation experience), automation training, and perceptions of aircraft systems (navigation, electrical, FMS, crew, engine and hydraulics). The two questionnaires were distributed randomly to an equal number of pilots in each of ten discussion sessions. Thirteen to 44 pilots participated in each session.

Coding Procedure

Pilot responses to the scenarios were categorized as "Complete" if they provided an explanation that described a cause and effect relationship between the conditions given in the scenario and the described actions of the autoflight system (AFS) that could have produced the behavior described in the scenario. Responses were categorized as "Correct" if they gave an accurate cause but did not describe how that cause led the AFS to produce the described behavior. A response was categorized as a "Solution" if it suggested a fix for the situation, but did not describe why it occurred. A response was categorized as a "Non-answer" if it ignored the conditions stipulated in the scenario. Responses were categorized as "Unknown" if they were not interpretable. Finally, responses were categorized as "Wrong" if they gave an implausible cause for the automation behavior, indicated that the pilot didn't know what could have caused the behavior, or gave no response.

Subjects

Two-hundred-and-two airline pilots completed the questionnaire. Twenty-four questionnaires had too much unclear or incomplete data to be used in the analyses. One hundred and seventy-eight questionnaires were used. Of these, 97 were Commanders (CMD) and 81 were First Officers (FO). Experience ranged from 205 to 20,500 hours and 0.2 to 42 years. Usable responses came equally from the two different questionnaire Groups (89 each).

Results

There were 81 First Officers (FO) and 97 Commanders (CMD) in the sample. As expected, First Officers reported fewer total flight hours ($t(176)=16.66$, $p<.001$) and fewer years of experience ($t(175)=13.44$, $p<.001$). On average, First Officers had been flying for 7.36 years (SD 7.43); Commanders had been flying for 22.6 years (SD 7.55). First Officers reported an average of 2418 flight hours (2793); Commanders reported having an average of 10,854 flight hours (SD 3776). Both First Officers and Commanders reported approximately the same recent flight time (FO: 271 hours, SD 98; CMD: 258 hours, SD 84; $t(176)=0.886$, n.s.). The pilots reported having substantial experience with flight in automated aircraft. Commanders reported having spent 88.5% of their flight time in automated aircraft. First Officers reported having spent 83.7% of their total flight time in automated aircraft, (difference by rank: $F(1,176)=3.91$, $p=.049$). Based on their reported hours, this means that on the average Commanders reported having 9,435 hours of flight experience in automated aircraft, whereas First Officers reported having 2,080 hours of flight experience in automated aircraft.

Sources of Automation Knowledge

In general, pilots reported knowing a "moderate" or "large" amount about aircraft automation (Mean 3.30, SD 0.828). Only 4.6% (8) of the pilots reported knowing only a "small" or "very small" amount about automation. When asked where they learned what they know about automation, the pilots reported learning most about automation from initial and recurrent airline training, materials provided by the airline, and their own experience. However, there were some differences between First Officers and Commanders. First Officers reported learning significantly more than Commanders from primary training (Means: CMD: 1.22, FO: 1.68; $F(1,172)=14.60$, $p<.001$), commercial training (Means: CMD: 1.95, FO: 2.47; $F(1,172)=11.01$, $p=.001$), initial airline training (Means: CMD: 3.02, FO: 3.69; $F(1,175)=28.99$, $p<.001$), and other pilots (Means: CMD 2.48, FO: 2.97; $F(1,166)=10.75$, $p=.001$). This may indicate a shift in the content of early training as advanced automation is becoming more prevalent in training aircraft.

