Pilot Detection of Masked Hazards and Failures in Manual and Automated Flight

MSc Thesis

S.J. van den Eijkel

TU Delft

Pilot Detection of Masked Hazards and Failures in Manual and Automated Flight

MSc Thesis

Thesis report

by

S.J. van den Eijkel

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on May 17th, 2023 at 13:00

Thesis committee:

Chair:	Prof. Dr. Ir. M. Mulder
Supervisors:	Dr. Ir. M.M. van Paassen
	Ir. O. Stroosma
	Dr. H.M. Landman
External examiner:	Dr. A. Sharpanskykh
Place:	Faculty of Aerospace Engineering, Delft
Project Duration:	October, 2021 - May, 2023
Student number:	4351150

Faculty of Aerospace Engineering · Delft University of Technology

Preface

Before you lies my thesis titled 'Pilot Detection of Masked Hazards and Failures in Manual and Automated Flight', which marks the end of my time at TU Delft. A time that has brought me a lot, both personally and professionally, but also a time in which I often doubted if it would ever come to an end. In these moments, and perhaps especially in this MSc Thesis, it was always my passion for aviation that has helped me through.

I enjoyed the opportunity to work with motivated airline pilots and experienced supervisors on such a complex subject as our relationship with automation. I do believe that there is so much more for us to learn on this subject, especially considering the future ahead. In Aerospace, we have decades of experience with automation usage. So now that we see more automation emerging in cars, image creation and text writing, I think it is our duty to make sure that we don't fall for the same ironies that we've seen before in aviation.

This thesis taught me something I knew already, that my heart is somewhere between the clouds. It is therefore that I will start flight school in July 2023 with KLM. I hope that one day I can combine both professions and bring pilots and engineers a little closer.

I would like to thank my supervisors for their guidance and for sticking with me in this time. Annemarie, your calm words and your willingness to look at the big picture as well as the details has been of great help. I would like to express my gratitude to the twenty pilots who took the time to come to Delft and without whose cooperation, this thesis would not have been possible. On a final note I would like to thank my mom, Nico, and my grandmother. If not for you, I would never have been able to finish this. And those close around me, with whom I often shared this process, probably more often than they would have liked, you have been of tremendous help.

Simon J. van den Eijkel
Delft, 10th of May, 2023

Contents

Glossary

AC Advisory Circular 21
AFCS Automatic Flight Control System 37
ASI Airspeed Indicator 33
CA Commercial Aviation 33
DUECA Delft University Environment for Communication and Activation 40
ECAM Electronic Centralised Aircraft Monitor 33
EFIS Electronic Flight Instrument System 37
EICAS Engine Indicating and Alerting System 33
FCU Flight Control Unit 34
FD Flight Director 23, 34
FGS Flight Guidance System 21
FMA Flight Mode Announciator 34
FMS Flight Management System 34
FO First Officer 33
GA General Aviation 18, 33, 36
IMC Instrument Meteorological Conditions 22
LOC Loss of Control 20
MCP Mode Control Panel 34
NTSB National Transportation Safety Board 17
PCM Perceptual Cycle Model 29
PF Pilot Flying 18
PFD Primary Flight Display 23, 33, 34, 37
PM Pilot Monitoring 33
SA Situation Awareness 21, 29, 30
SOPs Standard Operating Procedures 33
SRS SIMONA Research Simulator 40
TEM Threat and Error Management 31
UAS Undesired Aircraft States 31
VSI Vertical Speed Indicator 29, 33

List of Figures

List of Tables

1

Introduction

Part I

Scientific Paper

Pilot Detection of Masked Hazards and Failures in Manual and Automated Flight

S.J. van den Eijkel
Control & Simulation, Aerospace Engineering

Abstract — **This study investigated the effect of manual and autopilot control on hazard and failure detection in aviation, in cases where the autopilot can masks or diminishes cues of aircraft movement. This mechanism could result in loss of situation awareness, upsets and ultimately accidents. Twenty airline pilots participated in an experiment in which two scenarios, an engine failure and icing accumulation hazard, were flown in both manual and autopilot control conditions. Results show no significant difference in detection time for the engine failure scenario between the two modes, but a marginally significant difference in the icing accumulation scenario and a clear difference in the way pilots detected the failure/hazard. Manual control may provide clearer cues, but automation may lower workload and allow for better monitoring. Monitoring of the flight controls was found to be an important factor in hazard detection, with pilots who had their hands resting on the controls having significantly lower detection times. The study provides insights on the importance of active monitoring of flight controls and on the impact of non-backdriven controls. The findings suggest that understanding the advantages and disadvantages of autopilot use is crucial for training pilots effectively.**

I. Introduction

With the introduction of automation on the flight deck, came a promise of increased economy and precision of operations (Sarter, Woods, & Billings, 1997). The autopilot would reduce workload during flight and could ultimately replace the flight crew. This promise only partially came through: The autopilot has allowed for greater efficiency and greater flexibility in flight operations, for example, because it made the job of the Flight Engineer obsolete. It has also made aviation safer than with the previous generation of aircraft (Boeing Commercial Airplane Group, 1997). But, since this introduction of automation on the flight deck the role of humans using automation has been questioned (Draper, Young, & Whitaker, 1964): A taxonomy of problems with automation on the flight deck has been created (Jensen & Rakovan, 1995) and concerns about the relation between the crew and the automation are growing, as the misuse of automation is named more often as a cause in Loss-of-Control accidents (IATA, 2014), (Snow, 2015).

The introduction of the autopilot in the 1940s relieved pilots from direct control in the pitch- and roll axis. In the 1970s, a more advanced autopilot was introduced that could do more Flight Management tasks. These tasks include flying vertical and horizontal paths, instrument landings, and management and monitoring of subsystems, making the system more complex. This successfully lowered workload in the cockpit (Ephrath & Young, 1981), (Masalonis, Duley, & Parasuraman, 1999) and allowed for better monitoring accordingly (Hancock & Williams, 1993).

But, with these innovations, the role of the pilot has also shifted to that of a monitor, exception handler and manager of automated resources, a role that humans are arguably less suited for (Bainbridge, 1983). It has also increased the mental distance between the pilot and the aircraft. The pilot is not closing the control loops him/herself, but is an observer of these control loops. This has led to the introduction of new errors and problems as multiple studies have shown that pilots have a poor understanding of auto flight modes (Active Pilot Monitoring Working Group, 2014). This could lead to mode error or mode confusion (Sarter & Woods, 1995) and to loss of vigilance or overreliance on the autopilot system (Young & Stanton, 2002), (Parasuraman, Sheridan, & Wickens, 2000), (Masalonis et al., 1999).

Some question the degree of automation to be used on the flight deck and whether or not (partial) manual flight would actually be a better mode in some phases of flight. Earlier studies have shown that adaptive task allocation (i.e., switching between manual and automatic control in cruise) is a possible way to enhance monitoring of automated systems and increase vigilance (Davies & Parasuraman, 1982). Also, without an effective feedback loop, pilots may be unaware of their monitoring habits degrading and becoming ineffective (Active Pilot Monitoring Working Group, 2014). Another study compared differences in manual and automatic control in failure detection: In compensatory tracking tasks, manual controllers had lower detection times than monitors (Ephrath & Young, 1981), this was attributed to the availability of proprioceptive feedback. These disadvantages of automation and advantages of manual control are well represented in scientific literature and on the basis of these aforementioned studies, recommendations have been made for pilot training and airline operations.

However, a topic ill-served in literature is how the automation on the flight deck could hinder the flight crew in failure and hazard detection. It is known that flight crews tend to detect unexpected automation behavior from observations of unanticipated aircraft movements such as flight path deviations, speed deviations or unexpected control movements (Abbott, Slotte, & Stimson, 1996). But when the movement of the aircraft is counteracted by the auto pilot, the crew's sense-making activities and failure detection might be hindered. In others words; with the autopilot working as the engineers have intended, it might take away vital cues that help the pilot to detect a problem, with a partial or total loss of Situation Awareness (SA) as a result. This could be when the auto flight system initially masks an in-flight upset and makes the failure much more subtle, then suddenly disengages or is unable to maintain control when it runs out of control authority.

Therefore, we investigate in the current study whether pilots detect failures later when these are partially masked by automation as compared to manual flight. This will be tested in an experiment where licensed airline pilots will fly scenarios in which such masked failures and hazards will occur. We expect that without the movement of the aircraft as a cue, detection of the failure/hazard would be delayed, perhaps until the moment the upset is unavoidable.

The following section will explore this masking mechanism and give context to the problem, before proposing an experiment.

II. Background

We will describe two aircraft accidents to illustrate the problem of the autopilot masking a failure.

Case 1: *China Airlines Flight 006*, (February 1985). A Boeing 747SP was cruising at 41,000 feet en route to Los Angeles International Airport (LAX), when the aircraft suddenly experienced a loss of power on the number 4 engine. The crew noticed the problem and tried to relight the engine, but to no avail. The airspeed slowly decreased and when the captain disengaged the autopilot, the aircraft rolled over onto its back and plunged 30,000 feet towards the Pacific ocean. During the dive the aircraft went faster than its maximum allowed speed and reached accelerations of over 5G before the crew is regained control of the aircraft. On the ground, the severity of the incident was clear. The horizontal stabilizer of the 747 was partially broken off and the landing gear doors were missing.

The National Transportation Safety Board (NTSB) concluded that the number 4 engine initially 'hung' after wear on one of the fuel lines. But that inadequate actions of the crew made matters worse. The autopilot remained engaged during the malfunction. This meant that the autopilot was actively counter-steering against the roll-moment created by the hung engine and *"effectively masked the approaching onset of the loss of control of the airplane"*. When the autopilot was disengaged, the captain had no time to adapt to the new situation and did not counter steer in time to prevent the dive. The NTSB further concluded that over-reliance on the autopilot as well as fatigue, were a major contributor to this accident (NTSB, 1985).

Case 2: *Sriwijaya Air Flight 182*, (January 2021). A Boeing 737-500 departed Soekarno–Hatta International Airport in Jakarta. The aircraft climbed through 8,100 feet in a right roll angle of about 20 degrees. At that moment, the left thrust lever started to move backwards whilst the right lever remained at climb power. This led to a thrust asymmetry with which the autopilot was unable to maintain the right turn. Subsequently, the aircraft started a slow roll to the left that remained unnoticed to the crew, until a "bank angle" alert was issued. This alert came as a total surprise to the crew, that up until this point had poorly monitored the flight path of the aircraft. This startling effect and the fact that the control column of the aircraft was fully deflected to the right to correct for the asymmetric thrust made the crew think that the aircraft was banking too much to the right. This led the PF to give full left aileron that rolled the aircraft onto its back and in towards a dive that the crew were unable to recover from. All people on board perished in the accident (KNKT, 2021).

In both these accidents the crew experienced a heavy upset as a result of a loss of situation awareness and in both cases it was a relatively small failure that was completely masked by the autopilot. This allowed the error to progress and the aircraft to move towards the edge of the control envelop. In both cases, the transition to manual control led to the upset. One could argue, that if the crew of Flight 006 was flying the aircraft manually, some cues hinting to the problem would have been more clear cut. In both cases, when the Pilot Flying (PF) would manually fly and would be steering to the right to counteract the asymmetric thrust, we could perhaps hear him/her saying something in the context of: *"I seem to be correcting this thing more and more — I wonder what's happening?"* (Norman, 1990). To understand this masking effect, it is important to understand the mechanisms that allow the autopilot to mask hazards and failures.

Monitoring Frameworks

The Sensemaking Cycle (Billman, Mumaw, & Feary, 2020), shown in Figure 1, gives an insight into this mechanism. Each of the three outside triangles represents an activity that updates the Situation Model. The sensemaking process consists of identifying monitoring questions, gathering relevant evidence, and identifying needed actions. These activities are continuously performed to update the situation model. When the autopilot steers against the natural movement of the aircraft, it can diminish or remove cues to the pilot. This means that gathering the right evidence might not be possible or that the update situation model is update with the wrong information. This could then result in longer detection times and possibly loss of SA.

A model on which the Sensemaking Cycle is based is the Perceptual Cycle Model (PCM). The PCM, originally presented by (Neisser, 1976) and later updated and validated by (Plant & Stanton, 2014), is a model that can describe the mental model and decision process of pilots during failure detection (Plant & Stanton, 2013). It introduces the concept of "Schemata"; mental templates formed by earlier experience and used to interpret information and upcoming events. An anomaly, or a cue, can trigger sensemaking and when observations do not match the active schema, modifications to this schema or the selection of alternative can be made. Masked cues might delay the activation of the correct schema.

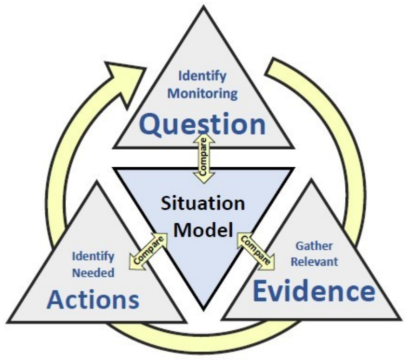


Figure 1. Situation Model and Sensemaking cycle.

Regulations

The problem of the autopilot masking cues has been recognized by airplane designers and certification bodies. In AC 25.1329-1C, which contains the acceptable means of compliance for airplane manufacturers of transport (Part 25) category airplanes, the problem is discussed:

> (2) *Masking of potential hazard.* It is not necessary that the Flight Guidance System (GS) always be disengaged when rare normal conditions that may degrade its performance or capability are encountered. The FGS may significantly help the flightcrew during such conditions. However, the design should address the potential for the FGS to mask a condition from the flightcrew or otherwise delay appropriate flightcrew action.

Furthermore, AC 23.14.19-2D shows means of compliance for flying in icing conditions for Part 23 airplanes (Normal, Utility, Acrobatic, and Commuter Category):

> The autopilot may mask tactile cues that indicate adverse changes in handling characteristics; therefore, the pilot should consider not using the autopilot when any ice is visible on the airplane.

These ACs advise against the use of the autopilot in icing conditions and recommend for a design to address other masking effects. Neither the certification standards nor the operating manuals provide the pilot with examples of what might happen when the autopilot remains engaged. They mention that the autopilot may mask 'tactile cues' but do not elaborate on what these cues might be. Neither provide the pilot with strategies to monitor the state of the aircraft in case the autopilot is engaged for longer than advised. Additionally, a study by Cole and Sand (1991), that analysed accidents in icing conditions, also suggested that pilots have a lack of understanding of the seriousness of an icing encounter on the performance of the aircraft. Therefore, is not unlikely that a pilot may find himself/herself in icing conditions with the autopilot turned on, or with an ongoing failure in the aircraft that the autopilot is successfully masking, without tactile cues or strategies to diagnose the problem. This study could provide tested examples that show the severity and consequences of this problem.

III. Methodology

A. Participants

The study involved a total of twenty participants, all of whom were licensed pilots with either an Airline Transport Pilot License (ATPL) or a Commercial Pilot License (CPL) from various airlines and varying type-ratings. All subjects participated voluntarily and were selected to ensure a homogeneous group with similar levels of proficiency in using autopilot systems. All but two participants had experience flying twin-propeller aircraft in their flight training, one participant continued flying twin-propeller aircraft and had more than 3000 hours of experience.

The average age of the group was 39.6 years with a standard deviation of 10.6 years. Nine participants were currently Captain, ten were First Officer and one was Second Officer. One participant had an active type-rating on an aircraft without backdriven controls, five more had a mixed type-rating that included aircraft without backdriven control. The distribution of their total flight hours can be found in figure 2.

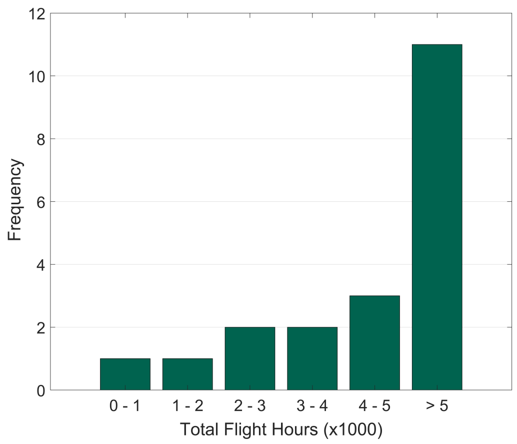


Figure 2. Total Flight Hours distribution.

B. Design

The experiment featured a within-subjects design to test whether the mode of flight control, namely fully manual or autopilot for heading and altitude, has an effect on the detection time for failures and hazards. Each participant was given a briefing and familiarisation flights after which they flew a number of scenarios that they were tasked to fly either manually or on autopilot, which was pre-determined for each scenario. They were told to stick to this instruction except for safety related concerns. They were also instructed to speak out loud as if there was a co-pilot flying with them and to call out any anomalies that they spotted. The design featured two scenarios in which a failure or hazard could be masked by the autopilot and their manual equivalent. The goal of these scenarios was to measure detection time and which cue led to the detection. Other scenarios were used to mitigate recognition and anticipation of the failures and hazards. The order of the scenarios was mixed for each participant to mitigate side-effects.

C. Apparatus

The SIMONA Research Simulator (SRS) at Delft University of Technology (Figure 3) was used to conduct the study (Stroosma, Van Paassen, &

Mulder, 2003). The SRS is a full-motion research simulator with a six-degrees-of-freedom hexapod motion system. The outside visuals are rendered with FlightGear and displayed on a collimated, 180 degrees horizontal by 40 degrees vertical field of view. A 5.1 surround sound system provided realistic 3D sound in the simulation. Participants were able to communicate with the experiment leader using a headset.

