Cognition, Technology & Work (2024) 26:267–279
https://doi.org/10.1007/s10111-024-00751-z

ORIGINAL ARTICLE

Using problem-based exploratory training to improve pilot understanding of autopilot functions

Jordy K. van Leeuwen[1] · Annemarie Landman[1,2] · Eric L. Groen[2] · Randall J. Mumaw[3] · Olaf Stroosma[1] · Marinus M. van Paassen[1] · Max Mulder[1]

Received: 10 July 2023 / Accepted: 5 January 2024 / Published online: 27 February 2024
© The Author(s) 2024

Abstract
Previous research indicated a need to improve pilot training with regard to understanding of autopilot logic and behavior, especially in non-routine situations. Therefore, we tested the effect of problem-based exploratory training on pilots' understanding of autopilot functions. Using a moving-base flight simulator, general aviation pilots ($n = 45$) were trained to diagnose failures either without foreknowledge and guidance (exploratory group), without foreknowledge but with some guidance (exploratory-guidance group) or with foreknowledge and full guidance (control group). They subsequently performed six test scenarios in which their understanding of the effects of failures was tested by requiring them to deduce the failures and select autopilot modes that were still functioning. Those who received exploratory training with guidance were significantly more likely than the other groups to diagnose failures correctly. The exploratory training group also selected the most appropriate functioning autopilot modes significantly faster than the control group. The results suggest that exploratory training with an appropriate level of guidance is useful for gaining a practical understanding of autopilot logic and behavior. Exploratory training may help to improve transfer of training to operational practice, and prevent automation surprises and accidents.

Keywords Automation surprise · Human–automation interaction · Situation awareness · Simulator training · Transfer of training

1 Introduction

Aircraft autoflight systems, encompassing the autopilot, autothrottle and the flight management system, are intended to decrease pilot workload and fatigue. Airline manufacturer policies state that "*The level of automation used shall be the most appropriate for the task at hand with regard to safety, passenger comfort, regularity and economy.*" (see, e.g., Goteman, 2018, p. 253 [Appendix]). Practically, this means that pilots hardly ever use manual control. They are automation managers, responsible for higher order goals.

Although flight deck automation has led to a significant increase in flight safety over the years (Allianz Global Corporate and Specialty 2014), aviation authorities and academia have also noted that there are issues with automation management (FAA 1996; Fletcher and Bisset 2017; Flight Deck Automation Working Group 2013; Joint Safety Implementation Team 2014; Sarter and Woods 1994; Sarter et al. 1997). One of these issues is that pilots often have an insufficient understanding of the automation's functions, logic, corollaries of actions, and of the interactions between different parts of the system such as sensors, automation logic (modes), and actuators. Between 2001 and 2007, researchers found that inadequate knowledge about the autoflight systems contributed to over 40% of accidents and 30% of serious mishaps [p. 201 (Flight Deck Automation Working Group 2013)]. Likewise, the Joint Safety Implementation Team (2014) found that automation confusion or lack of awareness contributed to 14 out of 18 investigated loss of control accidents (p. 5). Some prominent cases among these were Flash airlines 604, Colgan air 3407, Turkish airlines 1951 and West Caribbean Airways 708.

✉ Jordy K. van Leeuwen
vanleeuwen.jk@gmail.com

[1] Control and Operations Department, Delft University of Technology, Delft, The Netherlands

[2] Human Performance Department, TNO, Soesterberg, The Netherlands

[3] San José State University, San José, USA

Due to advances in technology, the number of possible automation modes and the complexity of their behavior and interactions have increased. The focus that aviation automation design puts on redundancy means that multiple pathways exist to perform the same task, each with their own dependence on other parts of the system, such as sensors, and providing varying levels of workload reduction to the pilots. The maneuvers to be performed, such as an altitude change, are well known and trained by the pilots. Multiple methods to perform such maneuvers are available, with varying reliance on on-board systems such as sensors and actuators, and varying levels of impact on pilot workload. External disturbances are mostly due to weather (wind, turbulence), which have a smaller impact on operations compared to disturbances encountered in the automotive domain.

Misunderstanding of automation logic in nominal or degraded operations may lead to mode confusion, in which a different autopilot mode is active than assumed (Flight Deck Automation Working Group 2013; Sarter and Woods 1995), or to automation surprises, in which the system behaves differently than is anticipated (Woods and Johannesen 1994). Common causes for automation surprises are indirect mode changes, automation cancelling pilot actions (de Boer and Hurts 2017; Dehais et al. 2015) or failures in sensors, automation hardware, or actuators. Automation surprise is still a common occurrence, as pilots reported experiencing about three automation surprises per year, on average (de Boer and Hurts 2017). In a survey, Holder (2013) found that 61% of pilots reported multiple issues when dealing with the automation in the first 6 months of flying their current type, and only 25% were of the opinion that they were adequately prepared.

Improvements in interface design as well as training have been recommended to mitigate this issue (Fletcher and Bisset 2017; Flight Deck Automation Working Group 2013; Sarter et al. 1997). Although training should not be used to compensate for bad design, optimizing training is a relatively time-, and cost-effective intervention for the present issues. Current pilot training focuses heavily on procedures and checklists, as these have proven to be extremely useful in time-critical, yet common, emergency situations (Degani and Wiener 1998). As a result, pilots can reproduce procedures very well, but may lack a higher order understanding and flexibility to generate solutions for *novel* or nuanced situations (Rasmussen 1983), meaning that valuable training time is possibly wasted.

Pilot resilience in dealing with autoflight failures may benefit from training that is, in part, exploratory in nature. The aim of exploratory training is to let trainees actively explore the task environment, solve authentic problems and construct knowledge through discovery. Teacher guidance can be extensive at first and reduce over time. [i.e., "scaffolding", (Vygotsky 1978)]. This contrasts with expository training, where knowledge is constructed through instruction, and with rote learning, where knowledge is constructed through memorization by repetition. Proponents of exploratory training argue that exploring and discovering solutions to problems makes the learned knowledge more meaningful. This is thought to lead to more active processing of relevant information, and better integration of the learned knowledge with existing schemata [see, Carolan et al. (2014)]. Learning through problem-solving has been used extensively in medical education, where skills of reasoning and hypothesis-testing to determine the underlying cause of observed symptoms are important (Barrows and Tamblyn 1980). In aviation, due to layers of automation, the flight system malfunctions and failures may similarly present themselves as symptoms instead of as clearly defined problems, making exploration through problem-solving a potentially effective approach to increase pilot resilience.

Evidence for the advantages of exploratory learning has been found, for instance, in the education of STEM (Hu et al. 2018; Martin et al. 2007; Ryan et al. 2004), but also for manual flying skills (Landman et al. 2018). Keith and Frese (2008) found a positive significant mean effect of exploratory training with encouragement to make errors in 24 identified studies. However, the effectiveness of exploratory training strongly depends on type of task, learner experience, and amount of guidance or freedom to explore (Carolan et al. 2014). Exploratory training with too little guidance is ineffective (Carolan et al. 2014; Mayer 2004), possibly because the training tasks require too much working memory for learning to take place (Kirschner et al. 2006). With too little guidance, trainees may simply be unable to solve even parts of the training problems, so that they do not experience and memorize any of the problem-solving rules (Mayer 2004). The training material then said to fall outside the "zone of proximal development", which is the zone between the trainee's independent capabilities and capabilities with the offered guidance (Vygotsky 1978), within which learning is hypothesized to be optimal, or alternatively involve too much cognitive load (Sweller 1994). Indeed, a meta-analysis by Carolan et al. (2014) indicated that exploratory training with little guidance is less effective than exploratory training with more guidance. No comparison was reported, however, between exploratory training with guidance and training without exploration.