Conceptions of Automation

Pilots were asked how reliable, predictable, complex, and understandable (to themselves and to pilots in general) they perceived the aircraft FMS and other aircraft systems to be. In general, pilots demonstrated a strong linear trend in their perceptions of aircraft systems (Linear trend $F(1,165)=94.32$, $p<.001$, $\varepsilon^2=.364$). Hydraulic systems were perceived to be the most reliable, predictable, and understandable, followed by engine systems, electrical systems, navigation systems, and Flight Management Systems. The crew was perceived to be the least reliable, predictable, and understandable. Overall, Hydraulic and engine systems were perceived to be the least complex, but pilots varied substantially in how complex they perceived the other systems to be. In general, Commanders perceived the aircraft systems to be more reliable ($\lambda=.887$, $F(6,169)=3.603$, $p=.002$) and understandable ($\lambda=.014$, $F(6,166)=1918.17$, $p<.001$) than did First Officers. In regards to the FMS, Commanders perceived the FMS to be more reliable (Means: CMD: 5.48, FO: 5.07; $F(1,176)=7.99$, $p=.005$), predictable (Means: CMD: 5.49, FO: 5.18; $F(1,174)=4.77$, $p=.03$), and understandable (Means: CMD: 5.52, FO: 5.20; $F(1,175)=4.17$, $p=.043$) than did First Officers.

Performance on Scenarios

Though encouraged to provide as many explanations for the scenarios as possible, 85% of the pilots provided two or fewer explanations per scenario (Mean 1.5, SD .67). No matter how many explanations they produced, the pilots tended to produce explanations for individual scenarios that were either consistently reasonable or not (see Table 1). More pilots were able to generate "complete" or "correct" responses for the Localizer scenario than for the MCP or Hill scenario. In turn, more pilots were better able to generate acceptable responses for the MCP and Hill scenarios than for the Runway and VNAV scenarios. Performance on the scenarios was not related to flight experience (Rank, years flying, flight hours, recent flight hours F(4,168)=0.58, n.s.), automation experience (F(1,167)=3.26, p=.07)[1], or source of automation training (F(3,164)=1.61, n.s.).

Table 1.

Explanation Quality by Scenario: Percent of Pilots Providing "Correct" Explanations

Appropriate Explanations		Scenario				
		Localizer	Takeoff Default	Altitude Capture	Runway	VNAV
All	Count	74_b	86_a	44_a	26_c	19_c
	Percent	83.1%	48.3%	49.4%	29.2%	21.3%
None	Count	8_b	57_a	37_a	52_c	54_c
	Percent	9.0%	32.0%	41.6%	58.4%	60.7%
Some	Count	7_b	35_a	$8_{b, c}$	$11_{a, b, c}$	$16_{a, c}$
	Percent	7.9%	19.7%	9.0%	12.4%	18.0%
Total	Count	89	178	89	89	89
	Percent	100.0%	100.0%	100.0%	100.0%	100.0%

Note: Within each row, cells with different subscripts are significantly different p<.05.

Pilots' conceptions of automation were related to performance on the scenarios (Criterion: total number of responses scored complete or correct; R^2=.293, F(6,166)=11.463, p<.001). Pilots who perceived the FMS as more predictable produced better explanations of the events described in the scenarios (b=.298, t(166)=2.249, p=.026). Given that pilots tended to produce explanations for individual scenarios that were either all "complete or correct" or all not "complete or correct," the set of explanations provided by each pilot for each scenario were categorized as either all "complete or correct" or not all "complete or correct". Using this classification, the effect of pilots' conceptions of the FMS on performance on the individual scenarios was examined using logistic regression. Perceived complexity and predictability of the FMS predicted performance on the VNAV scenario. Taking into account differences in the number of explanations given, pilots who perceived the FMS to be more complex (b=0.568, Wald $X^2(1)$=4.364, p=.037) and more predictable (b=1.18, Wald $X^2(1)$=4.896, p=.027) were more likely to provide complete or correct explanations. Perceived predictability also predicted performance on the takeoff scenario (b=0.590, Wald $X^2(1)$=5.823, p=.016). No other statistically significant relations between pilots' conceptions of automation and performance on individual scenarios were observed.

Discussion

Pilots' abilities to correctly explain the behavior of the FMS differed systematically across the scenarios. As the complexity of the automation increased, understanding decreased. The scenarios may be arranged in order of increasing complexity. The Localizer Intercept scenario is the least complex and it generated the greatest number of "complete" and "correct" responses. For the AFS to intercept the localizer, lateral control coupled to the approach track is required. The logic utilized by the AFS is relatively clear to the pilot and the issues are the same for the automated flight system as they are for a pilot manually executing this maneuver. If a pilot flying manually approaches the localizer course at too great of an intercept angle, too close to the ground station, and at too great a ground speed, the aircraft will overshoot the localizer course unless the pilot deviates from standard procedures by starting the turn early or using a steeper than standard bank angle. Hence, it is easy for the pilot to understand that the AFS must confront the same issues, but may not be able to alter the rules that it was programmed to obey.