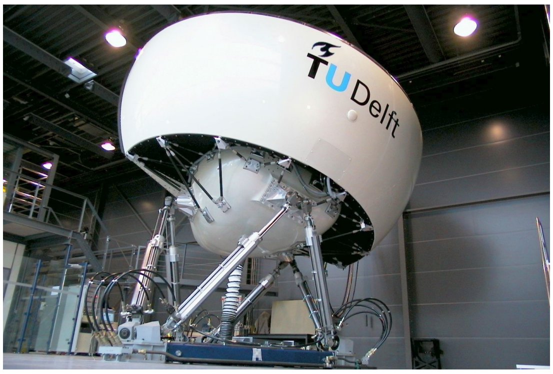


Figure 3. The Simona Research Simulator.

The aerodynamic model used in this study was a Piper Seneca III, a popular twin-engine propeller aircraft that is widely used in general aviation. The model's easy controllability allow for a quick familiarisation. The non-linear, six-degrees-of-freedom software model was originally developed by De Muynck and van Hesse (1990) and has been adapted to simulate failures by Koolstra, Herman, and Mulder (2015) and van Leeuwen (2020). For this study, the model has been enhanced with backdriven controls so that the input from the autopilot is observable to the pilot.

The Piper Seneca III was equipped with instruments similar to the Garmin G1000 avionics and an autopilot inspired by the Garmin GFC700 with several autopilot modes. The autopilot had elevator and roll control, but did not have auto throttles or rudder control. The G1000 avionics consisted of a Primary Flight Display (PFD) with speed tape, altitude tape, Horizontal Situation Indicator (HSI), Flight Mode Annunciator (FMA), Flight Path Vector (FPV) and Flight Director (FD), as well as controls for the autopilot system and an Multi-function Display (MFD) which displayed engine parameters (Manifold pressures, RPM, Cylinder Head Temperatures (CHT) etc.) as well as Outside Air Temperature (OAT) and Local Time.

Figure 4. The simulator's flight deck emulating the Piper Seneca III with Garmin avionics.

D. Procedures

Participants were given a short briefing and three familiarisation flights before they flew the experiment runs. There were a total of eight scenarios flown either in manual control or on autopilot, each lasting eight to twelve minutes. These scenarios are shown in Table 1 and will be further explained below. Each scenario began with a quick briefing on the flight state at the start of the scenario and what was expected from the participant during the scenario. The failure or hazard came unannounced and well after it was detected, the scenario was stopped. Each scenario concluded with a brief set of questions on what happened during the scenario. After the first three experiment runs, there was a 20-minute break. After the remaining eight scenarios, the session ended with a debrief in which each scenario and each failure or hazard was discussed. One session would last no longer than three hours.

Briefing

A 20 minute briefing was given at the start of the session. In this briefing, participants were told that the goal of the experiment is to evaluate the realism of the Piper Seneca III model and certain events that we can simulate with this model. This concealed that the real goal of this study is to investigate the failure and hazard detection process. This was done so that participants were not excessively primed on monitoring aircraft systems, so as to be more comparable to a real world scenario. The briefing then continued with an explanation of the aircraft systems and how to use the autopilot.

Participants were tasked to fly strictly in the mode that the experiment leader commanded at the start of each scenario. This would be either in manual control without FD, or automated in HDG and ALT mode. When using the autopilot, the participants had to control throttle settings themselves as there was no autothrottle. Participants were instructed that for safety related concerns, the autopilot could be disengaged at will.

Participants were instructed to talk as if there was a co-pilot next to them. That is to actively call out any actions they took or anomalies they spotted. These comments allowed us to determine detection times more accurately, as well as which primary cues led to the failure or hazard detection.

Familiarisation

After the briefing, each participant underwent three familiarisation flights to adjust to the Piper Seneca III aircraft. This was done to ensure that participants were comfortable and familiar with the aircraft before experiencing the hazard and failure scenarios. The first flight focused on manual control, the second flight on operating the autopilot, and in the third flight the participants experienced two failures (PFD screen turns off and double engine failure) to see if the participant would not fall silent but would continue to talk through the failure as instructed.

Failure/Hazard Scenarios

To effectively demonstrate the automation masking effect, a set of requirements was defined for the failure or hazard scenarios. These requirements were that each scenario must be an example of the automation masking effect, be realistic, be subtle, have comparable outcomes for manual and auto flight and have a detectable failure or hazard.

Based on these requirements, a one-engine out scenario was chosen. In this failure the autopilot compensates a roll angle. Icing accumulation resulting in loss of lift and elevator effectiveness was chosen to investigate a failure in which the autopilot compensates a pitch angle. These scenarios were chosen as they are realistic and can be introduced relatively gradually. The autopilot actively steers against the natural movement of the aircraft in both scenarios, hiding the fact that the aircraft is moving towards the edge of the control envelope. Other failures/hazards that were considered were a jammed elevator, trim runaway, and a mass shift in the longitudinal axis. In the lateral axis these are asymmetric flap deployment, jammed aileron, and jammed rudder. Note that the masking mechanism will be very similar for all failures/hazards.

Experiment Set-up

The experiment runs consisted of eight scenarios in total, four of which were used to test performance. These were the one-engine out scenario and icing accumulation resulting in loss of lift scenario, flown once in manual and once in automated control. These four scenarios were designed to be comparable in terms of duration, trigger, starting locations and instructions. A detailed description of these scenarios will follow below.

In addition to these failure scenarios, participants were also exposed to two scenarios in which no failure occurred and two scenarios in which a failure occurred that was not an example of the automation masking effect. Specifically, these scenarios involved a blocked pitot tube and a blocked static tube. This was done to prevent participants to be overly primed for hazard and failure detection and from recognizing patterns in the scenarios. These four scenarios were designed to make the set seem varied using different duration, trigger, starting locations and instructions. An overview of the experiment runs can be found in Table 1.

To the order of these eight scenarios were balance to mitigate order effects. So that for each participant that saw the failures on autopilot before manual control there was a participant that saw the failures on manual control first.

After each scenario, participants were asked if anything out of the ordinary had happened. If so, what they thought had happened, which cues were a first indication to this, what they did to further diagnose the problem, and finally explain which action they took to deal with the problem. They were also asked on the realism of the simulation, the flight model, and the event they experienced. These questions were designed to verify the cues that led to the detection of the failure and possibly

Table 1. Overview of experiment runs.

Mode of flight	Failure/Hazard
Autopilot	Left Engine Out
Manual	Left Engine Out
Autopilot	Icing Accumulation
Manual	Icing Accumulation
Autopilot	None
Manual	None
Autopilot	Clogged Pitot
Manual	Clogged Static

for participants to correct themselves if they did not speak up directly when the failure was noticed. The questions on the realism were asked for participants to maintain the idea that simulator realism and not failure detection was central in this study.

Left Engine Out Scenario

In the left engine out scenario, participants were instructed to fly various heading and flight level changes in the first 400 seconds of the flight until they were commanded an altitude change to 7,000 feet at 115 knots. When the aircraft passed 6,900 feet this triggered the system to start the failure exactly 30 seconds after this trigger, which made sure they were stable at 7,000 feet. At that moment, the engine power would decrease to zero in 60 seconds. The manual equivalent scenario was the same, except that the trigger would be after leveling off at 5000 feet.

This failure is immediately apparent on the MFD, where the manifold pressure drops from about 12 to 10, the lowest possible value, in about 14 seconds. Which is followed by a drop in RPM, 30 seconds into the failure. In the autopilot scenario, the control column would move from zero to 30 degrees right in 60 seconds to account for the engine malfunction. If the participant was to take no additional action to correct the engine failure, the autopilot would not be able to control the aircraft and would enter a left-hand spin 72 seconds into the failure. In manual control, no correction from the participant would lead to gradually increasing left-hand bank to 45 degrees and 10 degrees pitch down after 22 seconds.

Icing Scenario

In the icing scenario participants were instructed to climb to 7,500 feet after about 450 seconds of various heading and flight level changes. Passing 7,400 feet, the system was triggered to send the icing failure sequence to the aircraft after 40 seconds. The manual equivalent scenario was the same except for a commanded level change to 8,000 feet with the trigger at 7,900 feet. At this height, the OAT would be between -3 and 0 degrees Celsius.

When the failure was triggered the aircraft would start to slowly pitch up and lose elevator effectiveness. In autopilot, the pitch-up movement would be counteracted. This effect would gradually become greater in the 100 seconds that followed. In both scenarios, whether it was the autopilot or the participant correcting this pitch-up behaviour, the most forward position of the control column would be reached 92 seconds into the failure. Without throttling back or taking bank, the nose of the aircraft would pitch up until the aircraft would stall. Participants had no trim indicator or control surface deflection indicators available, but would be able to recognise this failure through the movement of the control column or that of the aircraft.

Dependent Measures

To investigate the effect of the use of autopilot on the failure and hazard detection process, the following variables were measured:

- *Detection Time:* The time between the start of the fault and the moment the pilot mentioned that something was wrong. This is the primary outcome measure for this study, it may help us as to understand if it was more difficult to detect a fault when flying on the autopilot. Determining the detection time was done after the experiment using all available data.
- *Primary Detection Cue:* The cue that is the trigger for the detection of the failure or hazard is of interest. Participants will be asked after each run if anything off-nominal had happened. If so, what they thought had happened and which cues were a first indication of this.

Next to these dependent measures, participants comments and remarks were recorded, particular comments of interest were also noted down manually. This to provide context to the dependent measures explained above. Additionally, when a scenario was flown manually, the participant's steering inputs and the state of the aircraft were also recorded. The control inputs also gave context to the pilot's comments and validated that the aircraft was in (partial) steady-state when the failure occurred. When the participant used the autopilot, it was noted whether or not their hands were on the controls to monitor the behavior of the autopilot.

E. Statistical Tests

In order to analyze the data collected in this study, several statistical tests was performed with SPSS (version 26). The data wwas checked for normality to start. It was expected that data did not meet the assumptions of normality. Therefore, a non-parametric Wilcoxon signed-rank test was used to compare the differences in detection times between the manual and automated flight conditions. When data for a single failure and flight mode will be analysed a Mann-Whitney test was used. The significance level (α) to reject the null-hypothesis was 0.05. A p value smaller than 0.6 is considered as a marginal significant effect.

IV. Results

A. Detection Times

Mean detection times, standard deviation and ranges for all failure/hazard scenarios are shown in Table 2. None of the failures/hazards went undetected. On average, the engine failure was detected marginally earlier when the autopilot was engaged, A Wilcoxon Signed-Ranks Test indicated that detection times were not significantly different ($Z = -.141$, $p < 0.888$)

In the icing accumulation scenario, participants were on average about eight seconds quicker in detecting the hazard when flying the aircraft manually. A Wilcoxon Signed-Ranks Test indicated that detection times for manual control were marginally significantly lower than for the autopilot ($Z = -1.894$, $p < 0.058$)

Detection times are shown in Figure 5 and Figure 6, using boxplots. Histograms and cumulative frequency plots for each failure/hazard scenario are shown in Figure 7 and Figure 8. When the Icing Accumulation scenario was flown using the autopilot, eight participants out of twenty noticed the hazard after the control column had deflected fully forward and the upset started, which was 92 seconds after the first movement of the control column.

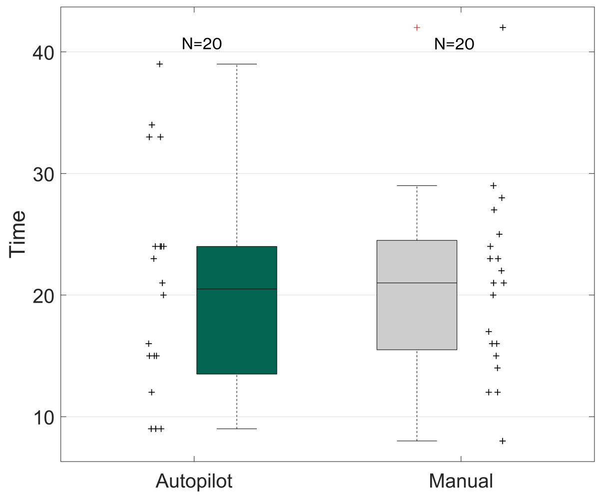


Figure 5. Boxplot of detection times for the engine failure scenario, for autopilot and manual flight.

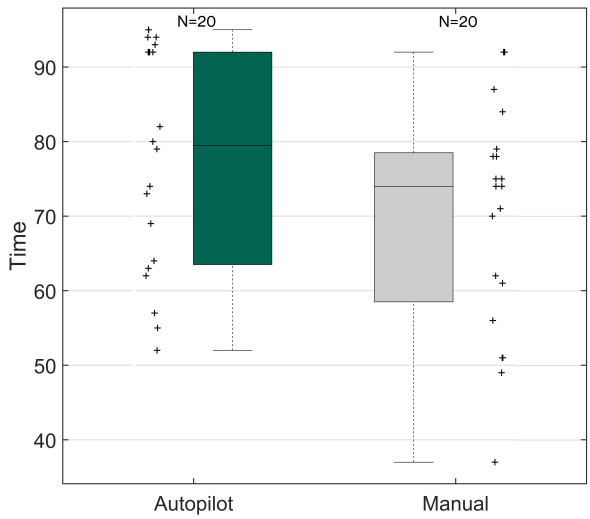
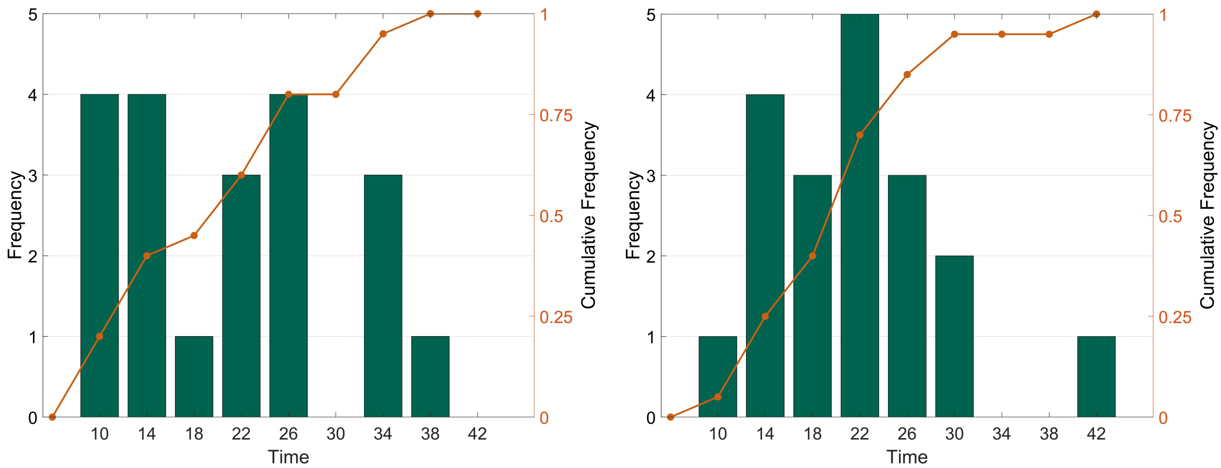


Figure 6. Boxplot of detection times for the icing accumulation scenario, for autopilot and manual flight

The results show one outlier in the detection time

Table 2. Detection time mean, median, standard deviation and range by failure/hazard type and flight mode

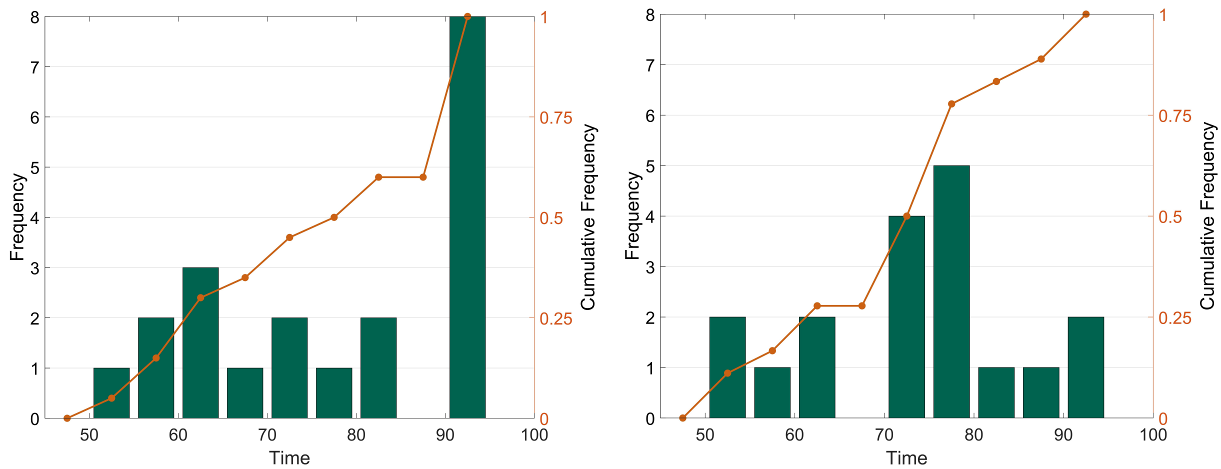
Failure/Hazard	Mode	Mean	Median	SD	Minimum	Maximum
Engine Failure	Autopilot	20.40 s	20.5 s	9.19	9 s	39 s
Engine Failure	Manual	20.75 s	21 s	7.62	8 s	42 s
Icing Accumulation	Autopilot	77.70 s	79.5 s	15.00	52 s	95 s
Icing Accumulation	Manual	69.80 s	74 s	15.10	37 s	92 s



(a) Autopilot Engine Failure

(b) Manual Engine Failure

Figure 7. Histogram and cumulative frequency plot for the engine failure scenario in (a) autopilot and (b) manual control



(a) Autopilot Icing Accumulation

(b) Manual Icing Accumulation

Figure 8. Histogram and cumulative frequency plot for the icing accumulation scenario in (a) autopilot and (b) manual control

for the manually flown engine failure. This particular participant had an active license and type-rating, but was not current as the participant had not flown for a number of years. The participant did neither produce an outlier in the autopilot flown engine failure, nor in the icing accumulation hazard scenarios.