Whereas training and expertise have been shown to positively affect automation management in different transportation sectors (Papadimitriou et al. 2020), the effect of different levels of exploration and guidance in training has not yet been investigated. Teachers in the aviation industry may feel hesitant to let pilots explore the system and consciously make errors in the simulator, as such behavior is not according to procedures. The goal of the current study was therefore to test whether a more exploratory form of

training indeed improves pilot understanding of automation as compared to more conventional, i.e., expository training. Insight into the structure of the system, as manipulated in the experiments by Kieras and Bovair (1984), was provided in "ground school" and kept equal between the different experimental groups.

We tested the following three hypotheses:

1. Exploratory training with limited instructor guidance is more effective than exploratory training without guidance (Carolan et al. 2014) and practicing solutions to problems without exploration. To test this, we compared performance between three training groups.
2. The effectiveness of exploratory training (with or without guidance) is highest in far transfer scenarios compared to near transfer scenarios [conform Keith and Frese (2008)]. The test scenarios therefore contained new failures and failures that were practiced in training.
3. We expected, in line with Carolan et al. (2014) that the effectiveness of exploratory learning would appear in problem-solving transfer tasks, but not in declarative knowledge, which was measured with a multiple-choice test.

2 Methods

2.1 Design

The study design was that of a semi-randomized controlled trial. Three groups received ab-initio training on autopilot functions. Two of the groups received exploratory training (see, Sect. 2.7), of which one explored freely with no guidance (Exploratory-NG group) and one received scripted guidance (Exploratory-G group). The third (Control) group, received solutions to the problems at the start of each training scenario and was not free to explore. Group performance was compared in a test immediately following the experimental training (see, Sect. 2.8). The test had two within-subject conditions: new scenarios and practiced scenarios. Performing a pre-training baseline test was not feasible as the sample group would have none to very little knowledge about autopilot functions to perform tasks. A power analysis was performed for a 3× 2 ANOVA with $\beta = 0.2$ and $\alpha = 0.05$. This indicated that a total sample size of 33 participants would be required to determine a medium-size effect (Cohens $d = 0.5$) of Group (i.e., the main focus of this study) with sufficient certainty.

2.2 Participants

Dutch general aviation pilots ($n = 45$) with an instrument rating participated in the experiment. Participants were included based on possessing an instrument rating and no type-rating. The instrument rating ensures that the participants were familiar with instrument procedures. Type-rated pilots are already highly experienced in the use of automation and were therefore excluded. Three balanced groups (Exploratory-NG, Exploratory-G and Control) of 15 participants each were formed based on the variables listed in Tables 1 and 2. Pilots' trait anxiety and fluid intelligence were measured after arrival on the experiment location using the State-Trait Anxiety Inventory (Spielberger et al. 1971) and the Raven's Advanced Progressive Matrices (John and Raven 2003), respectively. Re-assigning pilots to different groups based on these scores was not deemed necessary, as the scores did not diverge significantly (see, Table 2).

Table 1 An overview of the experience of the groups

Experience	Expl-NG	Expl-G	Control
Commercial pilot license	8/15	8/15	9/15
Glass cockpit experience	14/15	12/15	13/15
Generic autopilot experience	13/15	15/15	13/15
Type-specific avionics experience	9/15	8/15	9/15
Type-specific autopilot experience	3/15	3/15	5/15

Expl Exploratory

Table 2 Group comparison using Kruskal–Wallis test for age, flight hours, trait anxiety scores and intelligence scores

Experience	Expl-NG	Expl-G	Control	*p*
Age (years)	49.3 (24.0)	42.9 (22.9)	47.1 (23.5)	0.365
Experience (h)	1046 (1037)	862 (1190)	989 (761)	0.199
Trait anxiety (20–80)	28.4 (3.07)	28.1 (5.85)	28.3 (5.74)	0.296
Intelligence (0–12)	9.8 (2.01)	10.5 (1.15)	9.5 (1.67)	0.687

Expl Exploratory

2.3 Apparatus

The experiment was performed in the SIMONA Research Simulator at the Delft University of Technology. This is a full-motion simulator with a six-degrees-of-freedom hydraulic hexapod motion system. The simulator has a generic multi-crew flight deck including a control column (see, Fig. 1a) and rudder pedals with control loading, electrical pitch trim, throttles, and a gear lever. Outside vision is rendered with FlightGear on a collimated display with a 180° horizontal by 40° vertical field-of-view. Sound effects of (autopilot) alarms, gear retraction and wind and engine noise were played on a 5.1 surround sound system installed in the simulator. Participants did not wear a headset, but communicated via an intercom with the off-board experiment coordinator, who acted as the instructor during training.

(a) (b)

Fig. 1 **a** The experimental setup with: A. the control column, B. The Primary Flight Display, C. The Multi Function Display, D. The standby instrument, E. The throttle lever. **b** The Primary Flight Display used in the experiment

A Piper PA-34 Seneca III was simulated, a light multi-engine piston aircraft, with a non-linear, six-degrees-of-freedom software model developed by de Muynck and van Hesse (1990). The system and sensor failures used for this study were added to the model, as well as a three-axis autopilot.

Digital instruments were developed, inspired by the Primary Flight Display (PFD; see, Fig. 1b) and Multi Function Display (MFD, i.e., a moving map including a flight plan page) of the Garmin G1000. The PFD and the MFD were both presented on a 1024×768 pixels touchscreen in the simulator (see, Fig. 1a). A digital standby instrument with airspeed, attitude and altitude information was provided on a third screen.

2.4 Description of automation modes used in this study

An overview of the automation configurations (i.e., "modes") used in this study, as well as their functions, is provided in Table 3. Longitudinal modes control the aircraft pitch attitude angle, its altitude and speed. Lateral modes control the aircraft roll attitude and heading angles. As can be seen in the right-most column of Table 3, these modes can be ranked from "low level" to "high level". Higher level modes of automation take over more of the flying and navigation tasks from the pilot by using more (complex) on-board systems. Pilot workload is reduced more as a result. As an example of this, Fig. 2 shows the longitudinal automation modes (middle column), the actuators that are used (right column) and the sensors these modes depend on (left column). For full manual control, no autopilot functions are used. For manual control with flight director, the autopilot is still used to see which manual control inputs need to be made by following the flight director (FD; i.e., a flight instrument that calculates and shows the necessary attitude for the selected mode but does not execute it, see the magenta indications in Fig. 1b).