The Takeoff Default and Altitude Capture scenarios generated the next greatest number of "complete" and "correct" responses. In both of these cases, understanding the scenario requires retrieving from memory knowledge

[1] Reported automation experience was *negatively* related to overall performance on the scenarios, but this effect was not statistically significant at conventional levels.

of AFS logic for which there is no clear manual flight analog. The pilot in the Takeoff Default scenario wants to make an early turn to avoid a potential obstacle. In manual flight there is no limitation on making turns after takeoff. Airline procedures may recommend against making steep banking turns close to the ground, but ATC will routinely instruct aircraft to make early turns and fly offset departures to improve traffic flow. In the Altitude Capture scenario, the aircraft fails to respond to the pilot's command to change altitude because the AFS has transitioned into an "altitude capture" mode and is not responding to additional inputs. There is no manual flight analog for this behavior. In addition, both scenarios involve limitations that are not under the control of the flight crew. Many aircraft have a knob on the MCP that allows the flight crew to change bank limitations as required in the localizer scenario. However, in most aircraft, the altitude at which takeoff track limitations are released is neither annunciated nor adjustable. Sometimes the altitude at which an aircraft transitions to an altitude capture mode is annunciated. But because this altitude varies with the energy state of the aircraft, it is not consistent across flights and it cannot be set by the flight crew.

Pilots provided the fewest number of "complete" and "correct" responses to the Runway Change and VNAV scenarios. Both of these scenarios require knowledge of the operation of the Flight Management Computer. In the previous scenarios, targets and modes could be set directly using the MCP. In these scenarios, the pilots must interact with the AFS through the CDU. To understand the behavior of the AFS in these conditions requires that the pilot have and retrieve knowledge of the manner in which data are stored and used by the FMC and the way in which the FMC uses these data to make complex flight plan calculations. Furthermore, the results of these data manipulations depend on the specific position and energy state of the aircraft relative to the desired targets. For example, in the VNAV scenario the ability of the aircraft to make the desired crossing restriction depends on the position of the aircraft relative to the fix, the aircraft's altitude, ground speed, and energy state. The calculations performed by the FMC to determine whether the restriction will be met are hidden from the pilot. Sometimes a negative result is annunciated, but this is not always the case. To understand the problem in the Runway Change scenario, the pilot must know that approach procedures are typically built off of runways, and hence a change in runway may result in a discontinuity in the flight plan, and that when this occurs the FMC may not resolve the discontinuity and instead command the aircraft to continue on the last good heading. Furthermore, the problem will not occur if the change is made prior to a fix common to both approaches. Thus, pilots may perceive the FMC as pernicious, sometimes succeeding in make a successful runway change, sometimes not. In both of these scenarios, the data used is hidden from the pilot, making it difficult for a pilot to discern the nature of the problem from the information easily available in the cockpit.

Pilot performance on the scenarios was clearly related to the level of automation implicated in each scenario. However, on every scenario except for the Localizer Intercept scenario there was considerable variation between pilots. No measure of flight experience, automation training, or automation experience could explain this variation. There may be factors not ascertained in this study that could explain this variation. However, this result suggests that current automation training is not sufficient to ensure that pilots consistently develop a deep understanding of aircraft automation.

Manufacturers of automated systems have often suggested that automation can simplify pilots' tasks while improving precision and efficiency. However, pilots and researchers have repeatedly noted that while aviation automation has improved the efficiency and precision of operations, it has not reduced complexity. Indeed, automation may have increased the complexity of the pilot's job. Pilots often plead for manufacturers to make the automation simpler. There may be modifications to interfaces that would help simplify pilots' tasks. However, the complexity of the automation itself cannot be substantially reduced. It must be complex because the operational environment is complex and dynamic and the automation has been tasked with operating the aircraft in this environment with minimal pilot intervention.