B. Cues for Detection

The primary cues that led to detection are listed in Table 3. For the engine failure on autopilot, the drop in engine manifold pressure was named in thirteen out of twenty cases as the primary cue that led to detection, in five cases this was in combination with the speed. None of the participants named the movement of the control column as an indication of the failure. Twelve out of twenty participants had their hands on the control column at the onset of the failure.

For the engine failure in manual control, the aircraft movement or the need to correct for it was named as primary cue in thirteen out of twenty cases. In seven cases, it was the engine manifold pressure.

The icing accumulation on autopilot scenario was detected by the movement of the control column in all cases where detection happened before the upset, this was in twelve out of twenty cases. In the other eight cases, the pitch-up movement of the aircraft was the first indication that something was wrong. Monitoring the autopilot inputs with hands on the control column was done by thirteen out of twenty pilots. There was only one pilot that did not have his/her hands on the controls and noticed the problem before the upset. A Mann-Whitney U test was performed to evaluate whether these participants had lower detection times than participants who did not monitor the control column this way. Participants with their hands on the controls had a median detection time of 69 seconds (SD = 13.88). This was significantly lower as compared to a median detection time of 92 seconds (SD = 3.67) for participants who did not have their hands on the controls ($Z = -2.546$ $p = .008$). Participants without their hands on the controls were also more likely to end up in the upset.

The icing accumulation in manual control was detected in all cases by the need to trim down repeatedly. Some pilots reported putting no forces on the control column and not trimming down after a while to see what would happen.

C. Validation of scenarios

As participants were told that the realism of the simulation and in-flight events was the main research goal for this study, their feedback could be used to assess if the failures/hazard were presented realistically. None of the participants reported or showed signs of loss of immersion, motion sickness, or fatigue during the simulation. None of the participants had any issues to talk as if there was a co-pilot flying with them. Some participants did show signs of recognition of a failure/hazard when it was presented for the second time in a different flight mode: *“Here we go again with the engine"* and *“My stick is moving forward, so we might have the same problem as before"* were heard from a total of three participants.

The engine failure was correctly diagnosed by all participants before the scenario was stopped. One participants commented that the absence of warnings was unexpected, another that loss of power is accompanied with a more clear decrease in sound level. Most participants deemed the scenario a realistic slow engine failure.

The icing failure was often diagnosed as a trim runaway, jammed elevator, or change in center of gravity. Six pilots named icing as a plausible cause. The fact that the hazard was sometimes misdiagnosed does not pose an issue, as detection time is the primary outcome measure for this study and diagnosis follows after detection. Participants named the absence of a trim indicator and trim cut-out switches as a hindrance.

V. Discussion

Our findings suggest that there is a difference in primary detection cues between manual control and autopilot control, but that detection times are comparable: No significant difference was found in detection time for the engine out scenario, a marginally significant difference was found in the icing accumulation scenario. It is possible that two effects are at play: manual control giving clearer cues, but automation lowering workload and allowing for better monitoring. We see this back in comments that pilots made: *“With the autopilot it is much more subtle and you only really notice when you are in the upset.”*. When the upset finally happens, this might have a startling effect on the flight crew. Another participant said the following after the same scenario: *“Everything was*

Table 3. Primary Detection Cues for Failure/Hazard Scenarios

Failure/Hazard	Mode	Primary Cue	Ratio
Left Engine Out	Autopilot	Engine Manifold Pressure	13/20
Left Engine Out	Manual	Aircraft Movement	13/20
Icing Accumulation	Autopilot	Control Column Movement	12/20
Icing Accumulation	Manual	Pitch up tendency	20/20

hyper-stable, but still there was some forward autopilot input, which was weird". So the natural stability of the autopilot might enhance the prominence of disturbances that a failure/hazard introduces. These comments underlined advantages and disadvantages in the use of autopilot during failure detection. To use the autopilot correctly, we must understand these features and train our pilots accordingly.

The icing accumulation scenario did show a marginally significant difference in detection time, whilst the engine failure scenario did not. This might have been because the icing accumulation scenario is a better display of the masking mechanism, as there was only one cue that hinted towards this failure and the autopilot could successfully mask this cue. In contrast, in the engine failure scenario, there are many cues that point towards the failure. Also, because of the absence of autothrottles and automated rudder, even in autopilot scenarios, cues of aircraft movement such as a heading drift and yaw angle remain as cue. Therefore, the masking effect may be less dominant in this failure. Next to this, because the participants are actively closing the speed loop in the autopilot scenarios, the depletion of speed in the failure will perhaps quickly come to their attention. Further research into the masking effect should aim to better isolate this effect, perhaps by exploring fully automated flight with autothrottle and VNAV/LNAV.

In addition, the study found a difference in monitoring of the flight controls: In 25 out of 40 cases, pilots had their hands resting on the controls. Those who did had significantly lower detection time than those who did not. Pilots commented that it is the airline that recommends to keep the hands near the controls at low altitudes (below 10,000 feet). In the icing accumulation scenario, eight pilots did not detect the hazard before the upset. This was after the control column had fully deflected forward. The fact that such a primary part of the aircraft could escape the attention of the pilot for 92 seconds is interesting. Perhaps even more so considering pilots in the experiment are more focused on detecting potential failures, despite our best efforts to limit this. It raises the question if flight schools effectively train pilots to actively monitor the flight controls when flying on the autopilot. Further research into active monitoring of the flight controls could be useful, as well as the effect non-backdriven controls may have.

It is important to acknowledge that the experiment was conducted in a simulated environment, with a single-pilot and a twin-engine propeller aircraft with limited functionality. The ability to generalize the results to general practices of commercial aviation may therefore be restricted.

VI. Conclusion

This study investigated the effect of manual or automated control on detection times of hazards and failures on the flight deck. It was expected that the autopilot would reduce cues of aircraft movement and that this would lead to longer detection times as compared to manual flight. This study found that this is not necessarily true. The engine failure had comparable detection times across both modes. The icing accumulation hazard was more quickly detected when flying in manual control, but this was only a marginal significant effect.

The study did find a difference in cues that led to detection. In manual control, this would more often be cues of aircraft movement.

The findings showed a difference in pilot's monitoring behaviour. In 25 out of 40 cases, pilot's had their hands near the controls when the autopilot was on. In the icing accumulation scenario, those who did had significantly lower detection times and less often ended up in an upset situation. Eight pilots in this scenario only noticed the failure when the upset happened, 92 seconds after the start of the hazard.

This study shows that our relationship with the autopilot is complex. That the human-machine interaction between the pilot and the automated systems is one of benefits and drawbacks and that these are hard to isolate without getting interference from the other. Understanding these advantages and disadvantages of the autopilot and training our pilots accordingly, would be beneficial for flight safety.

Acknowledgments

The author would like to thank the Control and Simulation research group for the ability to conduct this study using the SIMONA Research Simulator. Furthermore, the author would like to express gratitude to befriended pilots who shared their knowledge and experience in preparation for this study. Finally, this study would not have been possible without the twenty pilots who voluntarily participated.

References

Abbott, K. H., Slotte, S., & Stimson, D. K. (1996). *The interfaces between flightcrews and modern flight deck systems* (Tech. Rep.). Federal Aviation Administration.

Active Pilot Monitoring Working Group. (2014). *A pratical guide for improving flight path monitoring* (Tech. Rep.). Flight Safety Foundation.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775-779. doi: https://doi.org/10.1016/0005-1098(83)90046-8

Billman, D., Mumaw, R., & Feary, M. (2020). A model of monitoring as sensemaking: Application to flight path management and pilot training. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 244-248. doi: 10.1177/1071181320641058

Boeing Commercial Airplane Group. (1997). *Statistical summary of commercial jet aircraft accidents: Worldwide operations* (Tech. Rep.).

Cole, J., & Sand, W. (1991). Statistical study of aircraft icing accidents. In *29th aerospace sciences meeting* (p. 558).

Davies, D., & Parasuraman, R. (1982). *The psychology of vigilance* [Book]. Academic Press London ; New York.

De Muynck, R., & van Hesse, M. V. (1990, June). *The a priori simulator software package of the piper pa34 seneca iii* (Tech. Rep.). TU Delft.

Draper, C. S., Young, L. R., & Whitaker, H. P. (1964). *The roles of man and instruments in control and guidance systems for aircraft.*

Ephrath, A. R., & Young, L. R. (1981). Monitoring vs. man-in-the-loop detection of aircraft control failures. In J. Rasmussen & W. B. Rouse (Eds.), *Human detection and diagnosis of system failures* (pp. 143–154). Boston, MA: Springer US. doi: 10.1007/978-1-4615-9230-3_10

Hancock, P., & Williams, G. (1993). Effect of task load and task load increment on performance and workload. In *Seventh international symposium on aviation psychology* (pp. 328–334).

IATA. (2014). *Loss of Control In-Flight Accident Analysis Report* (Tech. Rep.).

Jensen, R., & Rakovan, L. (1995). *Proceedings of the Eighth International Symposium on Aviation Psychology* (No. 1). Ohio State University, Department of Aerospace Engineering, Applied Mechanics and Aviation.

KNKT. (2021). *Aircraft Accident Investigation Report, PT Sriwijaya Air, Boeing 737-500, PK-CLC* (Tech. Rep.).

Koolstra, D. V. C. C., Herman, J., & Mulder, J. A. (2015). Effective model size for the prediction of the lateral controlenvelope of damaged aircraft. *AIAA Modeling and Simulation Technologies Conference*. doi: https://doi.org/10.2514/6.2015-2036

Masalonis, A. J., Duley, J. A., & Parasuraman, R. (1999). Effects of manual and autopilot control on mental workload and vigilance during simulated general aviation flight. *Transportation Human Factors*, *1*(2), 187-200. doi: 10.1207/sthf0102_7

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* New York: WH Freeman and Company.

Norman, D. (1990). The 'problem' with automation: Inappropriate feedback and interaction, not 'over-automation'. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *327*, 585-93. doi: 10.1098/rstb.1990.0101

NTSB. (1985). *Aircraft Accident Report, China Airlines Boeing 747-SP* (Tech. Rep.).

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000, 06). A model for types and levels of human interaction with automation. *IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans : a publication of the IEEE Systems, Man, and Cybernetics Society*, *30*, 286-97. doi: 10.1109/3468.844354

Plant, K., & Stanton, N. (2013, 06). What is on your mind? using the perceptual cycle model and critical decision method to understand the decision-making process in the cockpit. *Ergonomics*, *56*. doi: 10.1080/00140139.2013.809480

Plant, K., & Stanton, N. (2014, 12). The process of processing: exploring the validity of Neisser's perceptual cycle model with accounts from critical decision-making in the cockpit. *Ergonomics*, *58*, 1-15. doi: 10.1080/00140139.2014.991765

Sarter, N., & Woods, D. (1995, 03). How in the world did we ever get into that mode? mode error and awareness in supervisory control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*, 5-19. doi: 10.1518/001872095779049516

Sarter, N., Woods, D., & Billings, C. (1997, 01). Automation surprises. *Handbook of Human Factors and Ergonomics*, *2*, 1926-1943.

Snow, M. A. (2015). Preventing loss of control in flight. *Boeing Aero Magizine, No*, *57*, 9-12.

Stroosma, O., Van Paassen, M. M., & Mulder, M. (2003, 08). Using the simona research simulator for human-machine interaction research.. doi: 10.2514/6.2003-5525

van Leeuwen, J. (2020). The effectiveness of simulating active problem solving in pilot training to deal with automation surprises.

Young, M., & Stanton, N. (2002, 02). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human factors*, *44*, 365-75. doi: 10.1518/0018720024497709

Part II

Preliminary Analysis

*This part has been assessed for the course AE4020 Literature Study.

3
Introduction

It is February 19, 1985. A China Airlines Boeing 747SP is en route to Los Angeles International Airport (LAX). ‘Dynasty 006’ is cruising at 41,000 feet, 350 nautical miles northwest of San Francisco, when the aircraft suddenly experiences a loss of power on the number 4 engine. The crew notice the problem and try to relight the engine, but to no avail. The airspeed slowly decreases and when Captain Ho disengages the autopilot, the aircraft rolls over onto its back and plunges 30,000 feet towards the Pacific ocean. During the dive the aircraft goes faster than its maximum allowed speed and reaches accelerations over 5G before the crew is able to regain control of the aircraft. The crew declare an emergency and land safely at San Francisco. On the ground, the severity of the incident is clear. The horizontal stabilizer of the 747 is partially broken off and the landing gear doors are missing. How could such an advanced aircraft with an experienced crew enter a dive that could have ended disastrously?

Figure 3.1: China Airlines Flight 006

The National Transportation Safety Board (NTSB) concluded that the number 4 engine initially ‘hung’ after wear on one of the fuel lines. But that inadequate actions of the crew made matters worse. The autopilot remained engaged during the malfunction. This meant that the autopilot was actively counter steering against the roll-moment created by the hung engine and “effectively masked the approaching onset of the loss of control of the airplane”. When the autopilot was disengaged, the captain had no time to adapt to the new situation and did not counter steer in time to prevent the dive. The NTSB further concluded that over-reliance on the autopilot as well as fatigue, were a major cause of this accident (NTSB, 1985).

The accident illustrates some of the problems with automation in the cockpit. Using the Taxonomy of Flight Deck Problems and Automation Concerns (Jensen and Rakovan, 1995), we can identify the following relevant automation problems in this accident.

- Automation obscures its own state from pilots
- Automation obscures situation information from pilots

- Pilots do not perform as well when using automation
- Pilots have difficulty assuming control from automation
- Pilots are out of the control loop when they use automation
- Pilots place too much confidence in automation
- Pilots use automation when they should not
- Pilot situation awareness is reduced by automation
- Automation induces pilot fatigue

These problems show that automation is not the answer to everything and raises the questions if automation is not a hindrance, particularly in failure and hazard detection. One could argue, that if the crew of Flight 006 was flying the aircraft manually, some cues hinting to the problem would have been more clear cut. When the Pilot Flying (PF) would be steering to the right to counteract the asymmetric thurst, we could perhaps hear him/her PF saying something in the context of: “I seem to be correcting this thing more and more — I wonder what’s happening?”(Norman, 1990).
To understand what is happening here we have to understand the control loop.

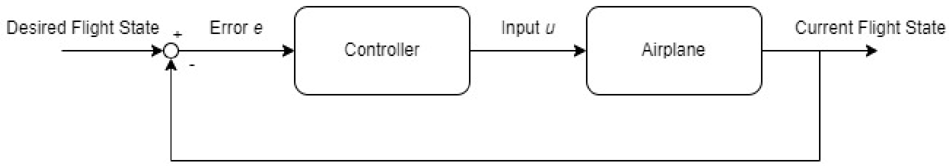


Figure 3.2: Basic Control loop Architecture

In Figure 3.2 we see a simplified version of a control loop for a controlled system. A desired flight state is compared to the actual flight state resulting in an error. This mismatch or error e serves as input to the controller, who will then generate a control input u and change the flight state. The combination of control and feedback is the control loop. When the aircraft is flying on manual control, the pilot is operating as the controller. The pilot will continuously observe the flight state and correct the airplane to its desired flight path. This is called pilot-in-the-loop. I.e. the human is an essential part of the control loop. When the aircraft is in auto flight, the autopilot will serve as the controller and close one or more control loops. This is called pilot-out-of-the-loop.

With this knowledge it can be hypothesized that if the pilot is not closing the control loop him/herself, essential cues that are used in the detection of faults and failures might be missed.
This problem might be even more prominent in General Aviation (GA). This because pilots have longer intervals between flights and training sessions and update their knowledge and improve their skills on their own. Therefore, individual abilities and judgement are critical for a safe flight (NTSB, 2014). This makes GA pilots more susceptible to automation problems.

3.1. Research Question and Objective

With this knowledge, the following **Research Question** is proposed for this study:

> ‘How can the use of the autopilot disguise the true state of an aircraft in the event of a failure or hazard?’

This question cannot be answered at once. To be able to do so, a number of sub-questions have to be answered. The first of which is ’What are possible mechanisms that allow the autopilot to mask hazards and failures?’. This report will aim to answer this question by explaining the inner workings of the autopilot and the failure and hazard detection process, as well as relevant literature. When these mechanisms have been identified, a follow-up questions would be how these mechanisms of automation masking can be demonstrated. Sub-questions that aid to answer this question are the following:

- Which observable metrics can be used to quantify the masking mechanisms of the autopilot?
- How can automation masking be isolated from other effects?
- What are requirements for experiment scenarios that demonstrate the masking effect?

The **Research Objective** is stated as:

> 'To uncover and display an undesirable characteristic of automation in aviation failure and hazard detection by means of an experimental comparison between manual and automated flight for failure detection.'

A possible result could be that automation does in fact have an effect on the pilots ability to perform failure and hazard detection. The experiment that is designed to demonstrate this effect could be an example for pilot training in the future. And should make pilots more aware of the potential masking effects of using basic automation and ways to cope with that, and to stimulate pilots to always attempt to "see through" the automation.