The failures used in this study for training and testing could be a failed sensor that precluded use of a high-level automation mode, a failure in the automation itself that forced the pilot to revert to manual control, or a failure in an actuator needed to effect the automation's steering commands, forcing the pilot to control the aircraft manually, possibly supported by guidance from the automation in the form of the Flight Director. In all cases the task environment did

Table 3 An overview of the autopilot modes with descriptions and level of automation (within the axis)

Axis	Automation mode	Description	Level
Lateral	Roll hold	Holds the current roll angle	Lowest
Longitudinal	Pitch hold	Holds the current pitch angle	Lowest
Lateral	Heading select	Assumes the selected heading	Low
Longitudinal	Altitude hold	Holds the current altitude	Low
Longitudinal	Vertical Speed	Assumes the selected vertical speed	Medium
Longitudinal	Altitude capture	When armed, will level off at the selected altitude	Medium
Longitudinal	Flight Level Change	Assumes the selected horizontal speed while climbing or descending to the armed altitude capture	High
Longitudinal	Vertical Navigation	Follows the vertical profile from a flight plan selected on the MFD	Highest
Lateral	VOR Navigation	Intercepts and assumes the heading of the selected outbound radial of a VOR (Very high frequency Omni-directional Range) beacon	Highest
Lateral	GPS navigation	Follows the flight plan selected on the MFD	Highest

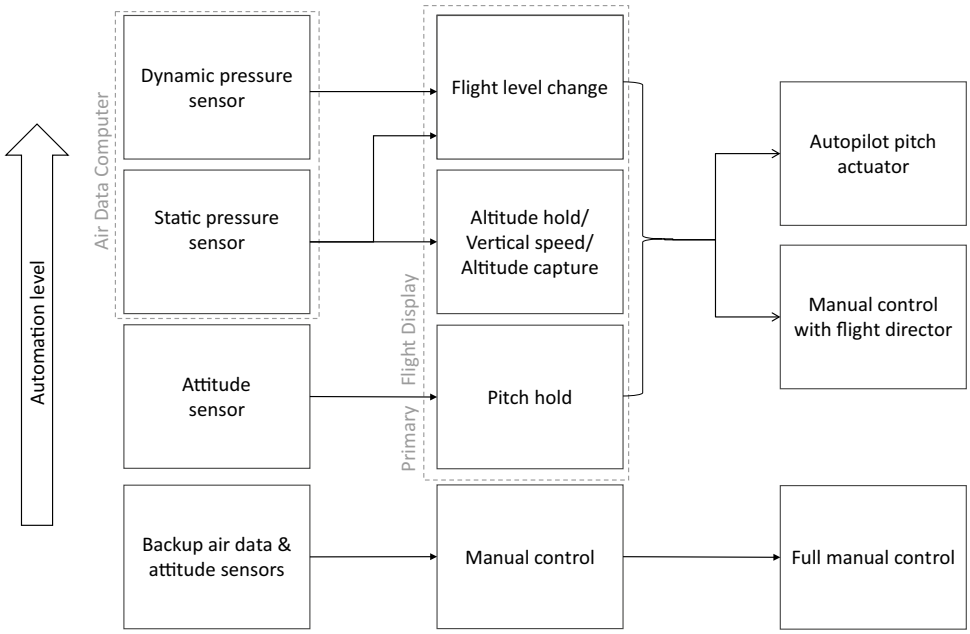

Fig. 2 Longitudinal automation overview, with sensors, automation, and actuators. Higher levels require less pilot workload, but rely on more (complex) subsystems

not change, and a viable solution was always available by selecting a different configuration of the automation.

2.5 Procedure

The procedure is illustrated in Fig. 3. The experiment started with a briefing and completion of the balancing tests (see, Sect. 2.2). Participants then received 15 min of ground school, in which the interface and basic working principles of the autopilot were explained. Pilots flew a 10-min familiarization flight in the simulator in the form of a circuit with a take-off and landing. This was followed by the automation training, consisting of five sections with a total duration of 2 h. Each section started with a “routine" scenario, in which pilots learned to use a new function of the autopilot through instructions and performing several tasks with the function. Then followed one or more “non-routine" scenarios where pilots had to respond to a failure that concerned the just-learned autopilot function. The task in the non-routine training scenarios, like in the upcoming test scenarios, was to reach each scenario’s objective by using as much of the automation’s unaffected functions as possible (i.e., the “highest functioning level of automation"). Although selecting higher modes can be very helpful in situations of high workload, it was acknowledged that this may not necessarily be the wisest approach if these simulated situations occurred in operational practice. Nevertheless, this instruction was necessary to obtain insight into the pilots’ understanding of the failure, and of its consequences for the system’s functioning. The total duration of the training was 2 h, which was the same for all training groups. A description of each scenario and the corresponding highest functioning level of automation is given in Sect. 2.7.

After the training, the participants completed the Rating Scale Mental Effort (RSME, Zijlstra and van Doorn (1985) and the Interest/Enjoyment sub-scale of the Intrinsic Motivation Inventory [IMI-IE, Ryan (1982)], and then received a 15-min break. Theoretical knowledge was then tested with a digital multiple-choice test with ten questions about the usage and working principles of the automation.

Then, participants were informed that their performance would be evaluated in six test scenarios (see, Sect. 2.8).

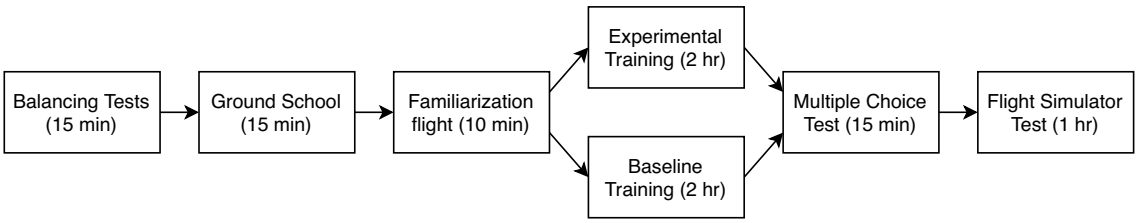

Fig. 3 Overview of the procedure

After each test scenario, participants were asked to diagnose the failure, and they rated their anxiety on an 11-point Likert-type scales ranging from 0 “not at all" to 10 “extremely" (Houtman and Bakker 1989). Surprise was rated on the same type of scale. At the end of the experiment, participants were debriefed about all scenarios.

2.6 Experimental manipulation

The difference between the groups concerned only the non-routine training scenarios. These scenarios were always performed twice in sequence. In the first trial, the Exploratory-NG and Exploratory-G groups tried to find the solution without any foreknowledge, whereas the Control group was explained beforehand what the problem was and how they should respond. Before the second trial, all groups were given this information so that all could practice the correct solution. Instructions were scripted on paper to prevent confounding effects of interactions with the instructor.

To initiate the exploration process, the Exploratory-NG and Exploratory-G groups were provided with the following instructions before the training started: 1. Notice autopilot behavior is off-nominal. 2. Identify which sensor or system is faulty. 3. Identify implications on autopilot performance. 4. Switch to alternative information source if possible, or switch to lower level automation.

However, what differentiated these two groups was that the Exploratory-G group received two hints throughout the first of each non-routine training scenario. The information presented in these hints was also provided to the other groups either after the scenario ended (Exploratory-NG) or before the scenario started (Control). The first hint, given circa 30–60 s after the failure, was developed to reduce the range of considered potential causes of the observed problem. The second hint, given either near or after the end of the scenario, was designed to reduce the considered potential solutions to the problem. An example of these hints are: 1. “Can we find out at which heading we are currently flying?” 2. “Did you have a look at the available information on the other screen?”. Pilots were then allowed for a moment to speculate and diagnose the problem, after which they were given the solution. This was intended to increase the chances that the pilots came in contact with the to-be-learned principles, that is, the systematic deduction of which system has failed based on the observed symptoms, and which consequences this has for the functioning of the automation. This contact should be a criterion for effective training (Mayer 2004).