One response to training complex automation has been to reduce the perceived complexity by limiting the training to the mechanics of executing particular procedures and limiting pilot discretion to the execution of only these procedures. However, this strategy limits pilots' understanding of the automation. When conditions arise that do not correspond to those covered by the trained procedures, the actions of the automation may surprise the pilot. Without a deeper understanding of how the automation operates, pilots cannot be expected to reliably generate explanations of the automation's behavior or to deal with it efficiently. Methods for automation education need to be developed that can help pilots build a *functional understanding* of their automation that allows them to anticipate automation actions and not simply respond with a small set of canned procedures. For pilots to construct adequate mental models of their automation, they do not need to know the intricacies of the underlying engineering, but they must understand how the system interacts with the environment – how it obtains information, what it controls, and what targets it is trying to achieve. Hence, for each automation mode, pilots must be trained to understand: 1) what is being controlled (e.g., pitch, thrust), 2) what data about the current state of the aircraft is being used (e.g., altitude from the Captain's radio altimeter, lateral position from GPS (Global Positioning System)), 3) what targets are being pursued (e.g., altitude from the MCP, speed from the FMC), and 4) what actions will be taken when the targets are achieved or fail to be achieved (e.g., proceed on heading, revert to programmed flight plan). Without this deep understanding, pilots will continue to be surprised by the automation.

References

Burki-Cohen, J. (2010). Technical challenges of upset recovery training: Simulating the element of surprise. *Proceedings of AIAA Guidance, Navigation, & Control Conference*: Toronto, Ontario, Canada, http://dx.doi.org/10.2514/6.2010-8008.

Degani, A., Shafto, M., & Kirlik, A. (1999). Modes in human-machine Systems: Constructs, representation, and classification. *International Journal of Aviation Psychology*, *9,* 125-138.

Feary, M., McCrobie, D., Alkin, M., Sherry, L., Polson, P., Palmer, E., & McQuinn, N. (1998). Aiding vertical guidance understanding. *NASA Technical Memorandum NASA/TM- 1998-112217*, Ames Research Center, Moffett Field, CMD.

Hoffman, R. & Lintern, G. (2006). Eliciting and representing the knowledge of experts. In K. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.) *Cambridge Handbook of Expertise & Expert Performance*. (pp. 203-222). New York: Cambridge University Press.

Javaux, D., & De Keyser, V. (1998). The cognitive complexity of pilot-mode interaction: A possible explanation of Sarter and Woods' classical result. In *Proceeding of the International Conference on Human-Computer Interaction in Aeronautics* (pp. 49-54). Montreal, Quebec: Ecole Polytechnique de Montreal.

Lee, H., Liu, O., & Linn, M. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, *24,* 115-136.

Norman, D. A. (1990). The 'problem' with automation: inappropriate feedback and interaction, not 'over-automation.' *Philosophical Transactions of the Royal Society B: Biological Sciences*, *327*(1241), 585-593.

Reveley, M., Briggs, J., Evans, J., Sandifer, C., & Jones, S. (2010). Causal factors and adverse conditions of aviation accidents and incidents related to integrated resilient aircraft control. *NASA Technical Memorandum NASA/TM-2010-216261.*

Sarter, N. B., & Woods, D. D. (1995). How in the world did I ever get into that mode? Mode error and awareness in supervisory control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 5-19.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. *Handbook of Human Factors and Ergonomics*, *2*, 1926-1943.

Sherry, L., & Mauro, R. (2014). Controlled Flight Into Stall (CFIS): Functional complexity failures and automation surprises. *Proceedings of the Integrated Communications Navigation and Surveillance Conference,* D1-1. IEEE.

Trippe, J. & Mauro, R. (2015). Understanding Automation Surprise: Analysis of ASRS Reports. Proceedings of the 18th International Symposium on Aviation Psychology, Dayton, Ohio.

Woods, D. & Sarter, N. (2000). Learning from automation surprises and "going sour" accidents. In Sarter, N. & Amalberti, R. (Eds.) *Cognitive Engineering in the Aviation Domain*. LEA: Mahwah, NJ.