This preliminary thesis proposes a plan to investigate the masking effect of automation in failure and hazard detection. Chapter 4 will provide context to this problem and the masking mechanisms. It will explain previous related incidents, masking in certification and previous literature that addresses failure and hazard detection. Chapter 5 will explore failure detection from the perspective of the pilot. It will explain vigilance of flight crew, and the effects that might influence this ability. It will also address monitoring in the cockpit, with an explanation of psychological frameworks for pilot monitoring as well as current monitoring practices on the flight deck and alerting of monitoring systems. Chapter 6 will propose an experiment to demonstrate the effect of the masking mechanisms and explain the design choices for the experiment. The conclusions of the report will be summarised in Chapter 7

4

Research Context and Literature Review

This chapter aims to give a broad overview of the research topic at hand. It will discuss the implementation of automation on the flight deck. The advantages that have come with that, but also criticism and disadvantages as described by literature. It will discuss the problem of masking in more detail, current certification standards that address the issue, as well as previous studies that investigated failure and hazard detection. The chapter will conclude with a summary of findings that are useful for this study and the experiment that this study proposes.

4.1. Automation in Aviation

With the introduction of automation on the flight deck, came a great promise of increased economy and precision of operations (Sarter et al., 1997. The autopilot would lift workload during flight and could ultimately replace the flight crew. This promise only partially came through: The autopilot has allowed for greater efficiency and greater flexibility in flight operations, for example, because it made the job of the Flight Engineer obsolete. It also made aviation safer than with the previous generation of aircraft (Boeing Commercial Airplane Group, 1997. Since this introduction of automation on the flight deck the role of humans in automation has been questioned (Draper et al., 1964. But there is also more recent worry over the relation between the crew and the automation as automation is named more often as a cause in Loss of Control (LOC) (IATA, 2014,(Snow, 2015

To understand the worries over automation use, we have to understand how automation was introduced and developed. Figure 4.1 shows a timeline of automation development.

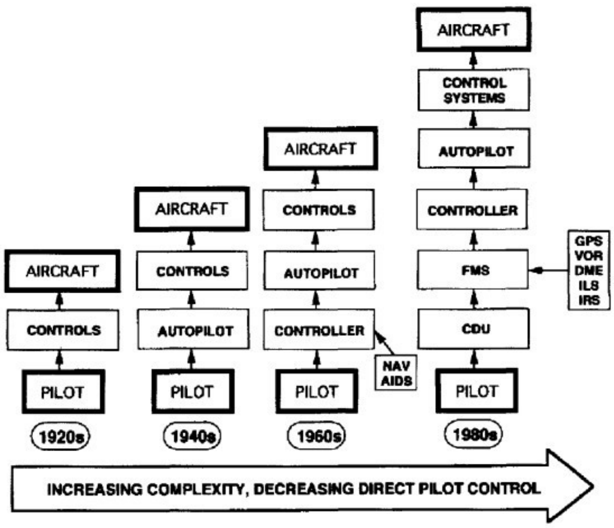


Figure 4.1: Trends in Cockpit Automation showing greater complexity and growing distance between pilot and aircraft (Billings, 1997)

In this figure we see two major breakthroughs: In the 1940s, the introduction of the basic autopilot relieved pilots from direct control in the pitch- and roll axis. In the 1970s, a more advanced autopilot was introduced that could do more Flight Management tasks. These tasks include flying Vertical/Horizontal paths, ILS landings, and management/monitoring of subsystems, making the system more complex. This graph shows the growing distance between the pilot and the aircraft. In the 1920s the pilot was simply a controller. From 1980 onwards, the role of the pilot has shifted to that of a monitor, exception handler and manager of automated resources (Bainbridge, 1983), a role that a human is arguably less suited for. This growing distance has led to the introduction of new errors and problems as multiple studies have shown that pilots have a poor understanding of auto flight modes (Active Pilot Monitoring Working Group, 2014). This could lead to mode error or mode confusion (Sarter and Woods, 1995) and to loss of vigilance or over-reliance on the autopilot system (Young and Stanton, 2002), (Parasuraman et al., 2000), (Masalonis et al., 1999). These topics are well represented in scientific literature and on the basis of these aforementioned studies, recommendations have been made for pilot training and airline operations. A topic that is represented less in literature is the masking effect of the autopilot.

4.2. The Masking Problem

This research will investigate whether or not the automation on the flight deck could hinder the flight crew in failure and hazard detection. In others words; whether or not it is possible that with the autopilot working as the engineers have intended, it might take away vital cues that help the pilot to detect a problem, with a partial or total loss of Situation Awareness (SA) as a result. This would be when the auto flight system initially masks an in-flight upset, then suddenly disengages or is unable to maintain control when it runs out of control authority. The autopilot has taken the pilot out-of-the-loop and masks the problem in a devious way.

This problem has been seen in aviation accidents before: China Airlines Flight 006 experienced a loss of power on engine number 4. The autopilot masked the severity of the problem by keeping the aircraft wings-level. The aircraft slowly approached the outskirts of the control envelope. When the crew disengaged the autopilot, the aircraft made a sharp roll to the right and lost 30,000 feet.
American Eagle Flight 4184 was operated by an ATR-72 when it had a severe icing encounter. The autopilot disconnected shortly after the ailerons deflected, an abrupt roll to the right followed that the pilots were unable to recover from (NTSB, 1994).

It is accidents and incidents like these, as well as empirical research that suggests that flight crews tend to detect unexpected automation behavior from observations of unanticipated aircraft movement. This could be flight path deviations, speed deviations or unexpected control movements (Abbott et al., 1996). But when the movement of the aircraft is counteracted by the auto pilot, the crew might remain unknowing of a problem. This mechanism, in which the autopilot takes the flight crew out of the control loop and in doing so, hinders the crew to detect failures and hazard, could have disastrous consequences. Therefore, it is clear that this phenomenon should be investigated.

4.2.1. Masking in Certification standards

The problem of the autopilot masking cues has been recognized by airplane designers and certification bodies. In Advisory Circular (AC) 25.1329-1C, which shows acceptable means of compliance for airplane manufacturers of transport (Part 25) category airplanes, the problem is discussed:

> (1) *Icing considerations.* The Flight Guidance System (FGS) performance and safety in icing conditions should be demonstrated by flight test and/or simulation tests and be supported by analysis where necessary. The implications of continued use of the automatic flight control elements of the FGS in icing conditions should be assessed. Ice accumulation on the airplane wings and surfaces can progressively change the aerodynamic characteristics and stability of the airplane. Even though the FGS may perform safely under these conditions, its continued use may mask this change, which, in turn, can lead to pilot handling difficulties and potential loss of control, should the autopilot become disengaged (either automatically or manually).
>
> (2) *Masking of potential hazard.* It is not necessary that the FGS always be disengaged when rare normal conditions that may degrade its performance or capability are encountered. The FGS may significantly help the flightcrew during such conditions. However, the design should address the potential for the FGS to mask a condition from the flightcrew or otherwise delay

appropriate flightcrew action. For discussion of alerting under such conditions, see Chapter 4, paragraph 45, FGS Alerting, Warning, Caution, Advisory, and Status, of this AC.

Furthermore, AC 23.14.19-2D shows means of compliance for flying in icing conditions for Part 23 airplanes (Normal, Utility, Acrobatic, and Commuter Category):

The autopilot may mask tactile cues that indicate adverse changes in handling characteristics; therefore, the pilot should consider not using the autopilot when any ice is visible on the airplane.

These ACs advises against the use of the autopilot in icing conditions and advises alerts for other masking effects. The discussion of alerts in certification and in practice will be discussed in Chapter 5. A similar recommendation, can be found in the operating manual of the Piper Seneca II:

Since the autopilot, when installed and operating, may mask tactile cues that indicate adverse changes in handling characteristics, use of the autopilot is prohibited when any of the visual cues specified above (cues that point to icing accumulation) exist. Or when unusual lateral trim requirements or autopilot trim warnings are encountered while the airplane is in severe icing conditions.

The certification standard and the operating manual agree on this subject; Both advise against the use of the autopilot in severe icing conditions. But in reality, it is not clear cut for the pilot to determine if these conditions persist. For this assessment, the pilot can also not fully rely on weather forecast, since icing conditions are hard to predict and prevail only locally. Furthermore, when the aircraft enters severe icing conditions, it will be tempting for the pilot to use the autopilot to make life easier. The aircraft will most likely be in Instrument Meteorological Conditions (IMC), where workload is usually higher and the autopilot can be of great help. Additionally, a study by Cole and Sand (1991), that analysed accidents in icing conditions, also suggested that pilots have a lack of understanding of the seriousness of an icing encounter on the performance of the aircraft.

Neither the certification standards nor the operating manual provide the pilot with examples of what might happen when the autopilot remains engaged, only that it could lead to LOC. They mention that autopilot may mask 'tactile cues' but do not elaborate what these cues might be. And neither provide the pilot with strategies to monitor the state of the aircraft in case the autopilot is engaged for longer than advised.

Therefore is not unlikely that a pilot may found himself/herself in icing conditions with the autopilot turned on, or with an ongoing failure in the aircraft that the autopilot is successfully masking, without tactile cues or strategies to diagnose the problem. This study could provide tested examples that show the severity and consequences of this problem.

4.3. Previous Failure Detection Studies

Previous experiments that compared manual and auto flight for failure and hazard detection have shown mixed results. In one experiment, participants were asked to perform a single-axis compensatory tracking task in which the dynamics of the controlled element would suddenly change. The human operators were divided into three groups: Active controllers, Inactive controllers and passive monitors, all were instructed to notify the experiment conductors when they would notice this change. Active controllers performed conventional compensatory tracking, in which the input was related to the output. The inactive controller commands did not have an effect on the observed error. In fact, the observed error was a replay of the error of the active controller, so they were being tricked to think that they inputs did have an effect. The passive monitor only observed the error of the active controllers. It was found that participants who had greater involvement in the control loops showed lower detection times (Ephrath and Young, 1981). Active controllers detected failures in about 1 second, the inactive controller took about 50 percent longer, the passive monitor required 3 to 5 times longer to detect the change in dynamics.

The same authors describe a similar experiment with similar results. In this experiment all participants performed three sets of compensatory tracking task runs. In the first run, the control inputs were related to the error signal and detection times were recorded. The second set started in the same way as the first set, but 2 seconds before the change in dynamics the screen would show a recorded error signal. The participants therefore became inactive controllers. In the third set, participant were asked to passively

monitor the error signal.
Results showed the superiority of the manual controller over the inactive controller and the passive monitor. The active controller detected the change within 1 second in approximately 70% of the runs. The inactive controllers and monitors required 3-5 seconds to report the change in dynamics in 70% of the runs.
The lower detection times of the manual controller were attributed to the availability of proprioceptive feedback. Though this result is valid only for single loop system failures, and does not compare to the complex nature of the flight deck, it is a good example of deteriorating detection times when the controller is taken out-of-the-loop.

In the same study, a follow-up experiment was conducted in a static simulator to investigate multi-loop failure detection. Participants were asked to fly an instrument approach from 12 miles out to touchdown in a mock-up Boeing cockpit. The experiment was conducted at four levels of control participation:

1. Monitoring, with autopilot coupling in all axes
2. Manual in the lateral axis, with autopilot coupling in the pitch axis
3. Manual in the pitch axis, with autopilot coupling in the roll axis
4. Fully manual

And three failure conditions were used:

1. No failure.
2. Failure in the lateral axis. The autopilot, if coupled, or the Flight Director (FD) steered the airplane away from the localizer course. The deviation was about 1.25 degrees 100 seconds after the failure started.
3. Failure in the pitch axis, identical in type to the lateral failure. Resulted in 0.35 degrees error 30 seconds after the failure started.

The use of the Flight Director is an important part here. The Flight Director is an overlay on the Primary Flight Display (PFD) that shows the required attitude of the aircraft to fly the desired flight path. The pilot can match the aircraft attitude with the FD to fly the desired course. The task of the pilot is therefore more like a tracking task. In the experiment, when a failure in the lateral axis occurs and the autopilot is only active in the pitch axis, the FD will show the deviation. But it is to the pilot to follow this command.

Detection times are plotted over workload in Figure **??** for manual and automatic control. It can be seen that failure detection of the passive monitors is superior to that of the active controller. The failures that went undetected were only when the failed axis was controlled manually. No failures were missed in an automatically-controlled axis. From the figure it is also clear that detection time increases with workload. It is hypothesized that in high-workload situations, the automation helps to lower the workload and the controller will have more mental capacity to monitor the state of the aircraft.

The authors of the aforementioned study published their initial findings in an earlier study (Ephrath and Curry, 1977). A study by Wickens and Kessel (1979) built upon the results of that study and attempted to verify and nuance some of its claims. This study hypothesized that the ability of the human to adapt to the dynamics, might hinder him/her in the detection of a failure. It also hypothesized that the use of proprioceptive feedback is reduced relative to visual information, particularly when both are available.
In this experiment, workload demands were experimentally manipulated, instead of assessed afterwards by letting the participants perform a critical side task next to the failure detection task. The participants were told that the side task was the loading task, allowing the detection task to fluctuate in response to the attentional resources available. The detection task consisted of a dual-axis tracking task, that was performed either actively or passively (e.g. as a monitor). It required the participant to match the position of a cursor with that of a target signal. The results showed that detection accuracy was better under automatic conditions, but that detection speed was superior in manual mode. This results was attributed to the availability of proprioceptive feedback in manual mode and is consistent with the conclusions of Ephrath and Young (1981). The author further hypothesized that ”the failure employed by Ephrath, a gradual displacement or bias of the flight path, is one for which no fundamental change in the operator’s transfer function was required to adapt. Therefore, proprioceptive channels probably conveyed little if any information relating to the occurrence of a failure . . . thereby eliminating one potential source of manual superiority.”

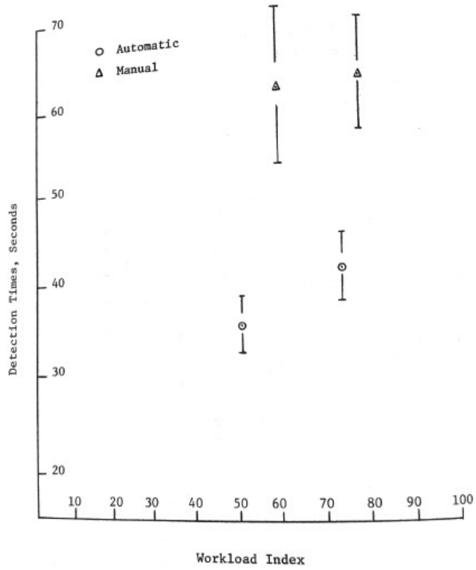


Figure 4.2: Detection Times Lateral (Yaw) Failures (Ephrath and Young, 1981)

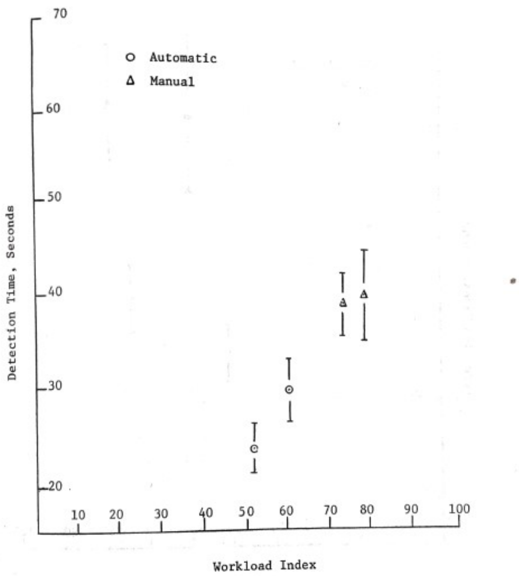


Figure 4.3: Detection Times of Longitudinal (Pitch) Failures (Ephrath and Young, 1981)

A two-part study by Beringer and Howard C. Harris (1999) investigated the response of pilot to autopilot malfunctions. Study 1 examined four failures: Runaway pitch-trim up, roll servo failure, roll sensor failure, pitch drift up. Study 2 examined four additional failures. Two that more immediately obvious, namely runaway pitch-trim down and runaway roll servo. And two subtler: Failed attitude indicator and pitch sensor drift down.
In both studies the pilots flew a Piper Malibu aircraft that was fitted with a KFC-150 autopilot and were asked to fly the whole flight using this autopilot. The failures were introduced during different phases of flight (straight and level, during descent, and during an ILS approach). The detection time, resolution strategy and related indices of performance were recorded. The results of this study can be found in Figure 4.4

This table shows the mean and median detection times for each failure and the range of these times per response category. These results show that the soft pitch malfunction was the most difficult to diagnose. Immediate disconnecters had an average of 17.7 seconds and manual overrides averaged 46.19. Three

Failure Type	Response Category	n	Response Time		Range	
			Mean	Med	Low	High
Command Roll	All (Disc)	29	16.5	8.5	1.8	107.1
	Immediate	18	5.9	5.9	1.7	11.8
	Manual Override	10	26.3	23.0	8.9	53.8
Soft Roll	Immediate	16	11.7	11.5	4.5	21.2
	Manual Override	13	37.5	26.0	13.2	85.1
Soft Pitch	Immediate	12	17.7	17.4	6.5	31.5
	Manual Override - 1	16	46.2	50.0	15.2	76.2
Pitch Trim Up	All (Disc)	25	10.5	6.9	0.2	39.2
	All (CB pull)	25	35.4	23.5	4.9	109.7
	All (CB lag)	25	25.0	15.7	0	102.3
	All (minus extremes)	23	22.7	15.7	5.1	71.3

Figure 4.4: Study 1 response time mean, median, and range by failure and response category types.

pilots never diagnosed this failure. The response time-distributions and cumulative frequency plot for this failure is shown in Figure 4.5.