2.7 Training scenarios

A detailed script of the training scenarios and instructions is available upon request from the authors. The training was split into three sections; Lateral, Vertical, and Navigation. Table 4 provides an overview of the scenarios and the principles the pilots were intended to learn from each scenario. All non-routine scenarios were stand-alone situations starting in stable cruise flight and consisted of a single simple task (e.g., climb and maintain 5000 ft. and fly heading 120°). Weather conditions for all training scenarios were good visibility, low turbulence and no wind.

In the first routine scenario, participants were taught how to engage and disengage the autopilot, make use of the Flight Director in manual flight, Control Wheel Steering (i.e., a button to temporarily disengage Roll and Pitch Hold so that roll and pitch can be adjusted), Roll Hold mode and Heading Select mode. These functions were covered in one continuous cruise flight. In the non-routine scenario following this routine scenario, the complete PFD failed. As illustrated in Fig. 2, this failure disables a large part of the automation and forces the pilot to control the aircraft fully manually using the backup instrumentation. In the second routine scenario, participants practiced with the Pitch Hold mode, Vertical Speed mode, Flight Level Change mode, Altitude Hold mode and Altitude capture, in one continuous cruise flight. Two non-routine scenarios followed. The first was an elevator actuator failure, which caused the autopilot to be unable to follow the selected vertical modes. It could still give the pilot guidance on the

Table 4 Non-routine training scenarios for each topic of the training, with the principles that were intended to be learned

Nr	Training topic	Failure	Principles taught
1	Lateral FD	PFD failure	One can always use the standby instrument to fall back on manual flying
2	Vertical FD	Elevator servo failure	Understanding how an issue with an actuator instead of with sensor data manifests itself and that the FD can then still be used as a reference for manual flight
3	Vertical FD	Blocked pitot tube	Deducing which sensor data are corrupted (airspeed) and that modes not using this data can still be used (Vertical Speed and lateral FD modes)
4	Navigation	VOR receiver failure	Deducing which sensor data are unavailable (VOR) and that modes not using this data (all besides VOR navigation) can still be used
5	Navigation	Air data computer failure	Deducing which sensor data are unavailable (airspeed and altitude) and which other modes can still be used (Pitch Hold and all lateral modes)

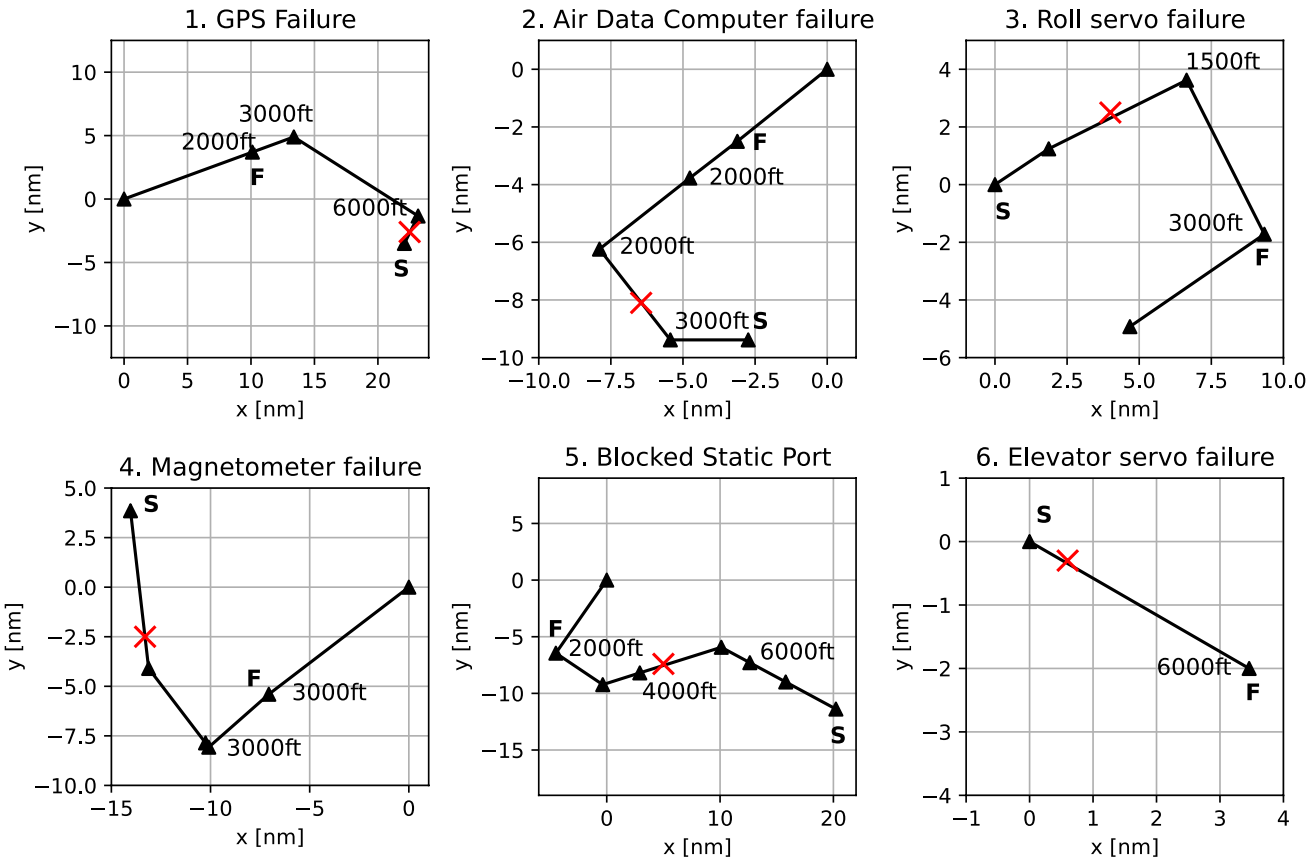

Fig. 4 Top-down view of the six test scenarios with relevant altitude restrictions. The location of the failure is indicated with a red X and the start and finish are indicated with an S and F, respectively

desired control inputs through the Flight Director, which the pilot could then effectuate manually.

The second non-routine scenario was a blocked pitot tube (dynamic pressure sensor), causing the autopilot to use corrupted airspeed data while in Flight Level Change mode. As a result of the standard pitot-static architecture found in all aircraft, indicated airspeed increases rapidly while the aircraft is climbing with a blocked pitot tube,

Table 5 The test scenarios with the corresponding automation failures and solutions

Nr	Failure	Familiarity	Highest functioning level of automation	Applicable trained principles
1	GPS failure	New	VOR Navigation with Vertical Speed, Flight Level Change and then Altitude Hold	Deducing which sensor data are unavailable (GPS) and which other modes can still be used
2	Air-Data-Computer failure	Practiced	Pitch Hold with GPS navigation	Applying the solution for non-routine training scenario 5
3	Roll servo failure	New	Manual flight with FD in Heading Select and Altitude Hold	Understanding how an issue with an actuator instead of with sensor data manifests itself and that the FD can then still be used as a reference for manual flight
4	Magneto-meter failure	New	Roll Hold with Vertical Speed and Flight Level Change, then Altitude Hold	Deducing which sensor data are unavailable (heading) and which other modes can still be used
5	Blocked Static Port	New	Pitch Hold with GPS navigation	Deducing which sensor data are unavailable (altitude) and which other modes can still be used
6	Elevator servo failure	Practiced	Manual flight with the FD in Heading Select and Altitude Hold	Applying the solution for non-routine training scenario 2

and vice-versa while descending. As illustrated in Fig. 2, this failure only disables the highest level of automation (Flight Level Change), as it relies on both dynamic and static pressure sensors. The slightly lower modes of Altitude Hold, Vertical Speed, and Altitude Capture remain available to the pilot, as do the full autopilot actuation system.

The last topic of the training, Navigation, covered the more complex lateral and vertical navigation modes of the FD. In the first routine scenario, participants were taught how to use the Navigation mode in combination with both Very high frequency Omni-directional Range (VOR) navigation aids and the Global Positioning System (GPS), and to use the course deviation indicator (i.e., an instrument which shows the lateral deviation from a selected course). In the subsequent non-routine scenario, the VOR receiver failed.

In the final routine scenario, the use of Vertical Navigation mode was taught during an Area Navigation (RNAV) approach. The accompanying non-routine scenario was an Air-Data-Computer failure, meaning that airspeed and altitude were unavailable. The pilot could now only rely on a relatively low level of automation (Pitch Hold), which only relies on the attitude sensor and not the dynamic and static pressure sensors. The autopilot actuation system still allowed them to fly the aircraft hands-off.

2.8 Test scenarios

The test consisted of six scenarios, each containing one failure which affected the behavior of the autopilot. Figure 4 shows a top-down view of the instructed flight path. Four failures were new, and two failures were repetitions of failures practiced in the non-routine training scenarios, although these occurred in a different location or phase of flight. Each scenario took place at a different location in the Netherlands, Germany or Belgium. Each test scenario covered a complete phase of flight with multiple way-points (with exception of test scenario 6 which was similar to the stand-alone training scenarios). For the scenarios that featured an approach, the appropriate approach plate was provided in paper form and available on a knee pad. Visibility was reduced to 20 km in the test scenarios, turbulence was increased to moderate and wind speeds were varied up to 10 knots. Weather information was always provided at the start of the scenario. Table 5 lists the six test scenarios with the solutions and learned principles that could be applied.

In test scenario 1, the GPS failed in the first leg of an RNAV approach without indication. This sensor failure only impacted the GPS Navigation mode. In test scenario 2, the Air Data Computer failed midway into the second leg of a new RNAV approach. This was an indicated subsystem failure impacting all the available vertical flight director modes except for Pitch Hold mode. The on-board automation automatically switched to Pitch Hold mode. In test scenario 3, the roll servo failed just after take-off, meaning that the ailerons remained at a one degree deflection. This was a non-indicated actuator failure which manifested itself by the inability of the autopilot to follow commands of the lateral FD (independent of the lateral mode selected). In test scenario 4, the magnetometer failed during an approach. This was an indicated sensor failure impacting all lateral modes except Roll Hold mode. The on-board automation automatically switched to Roll Hold mode. In test scenario 5, the static port was blocked in leg four of an RNAV approach. This was a non-indicated sensor failure which resulted in a frozen altimeter, an unreliable airspeed indicator, and the VNAV mode to never level off. This failure affected all vertical modes except the Pitch Hold mode. In test scenario 6, the elevator servo was blocked during a cruise climb. This was a non-indicated actuator failure which manifested itself by the inability of the autopilot to follow the command of the vertical FD (independent of the vertical mode selected).

2.9 Dependent measures

The state of the aircraft and pilot inputs were logged at 50 Hz for analysis. The following variables were obtained:

Problem diagnosis. After each test scenario, the participants were asked to describe what they thought the failure was. Their diagnosis was deemed correct if participants succeeded in naming either the malfunctioning sensor or system, or the limitation or abnormal functioning with the concerning axis. The proportion of correctly diagnosed scenarios was then obtained for the new scenarios and practiced scenarios.

Problem-solving time. The time in seconds between the moment of the failure and the first time the participant selected the highest level automation modes (both lateral and vertical) that were still functioning (see, Table 5). If participants never selected the highest modes, they were excluded from this measure, and reported separately. Test scenarios 2 and 4 could not be used to obtain this measure, as the on-board automation automatically switches to the highest functioning level when the failure occurs. For these two scenarios, instead the total time spent in one or more incorrect modes was taken up until the fixed scenario end at 150% of the nominal scenario length (i.e., reaching the destination without failure). Outcomes were excluded if participants never attempted a mode change in test scenarios 2 and 4, or if they never selected the highest functioning modes. These data are reported separately.

Mode changes. The total number of mode changes made during the scenario. This was averaged for the new and practiced scenarios.

Theoretical knowledge. The number of the correct answers out of the ten multiple-choice questions answered after the training. Examples of these questions are: "Which instruments rely on the Air Data Computers?" (Answer: Airspeed indicator, altimeter, vertical speed indicator). And: "Until when will the mode Flight Level Change (FLC) be active?" (Answer: Until the selected altitude is captured.)

Subjective measures. Mental effort, surprise and anxiety scores (see, Procedure) were averaged for the new and practiced test scenarios. The mental effort scores serve as an additional measure of the difficulty participants had with solving the test scenarios. Scores for surprise and anxiety serve as a manipulation checks of the test scenarios. The new failures should be more surprising than the practiced failures, and anxiety scores should ideally be high to indicate that scenarios were challenging.

2.10 Statistical analysis

The problem-solving times were first log-transformed to reduce the effect of outliers. They were then transformed into *Z*-scores, so that outcomes of different scenarios could be summed to obtain a composite score of the new test scenarios and practiced test scenarios. *Z*-scores of problem-solving time as well as scores on the rating scales were treated as ordinal measures. Problem diagnosis outcomes and Multiple choice test scores were tested for normality and treated as ordinal if not normally distributed.

For normally distributed and continuous data that were obtained in new scenarios and practiced scenarios separately, a Group (Exploratory-NG, Exploratory-G, Control) × Scenario type (New, Practiced) mixed model ANOVA was used. If data were obtained only for each group, a one-way ANOVA was used with the factor of Group. Significant effects were followed up by post-hoc pairwise comparisons, with Holm–Bonferroni correction.

Ordinal data were analyzed using Kruskal–Wallis tests for three groups. This was done separately for outcomes of new and practiced scenarios. Post-hoc pairwise comparisons of groups were performed using Mann–Whitney *U* tests with Holm–Bonferroni correction. Performance in new and practiced scenarios was compared using a Wilcoxon Signed Rank test.