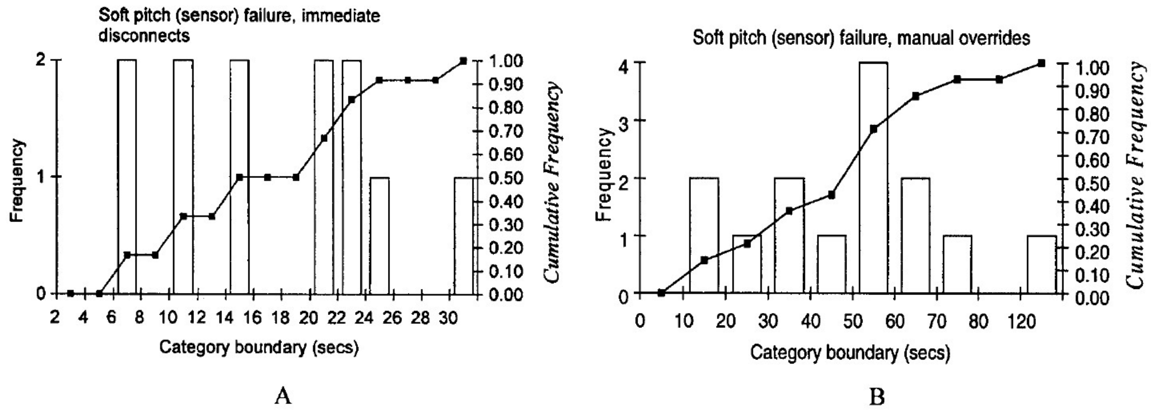


Figure 4.5: Soft-pitch (senso) response-time distributions and cumulative frequency plots for immediate disconnects (A) and manual overrides (B)

The soft-pitch and soft-roll failures are of most interest for this study, because they are akin to the failures and hazards that can be masked by the autopilot, as will be explained in Chapter 6. For both failures, a response time of over 11 seconds was found for those who disconnected the autopilot. For those who manually overrode the autopilot, both response times were over 35 seconds. Furthermore, the authors noted that some pilots employed the 'Wait-and-see' strategy. Seven pilots reported using this strategy and explained it as 'Fly through mild failures; disconnect for severe failures." this strategy is also known as 'Fly-through' or 'Diagnose then disconnect'. Such a strategy would heavily influence detection times. This is one of the reasons why it important for the experiment to consider detection time up to the moment of recognition and not of diagnosis.
More importantly, when a salient failure brings the aircraft to the edge of the control envelope in the real world, such a strategy might have disastrous consequences. Failing to immediately disconnect the autopilot and fly the aircraft manually would allow the failure to progress, a disconnect at a later stage might then be more difficult or perhaps impossible to correct.

Although the goals of these studies are similar to those of this study, there are important differences between them. Previous studies have shown how proprioceptive feedback can help manual controllers to detect failures in tracking tasks more quickly. Others have shown how an alleviation in workload can help passive monitors to detect failures more quickly in multi-loop tracking tasks. But an experiment wherein

the autopilot in a full-motion simulator works as intended, but in doing so, diminishes cues to the flight crew, has not been conducted.
Though, these studies are very useful on a practical level. They present tested scenarios and give insight how pilots will react to these failures. These aforementioned studies have detection time as their primary outcome measure. This is a useful, observable metric that can be used to quantify effects on failure and hazard detection. On an operational level, a longer detection time will allow a failure to progress, and potentially have more serious consequences. Therefore, detection time will be proposed as primary outcome measure to quantify the effect of the masking mechanisms of the autopilot.
Furthermore, these studies have presented ballpark figures for the detection time and have shown current standards in experiment design. These lessons will be taken in consideration for the design of this experiment.

This section has described the masking effect from the standpoint of aircraft manufacturers, certification bodies, and academia. It has provided this study with a number of lessons and considerations for the experiment design and shows the importance and usefulness of a study into this topic.
However, more lessons for the experiment can be drawn by considering the masking mechanisms from a pilot's perspective. The mental model of pilots and physiological effects on failure and hazard detection will be discussed in the follow section.

5

Vigilance and Monitoring

Failure detection in aviation is central in this study. The experiment that will be described in Chapter 6 will present experiment scenarios in which certain cues will be diminished or altogether removed depending on whether the pilot is flying the scenario manually or with the autopilot. Crucial in the ability of the pilot to detect failures is to remain vigilant and to observe cues by actively monitoring the systems and state of the airplane. The number and type of cues that will be available to the pilot is of key interest in this study, but to be able to single-out this variable in the results it is important to have a complete understanding of the pilots cue-management and other factors that may have an influence on this process.
This chapter aims to aid this understanding by explaining several factors that influence vigilance, lay out psychological frameworks that explain the pilot monitoring process, and dive into the current monitoring practices on the flight deck. This chapter will conclude with a summary of the findings and the consequences of this for the experiment.

5.1. Pilot Vigilance

Vigilance is an individual's ability to pay close and continuous attention to a field of stimulation for a period of time, watchful for any particular changing circumstances (Al-Shargie et al., 2019). For the experiment described in this report, this would the ability of the pilot to monitor the systems and the state of the airplane during cruise. There are a number of effects at play that influence this ability, some of which are interconnected. Although these effects might only lead to small vigilance decrement, the effects of this decrement on performance might be considerable.

This loss of vigilance can occur in any number of operating environments, but is most common when sustained attention of an extended period of time is required (Davies and Parasuraman, 1982). According to attentional resource theory, this occurs as a result of the loss of residual attention resources. Here, the rate at which these resources are depleted exceeds the rate that resupplies our system (Caggiano and Parasuraman, 2004).
This theory has been tested in aviation as well. A study by Wright and McGown (2001)) observed 12 pilots during long-haul flights. Various physiological parameters were measured such as brain electrical activity and head movement. The study concluded that after an extended period of time, performance in recognising and responding to changes in the system state degrades. In another experiment by Wiggins (2011)), participants were asked to perform an extended general aviation flight. In this flight, their ability to keep constant a number of parameters (e.g. altitude, velocity, pitch) was observed. It was found that altitude deviations became greater further into the flight. The result of this study can be found in Figure 5.1.

The authors of this study suggested that the type of task to be performed is of influence to vigilance. A comparative judgement task is where a variable is directly compared to another variable, for example when the heading bug is matched to the actual heading. A memory retrieval task is where a variable has to be monitored that can't be compared. For example an altitude that was given by ATC a few moments ago. It is the latter that imposes a relatively greater demand on attentional resources. This proposition was extended by See et al. (1995) suggesting that cognitive demand is greatest when there is a combination of memory load and a relatively high event rate.

Workload is another important variable for vigilance. Work overload can cause attentional resources to drain quickly and lead to lowered vigilance. A study by Wickens and Kessel (1979) has shown that

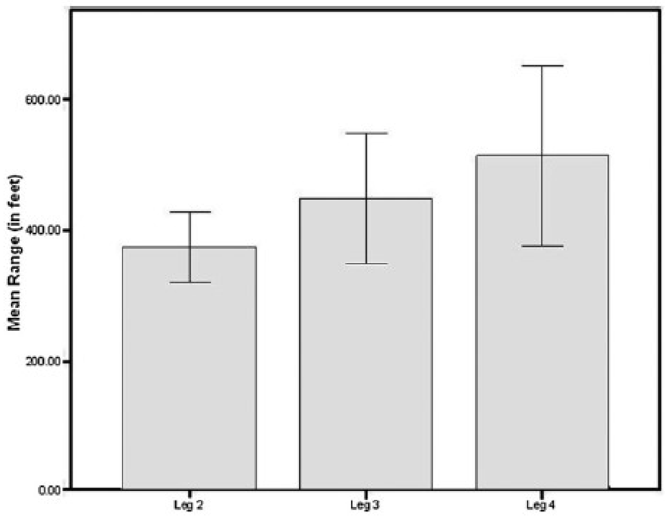


Figure 5.1: Mean absolute deviation in altitude of the aircraft over three legs in flight

operator workload affects detection of failures adversely. A study conducted by Smith (1979) observed a flight crew over longer periods of time whilst basic aircraft parameters and crew heart rates were recorded. The study showed that the number of errors increased with higher workload. A study by Hancock and Williams (1993) specifically investigated the effect of task load on monitoring and concluded that data suggest that that an increase in the task load produces an increase in both the time to react correctly to a monitoring cue.

But conversely, work underload can lead to overreliance, complacency, and boredom and lead to lowered vigilance as well. Malleable Attentional Resources Theory tells us that attentional capacity can change in size depending on the task demands, therefore mental underload can be explained by a lack of attentional resources (Masalonis et al., 1999).

The autopilot can cause such effects. The use of high-level automation leads to significantly lower objective and subjective workload for pilots. This could have an advantageous effect on their performance, but also lead to mental underload and overreliance. How to exploit the workload reducing advantages of automation whilst keeping the pilot in-the-loop in an ongoing challenge for aviation psychology (Adams et al., 1995).

One suggestion to have the best of both world came from Parasuraman et al. (1997). This study examined the use of adaptive task allocation (i.e. switching between manual and automatic control) as a possible way to enhance monitoring of automated systems and lower vigilance. The study built upon earlier studies that showed that task interruption of any form, including rest, has a beneficial effect on detection performance (Davies and Parasuraman, 1982). The former showed that adaptive task allocation helps to detect failures more quickly, both when the initial switch was made to manual control as well as after the second switch was made back to automated control.

For the experiment of this study these are useful observations. To be able to investigate the relation between the availability of cues and detection time, participants should have comparable levels of vigilance throughout the experiment and avoiding mental underload and overload should be one of the objectives of the experiment design. Loss of vigilance because of fatigue can be avoided if the experiment is not too lengthy, and with timely breaks. The fact that participants will fly scenarios manually and on autopilot should also aid this, as shown by Davies and Parasuraman (1982). The task demand can be increased with ATC commands such as level changes, heading changes, frequency changes and squawk code instructions. But, care should be taken as to not overload the participants. Since the study aims to investigate the effect of diminished cues and not the effect of workload.

In the experiment, pilots will fly manually and with the autopilot. We have seen that the use of the automation helps to lower the workload and that this reduction improves failure detection Masalonis et al. (1999), Ephrath and Young (1981). Therefore, because automation use is the variable of this study, it could be that a change in detection time could be effect of lower workload. Although this effect may be small, since workload levels in cruise condition should be comparable for manual and auto-flight, the effect of workload alleviation cannot be ruled out, unless workload is completely controlled.

5.2. Pilot monitoring Frameworks

An important part of the experiment will be to understand how pilots monitor the aircraft state to retain situation awareness. In other words, which cues do the pilots observe and process to get a mental model of the aircraft state. The following subsections will present a number of psychological frameworks that aim to describe this process. These frameworks will then be linked to explain the masking phenomenon.
The definition of monitoring is to ‘observe and check the progress or quality of (something) over a period of time’. Pilots can do this by observing the aircraft movement, gauges and instruments, and sounds. This information then has to be compared or verified with some sort of baseline. A mental model of the situation then has to be formed. Different studies give insight into this process.

5.2.1. Sensemaking Cycle

The Sensemaking Cycle (Billman et al., 2020), shown in Figure 5.2, gives a good insight into this process. Each of the three triangles represents an activity that updates the Situation Model. It consist of identifying monitoring questions, gather relevant evidence, and identify needed actions. These activities are continuously performed to update the situation model. In the cockpit this might happen in the following way: An aircraft is performing a flight level change. The pilot might therefore form a monitoring question. Perhaps ‘What is my vertical speed at this moment?’. A logical step after this would be to gather evidence to answer this question by looking at the Vertical Speed Indicator (VSI). The information that the VSI display might then be cause for identifying an action, such as changing the power setting.

The pilot will constantly go through these cycles to update his/her Situation Model. So what does this mean for the masking phenomenon? The autopilot can diminish or remove cues to the pilot. This means that gathering the right evidence might not be possible or that the updated situation model is update with the wrong information.
When one engine fails the aircraft will have the tendency to roll to one side. With the autopilot engaged, it will counter-steer against this movement and remove the attitude cue. With limited evidence available, it is hypothesized that the pilot might then lose SA.

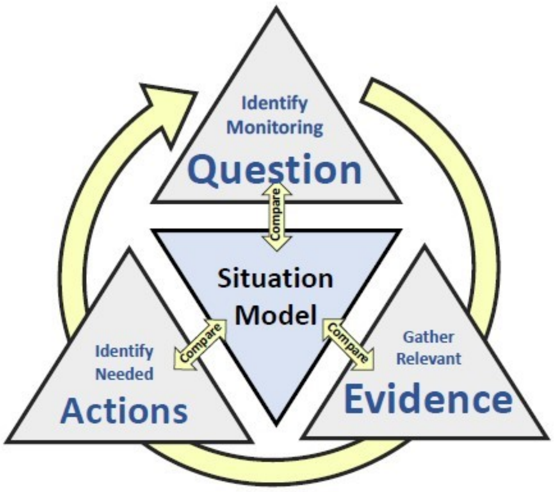


Figure 5.2: Situation model and sensemaking cycle.

5.2.2. Perceptual Cycle Model

A model that shows similarities with the Sensemaking Cycle is the Perceptual Cycle Model (PCM). The PCM, originally presented by Neisser (1976) and later updated and validated by Plant and Stanton (2014), is a model that describes the interaction of the human with the world. It argues that human thought and a person’s interaction with the world inform each other in a reciprocal, cyclical way. By considering the operator in its context, the interaction with the environment can be better understood. The PCM can be seen in Figure 5.3.

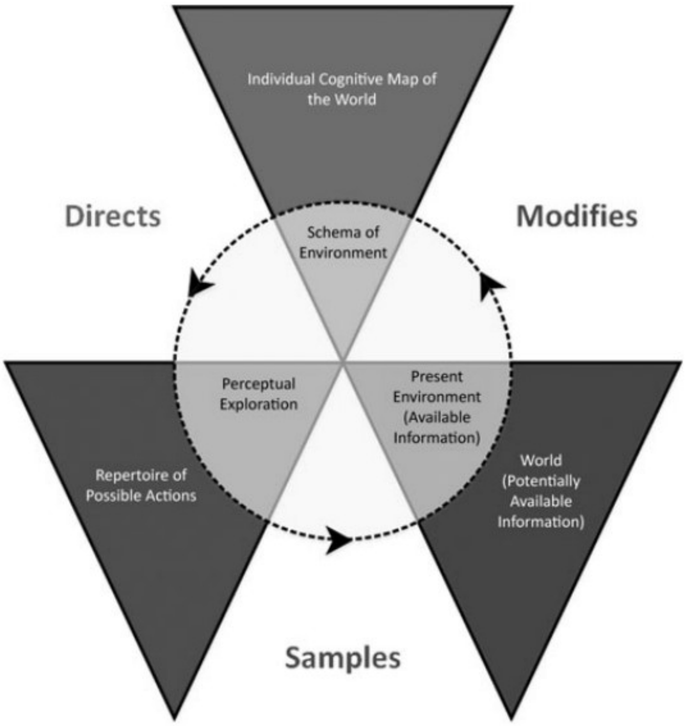


Figure 5.3: The Perceptual Cycle Model

Schemata are a concept that be thought of as mental 'templates' that are stored in the long-term memory. These templates are formed on the basis of earlier experiences and are used to interpret information, predict upcoming events, and focus attention. Schemata, Actions and the World form a cycle that is interrelated though top-down and bottom-up processing. Top-down processing occurs when a schema is activated by an observation or other stimuli, this schema can then be used to anticipate further events. Bottom up processing usually follows, hereby directed actions seek information from the environment and are interpreted in an existing schema. When observations do not match the active schema, modifications to this schema or the selection of alternative can be made. This model can and has been used to describe the mental model and decision process of pilots during failure detection (Plant and Stanton, 2013). In this study, the authors interviewed a helicopter pilot that experienced an incident during a search-and-rescue mission. The authors successfully applied the PCM to the data and concluded that the method was reliable.

5.2.3. Frame Theory

Another model that can be used to explain the loss of SA is frame theory. Frame theory has recently been used to explain behavior during surprise events (Landman et al., 2017). A frame is a mental structure based on previous experiences that can link individual data points together. Frames can be compared to schemata in the PCM. Such a frame can be useful in complex situation to describe what is happening; An engine failure in the world might reactivate a stored frame that was created in flight training. It is believed that frames are important to achieve high levels of SA. Frames are fed by incoming information (bottom-up). An anomaly, or a cue, can trigger sensemaking. Sensemaking activities that have been defined thus far are elaborating a frame, questioning a frame, preserving a frame, comparing frames, re-framing, and constructing or finding a frame.

Again, if we are looking to explain the masking phenomenon, this is where the automation would hinder the failure and hazard detection process. Faulty or limited data might activate the wrong frame or might delay the activation of the correct frame. As pilots learn from the reactions of the aircraft, removing those reactions might have negative impact (Wickens, 2003).

5.2.4. Threats and Error Management Model

The Threat and Error Management (TEM) model is a conceptual framework that assists in understanding the relationship between safety and human performance in operational contexts. The TEM model is originally developed for flight deck operations, as a product of collective industry experience. The framework and its usefulness are thoroughly explained by Maurino (2005). The TEM model can be used a safety analysis tool for single events, as is the case with accident analysis. It can also be used to analyse systemic patterns within a large set of events. For example, to understand ’just culture’ in an airline. Next to this, the model can be used as a licensing tool and a training tool in human performance.

The model consists of three components:

- Threats
- Errors
- Undesired Aircraft States (UAS)

Threats are “Events or errors that occur beyond the influence of the flight crew, increase operational complexity, and which must be managed to maintain the margins of safety.” Examples of threats are adverse meteorological conditions or congested airspace. Jump seat observations have noted threats occurring in 79% of all analysed flights (Helmreich, 2000). Some of these threats come unexpectedly, such as in-flight malfunctions. In which case flight crews must use their skill and knowledge to manage this threat. Other threats can be anticipated with briefing the flight crew response in advance. This could happen if the adverse weather is briefed prior to flight, the threat is then managed preemptively. Lastly, some threats are not directly obvious to the flight crew. These threats are called latent threats and may need to be uncovered by safety analysis. Examples of latent threats are equipment design issues, optical illusions, or shortened turn-around schedules.
For the experiment in this study, the function of the autopilot may be seen as a latent threat. The autopilot will function as intended, but this may still have adverse effects when it is coupled with a failure. An engine failure will also be a threat as well as the meteorological conditions in which icing can occur.