3 Results

3.1 Problem diagnosis

Figure 5 shows the average proportion of problems correctly diagnosed by the groups. A Kruskal–Wallis test revealed a significant effect of group, $H = 13.06$, $p = 0.001$. A Wilcoxon Signed Rank test revealed no significant difference between new and practiced scenarios, $Z = -0.31$, $p = 0.757$. Post-hoc comparisons of groups showed that Exploratory-G, mean = 74.3%, SD = 14.3, median = 85.7%, performed significantly better than Exploratory-NG, mean = 51.4%, SD = 21.4, median = 57.1%, $U = 44.5$, $p = 0.004$, and better than Control, mean = 48.6%, SD = 21.4, median = 42.9%, $U = 37.0$, $p = 0.001$. There was no significant difference between Exploratory-NG and Control, $U = 105.0$, $p = 0.751$.

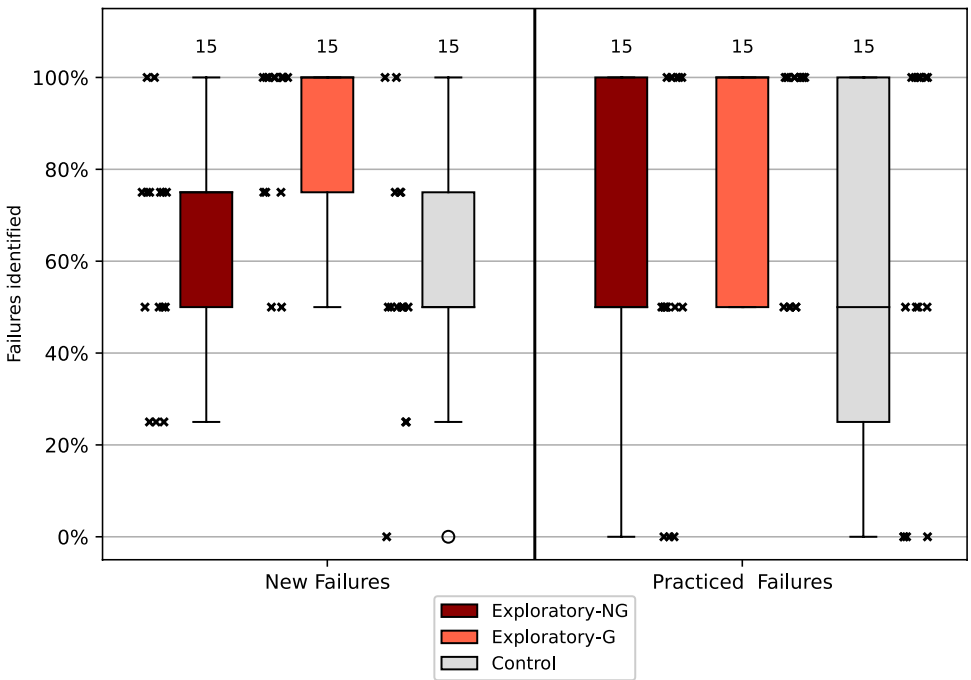

Fig. 5 Diagnostic performance for the new and practiced failures

3.2 Problem-solving time

There were too many missing cases to perform a repeated-measures analysis on new and practiced scenarios (25 valid pairs) due to pilots not finding the solutions within the allotted time or pilots not making any mode change in some scenarios. Therefore a Kruskal–Wallis test was performed for the composite Z-scores taken over all scenarios. These scores are shown in Fig. 6. There was a significant effect of Group, $H(2) = 7.50$, $p = 0.024$. Pairwise comparisons revealed no significant difference between Exploratory-G and Exploratory-NG, $p>0.999$, significantly higher performance in Exploratory-NG than Control, $p = 0.011$, and no significant difference after Holm–Bonferroni correction between Exploratory-G and Control, $p = 0.032$ (cutoff: $p = 0.025$). Exploratory-NG solved the problems on average in 124.5 s, SD = 70.2, median = 115.3, Exploratory-NG in 121.2 s, SD = 71.0, median = 92.6, and Control in 152.1 s, SD = 55.0, median = 156.0.

The median number of scenarios solved within the allotted time was 4/6 for Exploratory-NG, 5/6 for Exploratory-G, and 5/6 for Control. A Kruskal–Wallis test indicated no significant effect of Group, $H = 2.46$, $p = 0.292$. In scenario 2 and 4 (which started in the correct solution), 4/15 pilots in Exploratory-NG, 9/15 in Exploratory-G and 8/15 in Control did not attempt any mode changes. There was no significant difference between groups, $X^2(2) = 3.75$, $p = 0.153$.

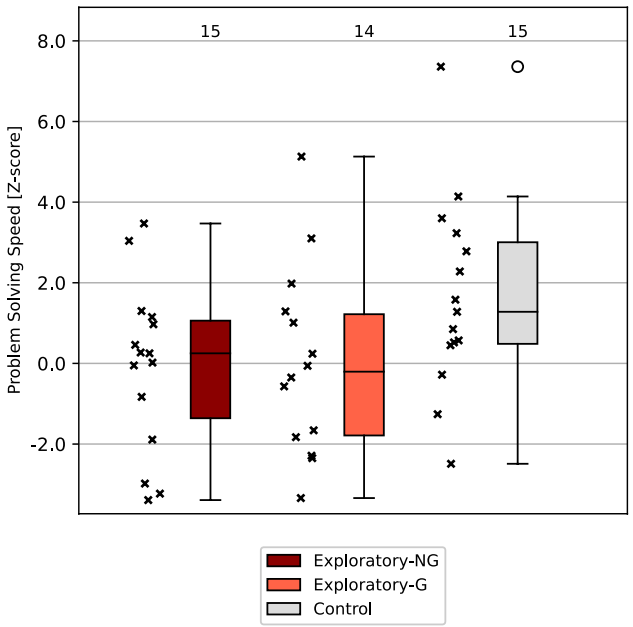

Fig. 6 Composite log-transformed Z-scores of all scenarios for the problem-solving time

3.3 Mode changes

Figure 7 presents the median number of mode changes in the new and practiced scenarios. There was no effect of Group in both the new, $H(2) = 3.21$, $p = 0.201$, and the practiced scenarios $H(2) = 0.018$, $p = 0.991$. A Wilcoxon Signed rank test indicated that significantly more mode changes occurred in the new scenarios than in the practiced scenarios, $Z = -2.33$, $p = 0.020$.

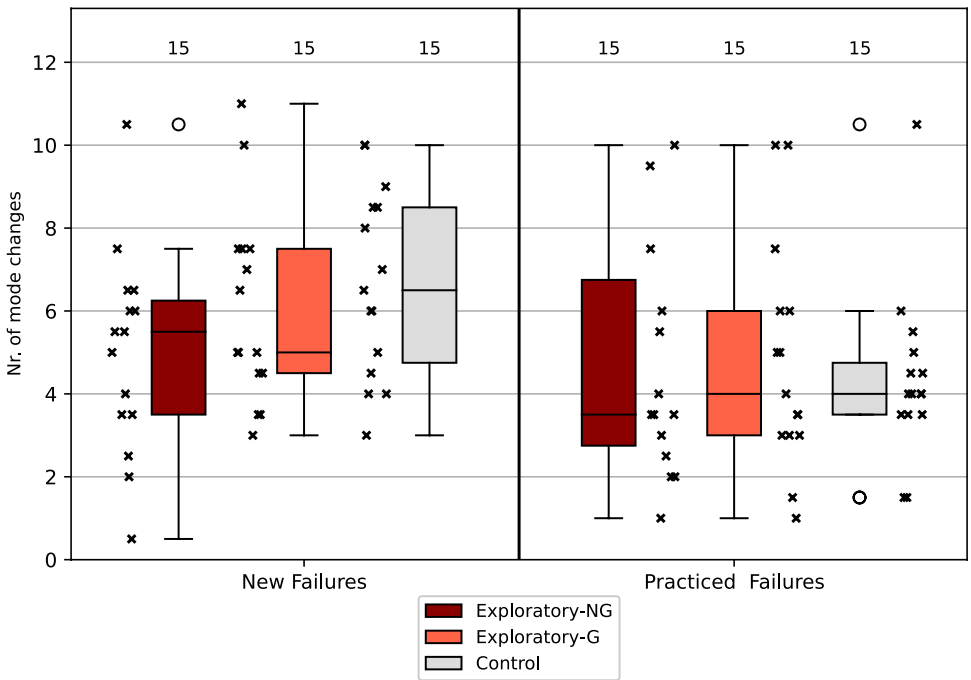

Fig. 7 Number of mode changes per scenario for new and practiced failures

3.4 Theoretical knowledge

The median score on the multiple-choice test was 7/10 for Exploratory-G and Control, and 6/10 for Exploratory-NG. A Kruskal–Wallis test revealed no significant effect of Group, $H(2) = 2.65$, $p = 0.266$.

3.5 Subjective measures

The subjective ratings of mental effort are shown in Fig. 8. There was no significant effect of Group on the mental effort ratings in the new scenarios, $H(2) = 1.15$, $p = 0.563$, nor on the practiced scenarios, $H(2) = 1.12$, $p = 0.546$. Mental effort ratings were significantly lower in the practiced scenarios, median = 40/150, than in the new scenarios, median = 73/150 $Z = -5.61$, $p<0.001$. The mental effort ratings averaged for the training scenarios are shown in Fig. 8. There was a no significant effect of Group, $H(2) = 5.71$, $p = 0.057$.

The new scenarios, median = 6.5/10, were significantly more surprising than the practiced scenarios, median = 4.5/10, $Z = -5.69$, $p<0.001$. Anxiety ratings were also significantly higher in new scenarios, median = 4.83, than in practiced scenarios, median = 2.65, $Z = -5.78$, $p<0.001$.

The median of Interest/Enjoyment ratings for the training was 44/49 for Exploratory-NG, 47/49 for Exploratory-G, and 44/49 for Control, with no significant effect of Group, $H(2) = 5.51$, $p = 0.064$.

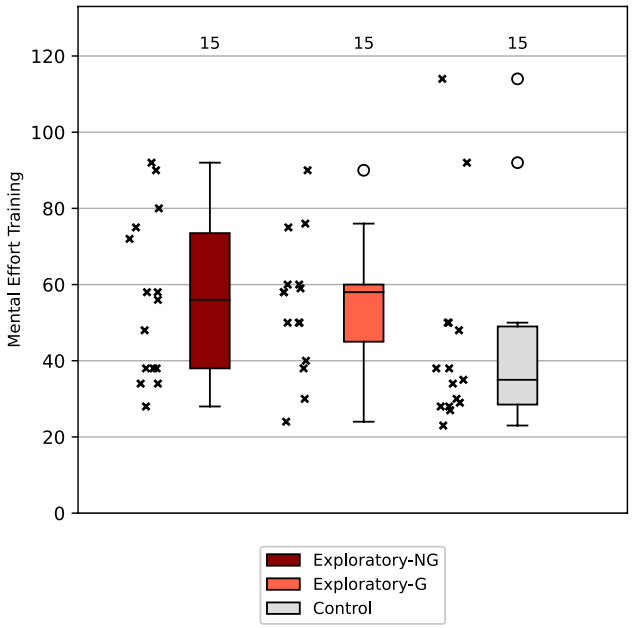

Fig. 8 Rating Scale Mental Effort scores for the training

4 Discussion

In line with Hypothesis 1, the results suggest that exploratory training with guidance led to significantly better performance in determining underlying autoflight issues in the test than those who received exploratory training without guidance and controls who received training without exploration. The time in which participants discovered solutions in the test was significantly (Exploratory-G) or nearly significantly (Exploratory-G) lower for the Exploratory training groups than controls. The number of mode changes performed and reported mental effort in the test did not differ significantly between the groups, indicating that all groups interacted to a similar extend with the autoflight system in the test.

The finding that more guidance increases the effectiveness of exploratory training is in line with a meta-analysis by Carolan et al. (2014). Adding guidance during exploratory training may have increased the chance that participants came into contact with the to-be-learned principles (Mayer 2004), instead of searching haphazardly within a large range of possible solutions. The absence of guidance may also have moved the Exploratory-NG group further away from the zone of proximal development, meaning that the training was perhaps too challenging for optimal skill acquisition (Kirschner et al. 2006; Vygotsky 1978). However, we found no evidence that mental load was too high for the Exploratory-NG group, since mental effort ratings during the training did not significantly differ between the Exploratory-NG and Exploratory-G groups. In contrast, these scores were significantly higher in both Exploratory training groups compared to the Control group, confirming that the exploratory training was more challenging, as intended. The subjective enjoyment of the training and training duration did not differ significantly between the groups, although there was a slight trend visible towards more enjoyment experienced by the exploratory training with guidance compared to others.

In contrast to Keith and Frese (2008), we found no evidence to support our Hypothesis 2, which stated that the effectiveness of exploratory training (with or without guidance) would be higher for far versus near transfer tasks. There were no specific effects on performance in new and practiced test scenarios. This could mean that the training transferred to new problems and practiced problems similarly, or that the practiced problems were not recognized from the training. The latter explanation is countered by the finding that participants found the new test scenarios significantly more surprising and mentally effortful.

Hypothesis 3, stating that the effects would appear in problem-solving tasks but not in tasks requiring declarative knowledge, was confirmed, as there were no significant differences between the groups in performance on the multiple-choice test. This result underlines that all groups

have received similar expository information during training, but it also confirms that exploratory training is not effective in increasing declarative knowledge (see, Carolan et al. (2014)).

The test scenarios appeared to be sufficiently challenging, as participants' subjective anxiety and surprise scores were around or over the mid-point of the rating scales. The scores were comparable to those in simulated mechanical failures (Landman et al. 2020). When discussing the goal of the experiment in the debriefing, most participants from all groups saw the potential benefit of using exploratory training in practice, if offered not at a too early stage of skill acquisition. Several participating certified flight instructors indicated that they were interested in implementing some form of exploratory training in their training programs.

Some limitations of the study are the limited training time, which means that conclusions about actual pilot training should be drawn with caution. The 2-h training was insufficient to train most participants to proficiency for the test scenarios, as some participants struggled to select the highest suitable level of automation even before the failure occurred, and several instances of mode confusion were observed. In reality, there is more training time available than in our study, allowing for more extensive exploratory as well as expository training. Also, the test was performed immediately after the training, whereas the long-term effects would be most interesting for operational practice. The sample group consisted of middle-aged general aviation pilots. Generalizations of the results to younger commercial pilots in initial training should be made with caution. Finally, the study was focused on measuring automation understanding, and therefore, required pilot behavior (i.e., selecting the highest suitable level of automation) that would not necessarily be the best course of action in real situations.