Errors are actions or in-actions from the crew that lead to deviations from the intention or expectation of the crew. These errors can often lead to undesired aircraft states. An error can come with or without a direct link from a threat or as part of an error chain. Examples of errors can be a unstable approach, flying in an incorrect automation mode or failing to do a checklist. Errors have been observed in 64% of analysed flights (Helmreich, 2000). The effect of the error on the safety of the flight depends on whether the flight crew detects and responds to the error before it leads to an undesired aircraft state. TEM aims to understand the error management and not solely the error causality.
For this experiment, an error could be the use of automation in a situation for which it is not intended. The use of automation is advised against during icing conditions or during a failure, but such an error might be easily made. Another error might be the failure to follow SOPs and cross-verify the state of the automation. This can be done by checking the steering columns input as well as the flight instruments.

Failing to manage the errors and threats could lead to UAS. Examples of this included lining up for the incorrect runway, penetrating restricted airspace or an unstable approach. A study by Helmreich (2000) observed undesired aircraft states in 32% of analysed flights.
In scenarios in which the automation masks an engine failure or the onset of ice this undesired state could be a resulting pitch or roll angle. Next to this, incorrect flight control configuration and incorrect automation configuration can also be seen as undesired states.

An important point that be made using these definitions is that flight crews should make a timely switch from error management to undesired aircraft state management. For example, when the autopilot has successfully masked the onset of icing and the situation has progressed to an undesired aircraft state (e.g. loss of lift, loss of altitude, operation outside of aircraft limits), the crew should initially focus on bringing the aircraft to a stable and predictable state. The management of the initial error (e.g. diagnosis of the problem and resolution) is of lower priority. Focussing on the initial error might lead to the crew getting ‘locked-in’ to error management and allowing the UAS and develop. For this experiment it will be interesting to see the participants threat, error and UAS management strategies. When and to what degree they are deployed and which external factors lead them to identify and manage these errors, threats and undesired aircraft states. This can be done by applying the TEM framework to the data.

5.2.5. Divergence

Recent research has coined the term 'divergence' to describe the incorrect mental model of the crew (Silva and Hansman, 2015). Figure 5.4 shows the timeline over which an incident might happen. The beginning of divergence is when the crew state assumption does not match with the actual state of the aircraft. With the frameworks that have been explained beforehand, this could also be dubbed an incorrect frame, or schema.

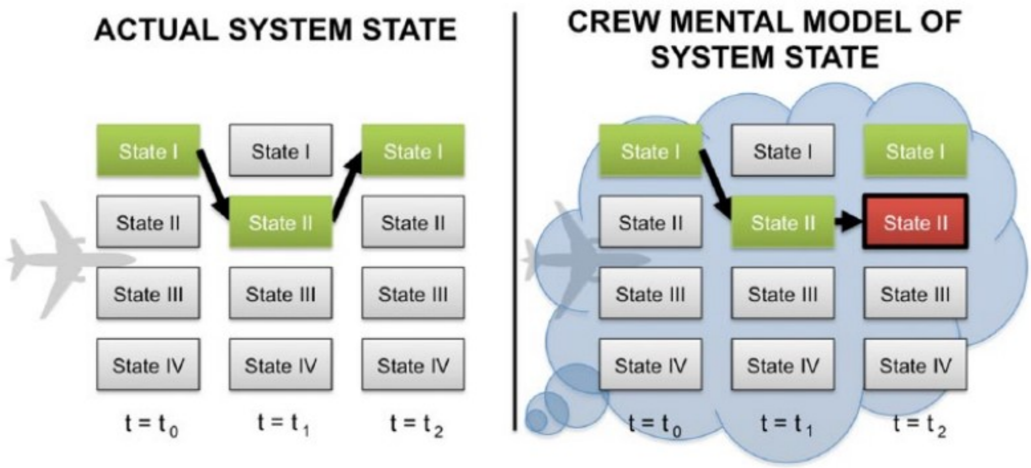


Figure 5.4: Timing of divergence, convergence, recovery, and impact for auto-throttle mode confusion cases. (Silva and Hansman, 2015)

This process is illustrated by Figure 5.5. In this figure we see a possible example of divergence. At time t = t0, the system state and the mental model match. Between time t = t0 and t = t1, the system state and the model transition to State 2. Divergence occurs are t = t2, when the system transitions to State 2 but the mental model does not follow.

The scenarios that have been developed for the experiment and that will be discussed in Chapter 6 are all divergence of the type D-1a. For this type, the actual state transitions without input from the crew, and the crew's mental model is not updated to reflect the transition. To test when and how divergence will happen, as well as to understand where causes and cognitive processes that lead to divergence, an analysis of automation-related accidents was conducted by the authors. In this analysis, it was found that divergence occurred in all analysed cases.

One could argue that divergence in failure detection is not as binary as suggested in the study by Silva and Hansman (2015). The flight crew might initially be indifferent to a change of state, for example an engine failure, and realise this failure only after some time. Although the failure is now identified and can be dealt with with the according procedures, the root cause of the problem (e.g. software glitch, wear on fuel lines, hydraulic issues, bird strike) might remain undisclosed to the flight crew. The crew will therefore not be able to achieve true convergence of the mental model and the actual state of the aircraft. Therefore, an unexpected failure is a good reason to divert or to return-to-base and assess the true state of the aircraft on the ground.

For this study, only the actions up to the identification of the divergence are considered. This means that only the time it will take for the crew to recognise that there is some problem is of interest. This because the primary objective of the study is to show a difference in detection time between manual and automatic flight. The phases that follow identification, namely diagnosis and recovery, are more complex and rely heavily on training and existing procedures. This would it considerably harder to measure detection time. A follow-up study could investigate diagnosis and recovery strategies.

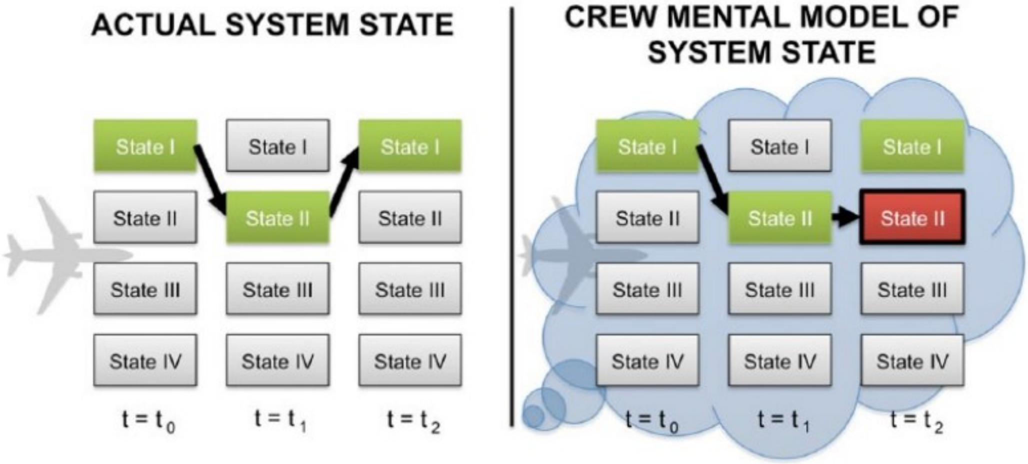


Figure 5.5: Schematic representation of divergence.

5.3. Pilot Monitoring Practices

As stated before, the monitoring of the flight path and the aircraft state is an important factor for the experiment. During the experiment runs, we would like the participants to behave as they would during a routine flight. Therefore, it is important to understand the Standard Operating Procedures (SOPs) that flight crew and airlines employ for monitoring, especially during the cruise phase. It is of interest which cues the pilots have available to make their mental image of the situation, and how they are actively sourced by monitoring. Which cues give the pilot on image of the state of the automation? And how the avionics alerting functions can help the pilots assess the state of the aircraft.

5.3.1. Monitoring in Commercial Aviation

In Commercial Aviation (CA), a flight crew almost always consist of two pilots. During such flights there is a clear distribution of roles over the two pilots. One pilot will act as the Pilot Flying (PF), responsible for the managing the current and projected flight path and energy of the aircraft at all times. The other is the Pilot Monitoring (PM), responsible for monitoring this flight path as well as communication with ATC (Flight Standards Service, 2015). Both roles can be assigned to either the Captain or the First Officer (FO). This means that both crewmembers will be monitoring the PFD and Engine Indicating and Alerting System (EICAS) in Boeing aircraft or the Electronic Centralised Aircraft Monitor (ECAM) in Airbus aircraft.
A quick field survey among KLM pilots was conducted to obtain more information on current monitoring practices. Three pilots with different type ratings (Boeing 737, Airbus A330, Embraer 175/195) reported that monitoring tasks during take-off and landing are well-defined. For example during take-off, the PM will focus solely on the engine indicators to spot an engine failure immediately. But that during cruise, there is no defined procedure for monitoring the flight path.

All pilots reported that their monitoring techniques are still very similar as they were trained to do in flight school. This is very useful for the experiment, as the experiment will be conducted with a twin-engine Piper Seneca, that is similar to aircraft used in flight schools.

5.3.2. Monitoring in General Aviation

In GA (and in pilot training), pilots are taught several monitoring techniques. The starting point of the monitoring process is to fly Pitch-Power. This means to match the Angle-of-Attack with the thrust setting of the engine and to verify this combination based on previous experience and with the primary flight instruments. If the pilot wants to climb to a different altitude, finding a pitch and power setting will have priority. This setting can then be adjusted for example by checking the VSI or Airspeed Indicator (ASI). During all stages of a VFR or IFR flight, the most important focus for the pilot will be the basic 6. This ‘six pack’ can be seen in Figure 5 and contains the following instruments:

- Airspeed Indicator (ASI)

- Altimeter
- Attitude Indicator (AI)
- Heading Indicator (HI)
- Vertical Speed Indicator (VSI)
- Turn Coordinator (TC)

Figure 5.6: Basic 6 of Flight Instruments

The instruments in Figure 5.6 can be scanned using different techniques. Namely, the T-scan, Radial or Hub-Spoke scan, and Inverted V-scan. There methods all revolve around the Attitude Indicator as the most important instrument to retain situation awareness. It is interesting to note that no procedures seem to be defined on how to monitor the state of the automation when the autopilot is engaged.

5.3.3. Monitoring of the Autopilot

Part of monitoring the systems and instruments on the flight deck involves monitoring the state of the autopilot. Understanding autopilot monitoring practices are essential for a successful conduction of this experiment as correct autopilot monitoring will serve as a vital cue in the failure scenarios. The potential masking effect will take away longitudinal and lateral attitude cues, but the input of the autopilot will be a replacement cue for the participants flying with the autopilot. How will pilots be able to observe the state and the inputs of the autopilot?

The KLM pilots who were asked on their monitoring practices named the Flight Mode Announciator (FMA) as the most important tool to monitor the state of the automation. The FMA is located on the top of the PFD and can be seen in Figure 5.7.

The FMA shows the current autothrottle mode, roll mode and pitch mode of the autopilot. The FMA is therefore useful against mode confusion. For the experiment, the FMA will be of lesser importance as the autopilot itself will not fail and mode confusion will not be investigated, as it is outside the scope of this study. The FMA will therefore not provide a cue of a failure.

The FD can also help pilots to see the intention of the autopilot. The FD is a magenta overlay on the PFD showing the required attitude for the intended flight path. This flight path comes directly from the inputs in the Flight Management System (FMS) or the Mode Control Panel (MCP) or Flight Control Unit (FCU). This can be seen in figure Figure 5.8. The FD can turned on or off at all times by the flight crew and can be used in two ways: During automated control it can serve as a cue to the intention of the autopilot; showing the attitude that the autopilot is steering towards. During manual control it will also show the required attitude for the intended path but it will be up to the pilot to follow the FD. In a way this reduces the task of the pilot to a tracking task and thus it is a hybrid form between manual and automatic control.

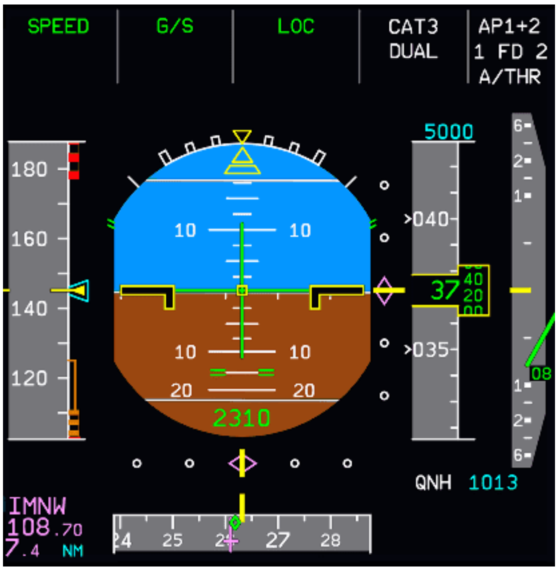


Figure 5.7: FMA presented on the top of the PFD

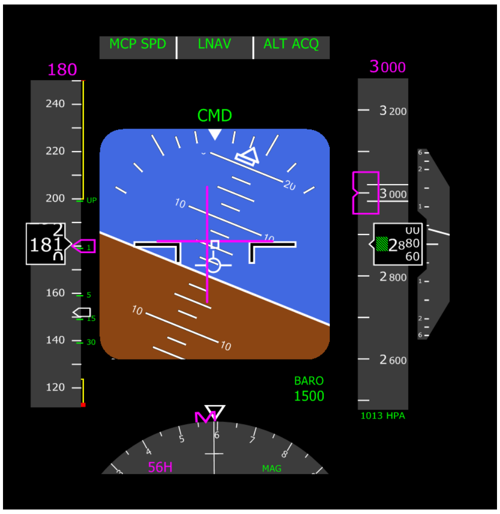


Figure 5.8: The Flight Director (FD) as a magenta overlay on the PFD

For this study, the level of automation is not of concern, only if the autopilot is used or not. Therefore, the FD will not be used in the manual flying scenarios.

More relevant for the experiment is the autopilot backdrivability. When the steering column is directly connected to the flight controls, the movement of the yoke will provide a visual cue of the commands and the magnitude of the control authority of the autopilot. When the aircraft is flown manually, the active stick can provide tactile cues that help pilots to retain SA (Hegg et al., 1994). Additionally, when the steering columns are slaved to each other, the PM will be able to 'see' the inputs of the PF.

The visual cue that the backdrivability provides allows for feedback on the autopilots input. This is an important cue for pilots to maintain SA or to diagnose problems with the aircraft; When the autopilot remains engaged during an engine failure, the autopilot will steer against the tendency of the aircraft to roll. The pilot will be able to see this roll input and will realise that there is some sort of control issue. Not all aircraft have this backdrivability of the autopilot. With the rise of fly-by-wire came sticks that are not directly connected with the controls and do not provide this feedback. These are so-called passive sticks and have been implemented in Airbus family aircraft and the F-16 Fighting Falcon among others. The differences between passive and active sticks can be seen schematically in Figure 5.9. An experimental comparison between the two, comparing the level of SA, would be interesting for further research, but is

outside the current scope of this study.
This study will use the backdrivability of the steering columns, this because this still is most common configuration. Especially in GA, passive sticks are the exception.

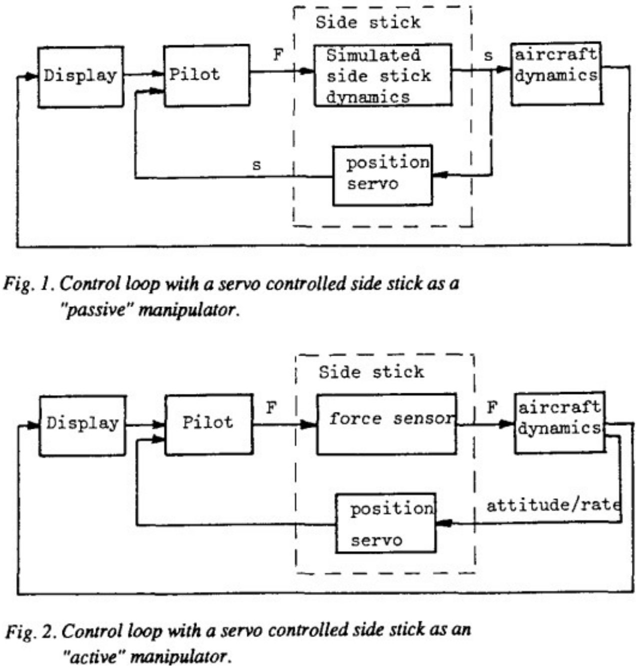


Figure 5.9: Schematic view of control loops for passive/active sticks (Hosman et al., 1990)

5.3.4. Alerting by automated systems

Pilots do not have to rely solely on their own ability to monitor the state of the aircraft, because with the rise of automation on the flight deck came intelligent systems that can help the pilot to assess a failure by alerting. Alerts can come in the form of annunciations, horns, or amber/red markings on the instruments.