5 Conclusion

In conclusion, this study suggests that it is beneficial to let pilots practice with automation in an exploratory manner, after sufficient knowledge has been acquired from expository training and a foundation of basic skills is present. The study also underlines the benefits of guidance in exploratory training. In operational practice, proper guidance in a coaching and non-judgemental manner may be a prerequisite for the effectiveness of exploratory training. In our study, exploratory training transferred to situations with new automation failures, indicating that it led to more generalized problem-solving skills. Such skills are especially useful for pilots, as training time in the simulator is expensive, and the range of specific events that can be practiced is limited. Of course, not all types of automation-related accidents can be prevented with better training, as some failures are simply too complex and too difficult to detect, let alone analyze (Sherry and Mauro 2014). Nevertheless, the results of this study are promising for the benefits of exploratory training in any domain in which understanding of autonomous systems, troubleshooting, and dealing with surprising situations are important. If applied correctly, exploratory training can be an efficient addition to existing training methods to increase resilience and prevent accidents due to automation surprises.

Acknowledgements The authors would like to thank NASA cognitive psychologist Dr. D. O. Billman for crucial insights she provided on designing the experimental training.

Data availability Raw data available upon request from the corresponding author.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Allianz Global Corporate & Specialty (2014) Global aviation safety study: a review of 60 years of improvement in aviation safety (Tech. Rep.). Retrieved from https://www.agcs.allianz.com/content/dam/onemarketing/agcs/agcs/reports/AGCSGlobal-Aviation-Safety-2014-report.pdf. Accessed 19 May 2020

Barrows HS, Tamblyn RM (1980) Problem-based learning: an approach to medical education, vol 1. Springer Publishing Company, New York

Carolan TF, Hutchins SD, Wickens CD, Cumming JM (2014) Costs and benefits of more learner freedom: meta-analyses of exploratory and learner control training methods. Hum Factors 56(5):999–1014

de Boer RJ, Hurts K (2017) Automation surprise: results of a field survey of Dutch pilots. Aviat Psychol Appl Hum Factors 7(1):28–41

de Muynck R, van Hesse M (1990) The a priori simulator software package of the Piper PA34 Seneca III (Unpublished MSc. report). Delft, The Netherlands: TU Delft

Degani A, Wiener EL (1998) Design and operational aspects of flightdeck procedures. In: The International Air Transport Association (IATA) Annual Meeting. Montreal, Canada: NASA

Dehais F, Peysakhovich V, Scannella S, Fongue J, Gateau T (2015) Automation surprise in aviation. In: Proceedings of the 33rd Annual ACM Conference on human factors in computing systems—CHI 15. New York, USA: ACM Press

FAA (1996) The interfaces between flightcrews and modern flight deck systems (Tech. Rep. No. 00784270). Federal Aviation Authority

Fletcher G, Bisset G (2017) Pilot training review – final report: recommendations and conclusions. West Sussex, UK. Retrieved from https://publicapps.caa.co.uk/docs/33/CAP1581FinalReport(P).pdf. Accessed 26 Feb 2020

Flight Deck Automation Working Group (2013) Operational use of flight path management systems (Tech. Rep.). Federal Aviation Authority

Goteman Ö (2018) Automation policy or philosophy? Management of automation in the operational reality. In: Dekker SWA, Hollnagel E (eds) Coping with computers in the cockpit. Routledge, pp 215–221

Holder B (2013) Airline pilot perceptions of training effectiveness (Tech. Rep.). Seattle, WA: Boeing Commercial Airplanes

Houtman I, Bakker F (1989) The anxiety thermometer: a validation study. J Pers Assess 53(3):575–582

Hu Y-H, Xing J, Tu L-P (2018) The effect of a problem-oriented teaching method on university mathematics learning. Eurasia J Math Sci Technol Educ 14(5):1695–1703

John, Raven J (2003) Raven progressive matrices. In: McCallum RS (ed) Handbook of nonverbal assessment. Springer, Boston, MA, pp 223–237

Joint Safety Implementation Team (2014) Airplane state awareness (Tech. Rep.). Retrieved from https://skybrary.aero/bookshelf/books/3000.pdf

Keith N, Frese M (2008) Effectiveness of error management training: a meta analysis. J Appl Psychol 93(1):59–69

Kieras DE, Bovair S (1984) The role of a mental model in learning to operate a device. Cognit Sci 8(3):255–273

Kirschner PA, Sweller J, Clark RE (2006) Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educ Psychol 41(2):75–86

Landman A, van Oorschot P, van Paassen MM, Groen EL, Bronkhorst AW, Mulder M (2018) Training pilots for unexpected events: a simulator study on the advantage of unpredictable and variable scenarios. Hum Factors 60(6):793–805

Landman A, van Middelaar SH, Groen EL, van Paassen MM, Bronkhorst AW, Mulder M (2020) The effectiveness of a mnemonic-type startle and surprise management procedure for pilots. Int J Aerosp Psychol 30(3):104–118

Martin T, Rivale SD, Diller KR (2007) Comparison of student learning in challenge-based and traditional instruction in biomedical engineering. Ann Biomed Eng 35(8):1312–1323

Mayer RE (2004) Should there be a three-strikes rule against pure discovery learning? Am Psychol 59(1):14–19

Papadimitriou E, Schneider C, Tello JA, Damen W, Vrouenraets ML, Ten Broeke A (2020) Transport safety and human factors in the era of automation: What can transport modes learn from each other? Acc Anal Prev 144:105656

Rasmussen J (1983) Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. IEEE Trans Syst Man Cybern SMC13(3):257–266

Ryan RM (1982) Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. J Pers Soc Psychol 43(3):450–461

Ryan SM, Jackman JK, Peters FE, Ólafsson S, Huba ME (2004) The engineering learning portal for problem solving: experience in a large engineering economy class. Eng Econ 49(1):1–19

Sarter NB, Woods DD (1994) Pilot interaction with cockpit automation ii: an experimental study of pilot's model and awareness of the flight management system. Int J Aviat Psychol 4(1):1–28

Sarter NB, Woods DD (1995) How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Hum Factors 37(1):5–19

Sarter NB, Woods DD, Billings CE (1997) Automation surprises. In: Handbook of human factors and ergonomics, vol 2. Wiley, pp 1926–1943

Sherry L, Mauro R (2014) Functional complexity failures and automation surprises: the mysterious case of controlled flight into stall (cfis). In: 18th International Symposium on Aviation Psychology, pp 488–493

Spielberger CD, Gonzalez-Reigosa F, Martinez-Urrutia A, Natalicio LF, Natalicio DS (1971) The state-trait anxiety inventory. Interam J Psychol 5(3–4):145–158

Sweller J (1994) Cognitive load theory, learning difficulty, and instructional design. Learn Instr 4(4):295–312

Vygotsky LS (1978) Mind in society: the development of higher psychological processes. Edited by: Cole M, John-Steiner V, Scribner S, Souberman E. Harvard University Press, Cambridge

Woods DD, Johannesen LJ (1994) Behind human error: cognitive systems, computers, and hindsight (Tech. Rep.). Dayton, USA

Zijlstra FRH, van Doorn L (1985) The construction of a subjective effort scale (Report). Delft University of Technology, Dept. Social Sciences and Philosophy, Delft, The Netherlands

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.