The certification standards that have been discussed in Chapter 4, have expressed the usefulness of alerts in masking conditions. According to AC 25.1329-1C, situations to consider for such alerts are the following:

> (a) *Sustained lateral control command.* If the autopilot is holding a sustained lateral control command, it could be indicative of an unusual operating condition for which the autopilot is compensating. Examples of such unusual operating conditions are asymmetric lift and/or drag due to asymmetric icing, fuel imbalance, or asymmetric thrust. In the worst case, the autopilot may be operating at or near its full authority in one direction. If the autopilot were to disengage while holding this lateral trim, the result would be that the airplane could undergo a rolling moment that could possibly take the pilot by surprise. Therefore, a timely alert should be considered to permit the crew to manually disengage the autopilot and take control prior to any automatic disengagement that might result from the condition.
> (b) *Sustained pitch command.* If the autopilot is holding sustained pitch command, it could be indicative of an unusual operating condition (for example, inoperative automatic horizontal trim) for which the autopilot is compensating. If the autopilot were to disengage while holding this pitch command, the result would be that the airplane could undergo an abrupt change in pitch that could possibly take the pilot by surprise. Therefore, a timely alert should be considered to permit the crew to manually disengage the autopilot and take control prior to any automatic disengagement that might result from the condition.

In other words, when the autopilot is asserting a sustained command in either the longitudinal or lateral axis, potential masking conditions may be in effect, and timely alerts should be considered to make pilots aware of this.

The aforementioned AC does not describe what ‘timely’ is. Turbulence or changing winds might lead to sustained pitch/lateral control commands in a natural way. If false alerts are to be avoided, the time it will take before an alerts activates might be substantial.

The Garmin G1000 is an Electronic Flight Instrument System (EFIS) that has the ability to do alerting. The function and form of these alerts will now be discussed as a mock-up G1000 will be used in the experiment.

Alert Condition	Annunciation	Description
Aileron Mistrim Right	AIL→	Roll servo providing sustained force in the indicated direction
Aileron Mistrim Left	←AIL	
Elevator Mistrim Down	↓ELE	Pitch servo providing sustained force in the indicated direction
Elevator Mistrim Up	↑ELE	
Pitch Trim Failure)	PTRM	If AP engaged, take control of the aircraft and disengage AP
Roll Failure	ROLL	Roll axis control failure; AP inoperative
Pitch Failure	PTCH	Pitch axis control failure; AP inoperative
System Failure	AFCS	AP and MET are unavailable; FD may still be available
Preflight Test	PFT	Performing preflight system test; aural alert sounds at completion Do not press the **AP DISC** Switch during servo power-up and preflight system tests as this may cause the preflight system test to fail or never to start (if servos fail their power-up tests). Power must be cycled to the servos to remedy the situation.
	PFT	Preflight system test failed; aural alert sounds at failure

Figure 5.10: Table of G1000 Alerts

Figure 5.10 shows the possible annunciations of the G1000 (Garmin, 2011). Only one annunciation occurs at a time and messages are prioritized on criticality. The alerts for Aileron Mistrim and Elevator Mistrim are of interest in this study. According to Garmin, these alerts fall in the ‘Caution’ category. Hence, the alert is in yellow. Alerts in the ‘Warning’ category are marked red.
The alerts will appear in the Automatic Flight Control System (AFCS) System Status Field on the PFD. This is in the top left corner of the PFD as can be seen in Figure 5.11.

Although it is known how the G1000 will warn for sustained control in the lateral and longitudinal axis, it remains unknown when, and based on what condition the instrument will warn. Therefore, it remains unclear how the word ’timely’ is interpreted for the G1000.

This, and the fact that the G1000 has one of the most advanced AFCS systems but that most general aviation aircraft do not fly with such a system, makes that this study will not consider alerts in the experiment. The certification standards and implementation leave enough space to investigate the potential masking effect without alerts. Even more so, they indicate that the understanding of the masking effect is limited, as no examples of what pilots can expect nor strategies to deal with this issue are explained.

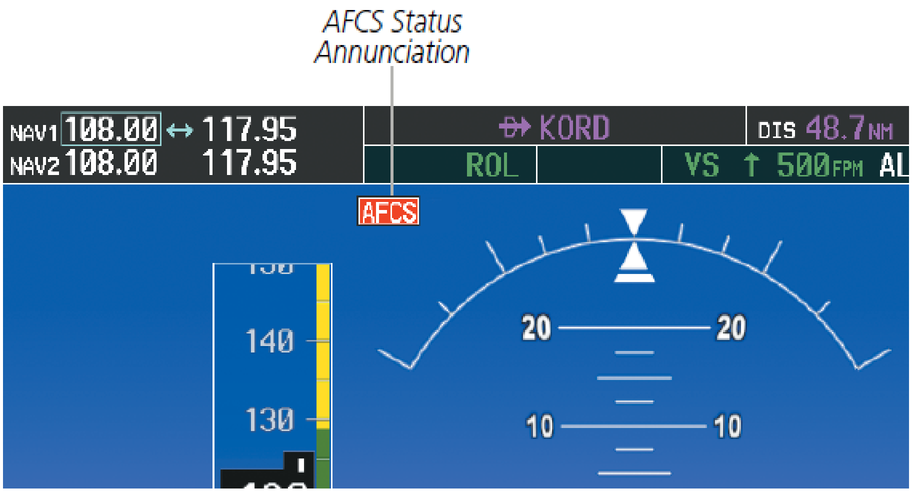


Figure 5.11: Location of Alert on PFD

6 Experiment Design

This section will describe the design of the experiment, and the underlying choices that have led to these choices. It will described the requirements for failure scenarios and the selection thereof, the hypotheses, the participants, the hardware and software that will be used as well as the flight model, procedures and how a possible scenario might be tested, the experiment matrix and finally the dependent measures.

6.1. Requirements for Failure Scenarios

The selection of scenarios is the most critical part of this research, as the scenarios should be apt to demonstrate the masking mechanisms and quantify its effect. Since it is so important, and most time of the project was spend on the selection of these scenarios, it would be suiting to explain the choices that have been made in this process. To investigate the automation masking effect effectively, a number of requirements for the failure scenarios have been defined:

- Each scenario shall be an example of the automation masking effect.
- Each scenario shall be realistic (ideally an example of a real-world incident is found).
- Each scenario shall be subtle (i.e. with an expected detection time that is significantly longer than the time for a human reflex).
- Each scenario shall have comparable outcomes for manual- and auto flight
- Each scenario shall have a detectable failure or hazard.
- Each scenario shall be unique.

Although some of these requirements might seem trivial, the implications of these requirements make that only a limited number of scenarios are available. For example, all failures where the automation itself will fail would not be suitable for the experiment since there is no comparable situation in manual flight. A scenario is which a secondary subsystem will fail, but that the crew does not detect because they are busy with the autopilot, would be a scenario better suited to investigate over-reliance or loss-of-vigilance. A failure that is induced very quickly would be more suited for startle and surprise research. Finally, the requirement for a realistic scenario ensures that the participants will not experience a trap, carefully set like a mechanical clockwork, but with no application to the real world.

6.2. Selection of Failure Scenarios

The Automation Masking effect can occur when the autopilot obscures cues in failure and hazard detection. For an aircraft with a basic autopilot without autothrust and without rudder servos, the autopilot can take away these cues only in the pitch- and roll axis. With this knowledge as well as the other requirements stated in Section 6.1, a selection of the failure scenarios could be made. It was chosen that a one-engine out scenario would be an apt choice for the roll axis failure and that icing accumulation, resulting in loss of lift in the longitudinal direction would be fit for the failure in the pitch axis. In both scenarios the autopilot would actively steer against the natural movement of the aircraft. By doing so, the autopilot is hiding in plain sight that the aircraft is moving towards the end of the control envelop. Both scenarios are also realistic and can be relatively slowly introduced.
Other failures in the pitch and roll axis were also considered. Two-engines out, a pitch servo failure, a roll

servo failure, a rudder servo failure, or a mass shift were other options for the scenarios. But it must be noted that all of these failures have a similar effect in terms of detection cues. For all these examples, the autopilot will counter-steer in a particular axis to hide the failure. Since detection, and not diagnosis of the problem is the scope of this study, the origin of the failure is of secondary importance.

6.3. Participants

In the experiment, about 30 pilots will be asked to perform flight scenarios in the SIMONA Research Simulator at Delft University of Technology. These pilots shall be in the possession of an ATPL and, possibly, the same type-rating. This group is selected so that it can be ensured that the group is homogeneous and that all pilots have a similar proficiency in using the autopilot. Also, this group will have experience from their initial training with flying a twin-propeller aircraft, which will be used in the experiment. Pilots will be asked about their flight hours, recent experience and recent twin-propeller experience to ensure that there is no significant difference in proficiency in using the autopilot, as this might influence the results in an unwanted manner.

6.4. Facilities and Equipment

In order to perform this experiment these scenarios have to be programmed in C++. Delft University Environment for Communication and Activation (DUECA)) is the framework for the implementation of the real-time program that is used. The scenarios are heavily based on the work of Landman et al. (2017), but have been modified to match the goals of this study. The experiment will be performed using the SIMONA Research Simulator (SRS).

The aerodynamic model that will be used in the experiment is the Piper Seneca II, a two-engine propeller aircraft equipped with a GFC700 autopilot and displays based on the Garmin G1000 avionics. In this research, a plethora of possible failures and faults have been programmed that can serve as a basis for this experiment. The Piper Seneca II setup is general and representative, therefore the conclusions of this experiment will be applicable to either General Aviation and Commercial Aviation.
The Piper Seneca II is fitted with a basic autopilot that can control elevator and aileron servo inputs, as well as several autopilot modes such as Flight Level Change, Vertical Navigation, Horizontal Navigation etc. The aircraft is not fitted with auto throttles or the ability for the autopilot operate the rudder.
Another important reason to use the Seneca is the continuity of the research. At the faculty of Aerospace Engineering at Delft, University of Technology former studies have used the Piper Seneca II model in experiments on startle and surprise (Landman et al., 2017) and automation surprises (van Leeuwen, 2020), among others. The aircraft is therefore well known and free of any teething problems. Furthermore, the results and recommendations can then more easily be a starting point for further studies.

Figure 6.1: The Piper Seneca II aircraft

6.5. Procedures and Design

All pilots will receive a familiarisation training, so that they can become accustomed to the procedures in the Seneca. After this familiarisation, they will perform a number of scenarios in which certain faults might occur. They are informed beforehand that the experiment is about testing the fidelity of simulated events, but that scenarios can also be uneventful. The pilot will also be asked to talk through his thinking process, especially when something seems off. This will help to get an insight which cues help the pilot to diagnose the problem, and their mental models of the sit.

They will then perform eight experiment runs, four of which are test scenarios featuring severe icing accumulation or an engine failure. All participants will perform those scenarios both manually and on autopilot. The other fours runs are used as a distraction, so that the participants do not recognise which failure will happen and the experiment goal is disguised.

Participants will have different cues available to them to recognize the failure or hazard, depending on whether they are flying manually or on the autopilot. When the autopilot is used, the masking effect will take away cues in the longitudinal or lateral attitude of the airplane.
A possible flight for the manual condition might look as follows: The scenario is started at cruise altitude. The participant will fly straight and level to give the participant time to get situation awareness. From there, there will be a number of instructions that can be expected on a normal flight. For example a Flight Level change, heading change, frequency change or squawk code. After a predefined time, the aircraft will slowly lose power on the left engine. The pilot will have a number of cues available to diagnose this problem: 1. Aircraft will yaw and roll to the left 2. Sound of the engine will change 3. Pilot can see the engine windmilling after the power is gone 4. Engine indicators will show decreased power.
The pilot will have to gather this evidence and update his mental model. When the pilot announces that he has noticed something is wrong, the time will be noted and the scenario will be ended.

For the participant flying the aircraft on autopilot this scenario is very similar. However, because of the masking effect of the autopilot, the aircraft will not yaw and roll to the left as much as in the manual scenario, as the autopilot will correct the roll with automatic aileron input meaning only a slight yaw angle will remain for the pilot as a cue of attitude change. This can be seen in Figure 6.2, the aircraft is wings-level but has slightly drifted to the right and has a yaw angle. The pilot will however have an additional cue available to be able to see this correction from the autopilot on the steering column. As the autopilot moves the ailerons, the steering column will move accordingly. The other cues of the engine failure are the same as in the manual scenario.

Figure 6.2: PFD of G1000 50 seconds into the Engine Out scenario. No roll angle visible, FPV indicates remaining sideslip

6.6. Dependent Measures

In each scenario it is measured if, and to what degree, the participant succeeds in the recognition of a fault. In order demonstrate the effect of the use of autopilot on the failure and hazard detection process, the following variables are measured:

- *Detection Time.* The time between the start of the fault and the moment the pilot mentions that something is wrong. This is the primary outcome measure for this study. This parameter will be an important marker to understand if it is more difficult to detect a fault when flying on the autopilot
- *Performance score.* After each scenario, the participants are asked to write down if they noticed anything strange, and if so: what it was, what cues made them notice this, and what they think the underlying cause was. The extent to which the participant has noticed the fault divided into four levels, meaning that this will be an ordinal measure:
    1. The participant has failed to spot the fault.
    2. The participant has noticed certain cues, but did not notify the experiment leader.
    3. The participant did notice and announced the fault, but cannot recall the cues that helped him/her to this.
    4. The participant has correctly identified the fault and is able to explain which cues brought him/her to this conclusion.
- *Participants comments and remarks.* The comments and remarks of the pilot will be recorded to get an insight in the thought process of the participant. This can be an indication of which cues the pilot predominantly uses to understand that a fault is happening. And will also help to apply psychological frameworks such as TEM and PCM to the data.
- *Control inputs.* Additionally, when a scenario is flown manually, the pilot's steering inputs will also be measured. It could be the case that the pilot might correct the disturbance of the fault unknowingly. In this case, the pilot would be masking the fault themselves, possibly degrading their own ability to diagnose a problem. The control inputs will also give context to the pilot's comments and validate that the aircraft was in (partial) steady-state when the failure occurred

6.7. Experiment Matrix

The Experiment Matrix shows a possible order in which participants will experience the scenarios. Excluding the familiarisation flights, the order will be mixed quasi-randomly so that learning effects are mitigated. The experiment matrix is designed in such a way that participants will fly both in manual- and automatic mode, so that a potential skill difference between pilots would not affect the data.
The first four runs will be used as familiarisation, so that the participants can become accustomed to both manual and automatic flight. They will also see a failure, so that they know what is expected from them when these failures happen. Runs that will be used to answer the research question named in Chapter 3 are runs 6, 8 10 and 12. In runs 5, 7, 9 and 11 pilots will experience either nothing or a 'Filler' at the end of the run. The filler is a generic failure that the participants will have to detect that is not used to verify the hypothesis. These fillers are used so that participants will not recognize patterns in the experiment matrix and will keep participants 'on edge'. Between run 6 and 7, there will be a 30 minute break. In total, the experiment should take 140 minutes.

Table 6.1: Experiment Matrix for four familiarisation runs

Run	Mode	Scenario	Time [min]
1	Manual	Familiarisation Flight	10
2	Automatic	Familiarisation Flight	10
3	Manual	Failure Practice	5
4	Automatic	Failure Practice	5

Table 6.2: Experiment Matrix for eight experiment runs

Run	**Mode**	**Scenario**	**Time [min]**
5	Manual	Filler Failure	10
6	Manual	Icing	10
Break	-	-	30
7	Automatic	Filler Failure	10
8	Automatic	Engine Failure	10
9	Manual	No Failure	10
10	Manual	Icing	10
11	Automatic	No Failure	10
12	Automatic	Engine Failure	10

7 Conclusion

This preliminary thesis report has investigated the use of automation on the flight deck and proposes an experiment to examine the effect thereof on failure and hazard detection. It has considered the masking mechanism of the autopilot by exploring related accidents, certification standards and previous research. It was found that the autopilot, working as intended, can take away vital cues in the failure and hazard detection process. The autopilot can do so by counter-steering the natural movement of the aircraft during a failure, taking away longitudinal or lateral attitude cues that hint towards that failure. Previous research has used detection time as primary outcome measure, as a swift recognition of the problem is vital for safe operation of the flight. Previous research has provided a ballpark figure for detection times and has given an insight into current experiment design standards and pilot decision making processes. Crucially though, it hasn't investigated the effect of limited cues on the failure and hazard detection process.
This report has further considered the masking effect from a pilots perspective. It has explained the effect of task type, task duration and workload on pilot vigilance and recommended that these parameters should be, if possible, kept at a constant in the proposed experiment. Next to this, psychological frameworks that can be used to explain the mental model of pilots have been described. These include the PCM, TEM model and Sensemaking cycle. Following this, the report investigated current monitoring practices in General and Commercial Aviation by means of a field survey. It was found that monitoring practices of airline pilots are similar to the techniques used in ATPL training, often done in single- or twin-engined propellor aircraft. And that monitoring relies heavily on checking the basic flight instruments. It was found that the state of the automation can be monitored by observing the FMA and FD, but that regularly checking the autopilot inputs through the backdrivability of the control column is not part of current operations. The report went on by explaining how state-of-the-art avionics can help flight crew retain situation awareness. To demonstrate the effect of the autopilots masking mechanisms on failure and hazard detection an experiment is proposed. This experiment aims to single-out the effect of limited cues by making a comparison between manual and automatic flight. In this experiment, 20 airline pilots will be asked to fly the SIMONA Research Simulator at Delft University of Technology. After four familiarisation flights in the Piper Seneca II, they will experience eight experiment runs in which a failure might happen. These failure scenarios have been carefully selected after a number of requirements were set. A single engine failure and severe icing accumulation were chosen as scenarios to demonstrate the masking effect. The report further recommends that the other four runs should be used to mitigate possible confounds. These include recognition of failures due to repetition and too much focus of the pilots on detecting a failure. The report further selected four dependent measures for the experiment: Detection time, Performance score, Participants comments and remarks, and Control Inputs. The results of this experiment and the scenarios that have been developed to conduct the experiment could become a future example for pilot training. This should make pilots more aware of the potential masking effect of using automation and stimulate pilots to always attempt to ”see through” the automation.

A

Project Gantt Chart

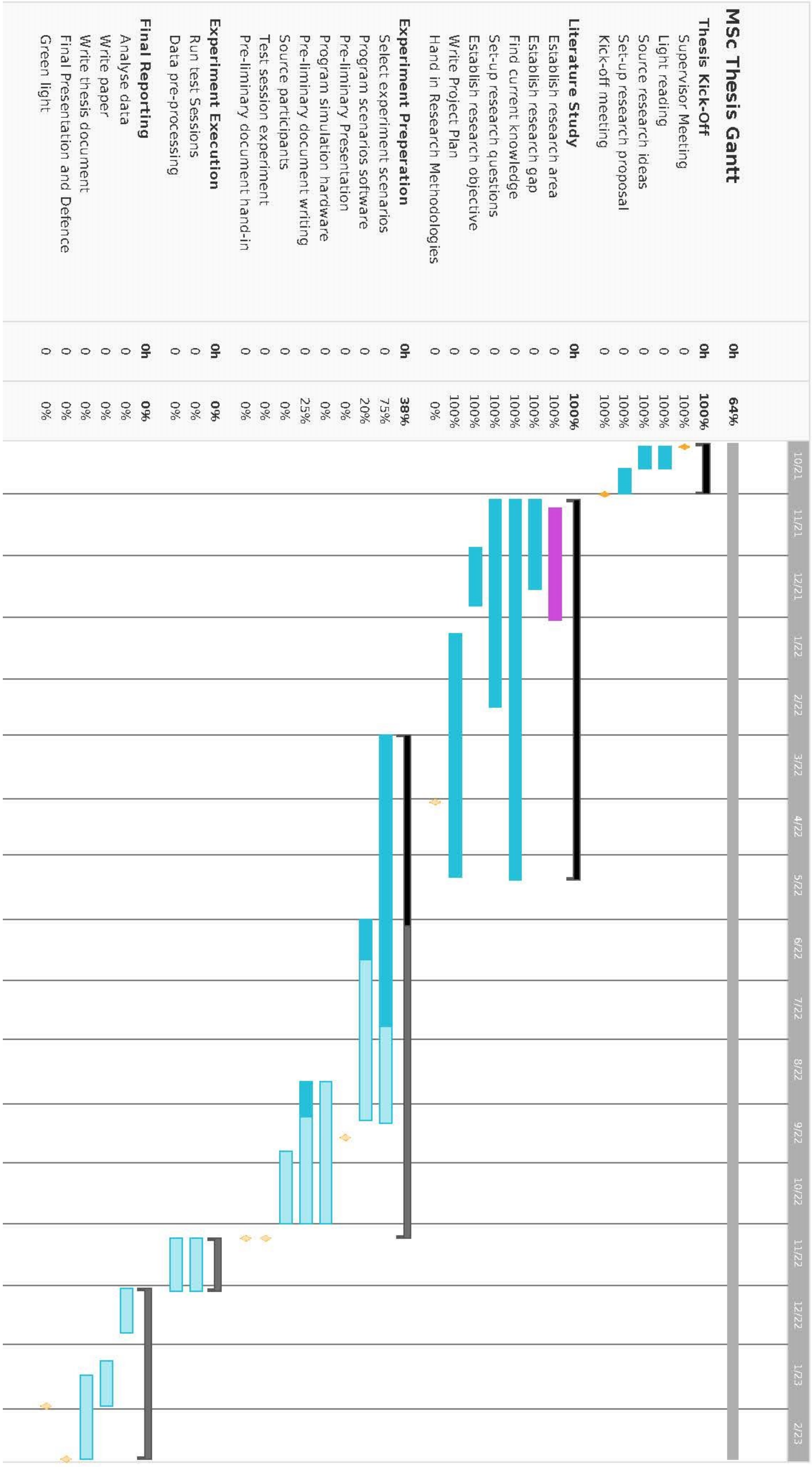


Figure A.1: Gantt Chart

Bibliography

NTSB. (1985). *Aircraft Accident Report, China Airlines Boeing 747-SP* (tech. rep.).

Jensen, R., & Rakovan, L. (1995). *Proceedings of the Eighth International Symposium on Aviation Psychology*. Ohio State University, Department of Aerospace Engineering, Applied Mechanics; Aviation.

Norman, D. (1990). The 'problem' with automation: Inappropriate feedback and interaction, not 'over-automation'. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *327*, 585–93. https://doi.org/10.1098/rstb.1990.0101

NTSB. (2014). *Preventing loss of Control in Flight in General Aviation, Most wanted list of transportation Safety improvements* (tech. rep.).

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

Boeing Commercial Airplane Group. (1997). *Statistical summary of commercial jet aircraft accidents: Worldwide operations* (tech. rep.).

Draper, C. S., Young, L. R., & Whitaker, H. P. (1964). *The roles of man and instruments in control and guidance systems for aircraft.*

IATA. (2014). *Loss of Control In-Flight Accident Analysis Report* (tech. rep.).

Snow, M. A. (2015). Preventing loss of control in flight. *Boeing Aero Magizine, No*, *57*, 9–12.

Billings, C. E. (1997). *Aviation automation: The search for a human-centered approach* (Vol. 36). Taylor Francis.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/https://doi.org/10.1016/0005-1098(83)90046-8

Active Pilot Monitoring Working Group. (2014). *A pratical guide for improving flight path monitoring* (tech. rep.). Flight Safety Foundation.

Sarter, N., & Woods, D. (1995). How in the world did we ever get into that mode? mode error and awareness in supervisory control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*, 5–19. https://doi.org/10.1518/001872095779049516

Young, M., & Stanton, N. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human factors*, *44*, 365–75. https://doi.org/10.1518/0018720024497709

Parasuraman, R., Sheridan, T., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans : a publication of the IEEE Systems, Man, and Cybernetics Society*, *30*, 286–97. https://doi.org/10.1109/3468.844354

Masalonis, A. J., Duley, J. A., & Parasuraman, R. (1999). Effects of manual and autopilot control on mental workload and vigilance during simulated general aviation flight. *Transportation Human Factors*, *1*(2), 187–200. https://doi.org/10.1207/sthf0102_7

NTSB. (1994). *Aircraft Accident Report, In-flight Icing Encounter and Loss of Control* (tech. rep.).

Abbott, K. H., Slotte, S., & Stimson, D. K. (1996). *The interfaces between flightcrews and modern flight deck systems* (tech. rep.). Federal Aviation Administration.

Cole, J., & Sand, W. (1991). Statistical study of aircraft icing accidents. *29th Aerospace Sciences Meeting*, 558.

Ephrath, A. R., & Young, L. R. (1981). Monitoring vs. man-in-the-loop detection of aircraft control failures. In J. Rasmussen & W. B. Rouse (Eds.), *Human detection and diagnosis of system failures* (pp. 143–154). Springer US. https://doi.org/10.1007/978-1-4615-9230-3_10

Ephrath, A. R., & Curry, R. E. (1977). Detection by pilots of system failures during instrument landings. *IEEE Transactions on Systems, Man, and Cybernetics*, *7*(12), 841–848. https://doi.org/10.1109/TSMC.1977.4309640

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*, *9*(1), 24–34. https://doi.org/10.1109/TSMC.1979.4310070

Beringer, D. B., & Howard C. Harris, J. (1999). Automation in general aviation: Two studies of pilot responses to autopilot malfunctions. *The International Journal of Aviation Psychology*, *9*(2), 155–174. https://doi.org/10.1207/s15327108ijap0902_5

Al-Shargie, F., Tariq, U., Mir, H., Alawar, H., Babiloni, F., & Al-Nashash, H. (2019). Vigilance decrement and enhancement techniques: A review. *Brain Sciences*, *9*(8).

Davies, D., & Parasuraman, R. (1982). *The psychology of vigilance*. Academic Press London ; New York.

Caggiano, D., & Parasuraman, R. (2004). The role of memory representation in the vigilance decrement. *Psychonomic bulletin review*, *11*, 932–7. https://doi.org/10.3758/BF03196724

Wright, N., & McGown, A. (2001). Vigilance on the civil flight deck: Incidence of sleepiness and sleep during long-haul flights and associated changes in physiological parameters [PMID: 11214900]. *Ergonomics*, *44*(1), 82–106. https://doi.org/10.1080/00140130150203893

Wiggins, M. (2011). Vigilance decrement during a simulated general aviation flight. *Applied Cognitive Psychology*, *25*, 229–235. https://doi.org/10.1002/acp.1668

See, J., Howe, S., Warm, J., & Dember, W. (1995). Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin*, *117*, 230–249. https://doi.org/10.1037/0033-2909.117.2.230

Smith, H. P. R. (1979). A simulator study of the interaction of pilot workload with errors, vigilance, and decisions.

Hancock, P., & Williams, G. (1993). Effect of task load and task load increment on performance and workload. *Seventh International Symposium on Aviation Psychology*, 328–334.

Adams, M. J., Tenney, Y. J., & Pew, R. W. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors*, *37*(1), 85–104. https://doi.org/10.1518/001872095779049462

Parasuraman, R., Mouloua, M., & Molloy, R. (1997). Effects of adaptive task allocation on monitoring of automated systems. *Human factors*, *38*, 665–79. https://doi.org/10.1518/001872096778827279

Billman, D., Mumaw, R., & Feary, M. (2020). A model of monitoring as sensemaking: Application to flight path management and pilot training. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 244–248. https://doi.org/10.1177/1071181320641058

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* WH Freeman; Company.

Plant, K., & Stanton, N. (2014). The process of processing: exploring the validity of Neisser's perceptual cycle model with accounts from critical decision-making in the cockpit. *Ergonomics*, *58*, 1–15. https://doi.org/10.1080/00140139.2014.991765

Plant, K., & Stanton, N. (2013). What is on your mind? using the perceptual cycle model and critical decision method to understand the decision-making process in the cockpit. *Ergonomics*, *56*. https://doi.org/10.1080/00140139.2013.809480

Landman, A., Groen, E., Van Paassen, M. M., Bronkhorst, A., & Mulder, M. (2017). Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*, 001872081772342. https://doi.org/10.1177/0018720817723428

Wickens, C. D. (2003). Pilot actions and tasks: Selections, execution, and control. *Principles and Practice of Aviation Psychology*, 239–263.

Maurino, D. (2005). Threat and error management (TEM). *Canadian Aviation Safety Seminar (CASS)*, 18–20.

Helmreich, R. L. (2000). On error management: Lessons from aviation. *Bmj*, *320*(7237), 781–785.

Silva, S., & Hansman, R. (2015). Divergence between flight crew mental model and aircraft system state in auto-throttle mode confusion accident and incident cases. *Journal of Cognitive Engineering and Decision Making*, *9*. https://doi.org/10.1177/1555343415597344

Flight Standards Service. (2015). *Safety alert for operators 15 011* (tech. rep.). Federal Aviation Administration.

Hegg, J. W., Smith, M., Yount, L. J., & Todd, J. (1994). Features of active sidestick controllers. *AIAA/IEEE Digital Avionics Systems Conference. 13th DASC*, 305–308.

Hosman, R. J. A. W., Benard, B., & Fourquet, H. (1990). Active and passive side stick controllers in manual aircraft control. *1990 IEEE International Conference on Systems, Man, and Cybernetics Conference Proceedings*, 527–529. https://doi.org/10.1109/ICSMC.1990.142165

Garmin. (2011). *G1000 integrated flight deck pilot's guide* (tech. rep.).

van Leeuwen, J. K. (2020). *The effectiveness of simulating active problem solving in pilot training to deal with automation surprises* (Master's thesis). TU Delft.

Part III

Appendices

A

Scenario Description

This appendix contains the run sheets and .sce files. The run sheets were used by the experiment leader to configure for each run and give timely heading and altitude changes. The .sce files were programmed for DUECA and show settings and triggers for the scenarios.

Run 1

Description: Autopilot, Left engine out

Initialisation	
Mode	Autopilot
Scenario	901_AP_Eng
INCO	Clean_125kt_5000ft
Gear	UP
Flaps	UP
Throttle	55%

Significant Weather	
Turbulence intensity	0.1
Wind vel (m/s)	7
Wind dir degr	200

Instructions
Ensure proper **flight configuration** Select HDG and ALT T=30 'PH-SRS, speed 140 knots' T=90 'PRS, turn right heading 280' T=200s 'PRS, turn left heading 220' T=250 'PRS, speed 115 knots' T=270s 'PRS, climb and maintain 7000' Altitude = 6900 + 30s **EVENT** Left engine to 0 power in 60s

```
# Scenario
id                      901
inco                    Clean_125kts_5000ft.inco
APstate                 1
verticalFDstate         1
altitudeset             1
lateralFDstate  1
CDI                     0
fp_name         0
lat_zero                0.8990575503090124
lon_zero                0.06505984928938273
alt_zero                -3.35
psi_zero                0.0


windEvent
eventtime               0
enable_turb             1
turb_int                0.05
wind_vel                7
wind_dir                200
enable_windshear        0
fg_visibility           100000

massEvent
eventaltitude           6900
shift_x         0.0
shift_y         0.0
shift_z         0.0
shift_mass              0.0
shift_time              0.0

engineEvent
eventtime_after 30
power_left              0
power_right             1
max_rpm_left            2800
max_rpm_right           2800
engine_time_left        60
engine_time_right       0
```

Run 2

Description: Manual, Left engine out

Initialisation	
Mode	Manual
Scenario	902_M_Eng
INCO	Clean_125kt_3000ft_zeeland
Gear	UP
Flaps	UP
Throttle	55%

Significant Weather	
Turbulence intensity	0.1
Wind vel (m/s)	7
Wind dir degr	200

Instructions
Ensure proper **flight configuration** T=60 1:00min ‘PH-SRS, descend and maintain 1000’ T=180s 3:00min ‘PRS, turn left heading 120’ T= 300s 5:20min ‘PRS, turn right heading 220’ T= 330s ‘PRS, climb and maintain 5000’ Altitude = 4900 + 40s **EVENT** Left engine to 0 power in 60s

```
# Scenario
id                      902
inco                    Clean_125kts_3000ft_zeeland.inco
APstate                 0
verticalFDstate         0
altitudeset             1
lateralFDstate  0
CDI                     0
fp_name         0
lat_zero                0.8990575503090124
lon_zero                0.06505984928938273
alt_zero                0.0
psi_zero                0.0

windEvent
eventtime               0
enable_turb             1
turb_int                0.05
wind_vel                4
wind_dir                200
enable_windshear        0
fg_visibility           100000
fg_cloud0_alt           2000

massEvent
eventaltitude           4900
shift_x         0.0
shift_y         0.0
shift_z         0.0
shift_mass              0.0
shift_time              0.0

engineEvent
eventtime_after 40
power_left              0
power_right             1
max_rpm_left            2800
max_rpm_right           2800
engine_time_left        60
engine_time_right       0
```

Run 3

Description: Autopilot, Icing

Initialisation	
Mode	Autopilot
Scenario	903_AP_Ice
INCO	Clean_125kts_5000ft
Gear	UP
Flaps	UP
Throttle	55%

Significant Weather	
Turbulence intensity	0.1
Wind vel (m/s)	7
Wind dir degr	200

Instructions
Ensure proper **flight configuration** Select HDG and ALT T=30 ‘PRS, turn right heading 100, descend to 4000’ T=200 ‘PRS, turn right heading 180, climb and maintain 7500 Altitude = 7400 + 30s **EVENT** Icing accumulation in 60s

```
# Scenario
id                      903
inco                    Clean_125kts_5000ft.inco
APstate                 1
verticalFDstate         1
altitudeset             1
lateralFDstate  1
CDI                     0
fp_name         0
lat_zero                0.8990575503090124
lon_zero                0.06505984928938273
alt_zero                -3.35
psi_zero                0.0

windEvent
eventtime               0
enable_turb             1
turb_int                0.05
wind_vel                5
wind_dir                200
enable_windshear        0
fg_visibility           100000
fg_cloud2_alt           6000

massEvent
eventaltitude           7400
shift_x         0.0
shift_y         0.0
shift_z         0.0
shift_mass              0.0
shift_time              0.0

icingEvent
eventtime_after 30
ice_added_mass          100
ice_effectivity 0.1
ice_grow_time           100
```

Run 4

Description: Manual, Icing

Initialisation	
Mode	Manual
Scenario	904_M_Ice
INCO	Clean_125kts_5000ft
Gear	UP
Flaps	UP
Throttle	55%

Significant Weather	
Turbulence intensity	0.1
Wind vel (m/s)	7
Wind dir degr	200

Instructions
Ensure proper **flight configuration** T=30s 'PRS, climb and maintain 7000' T=60s 'PRS, turn left heading 190 T=270s 'PRS, descend and maintain 6000' T=400 'PRS, turn right heading 360' T=450 'PRS, climb and maintain 8000' Altitude = 7900 + 35s **EVENT** Icing accumulation in 60s

```
# Scenario
id                      904
inco                    Clean_125kts_5000ft.inco
APstate                 0
verticalFDstate         0
altitudeset             1
lateralFDstate  0
CDI                     0
fp_name         0
lat_zero                0.8990575503090124
lon_zero                0.06505984928938273
alt_zero                -3.35
psi_zero                0.0

windEvent
eventtime               0
enable_turb             1
turb_int                0.05
wind_vel                5
wind_dir                200
enable_windshear        0
fg_visibility           100000
fg_cloud2_alt           6000

massEvent
eventaltitude           7900
shift_x         0.0
shift_y         0.0
shift_z         0.0
shift_mass              0.0
shift_time              0.0

icingEvent
eventtime_after 35
ice_added_mass         100
ice_effectivity        0.1
ice_grow_time          100
```

B
Briefing

The following appendix will show the slides that were used in the briefing before the experiment.

Briefing PH-SRS

TU Delft Aerospace Engineering
Simon van den Eijkel

Doelen van vandaag

- Het realisme testen van verschillende in-flight events
- Tijdens manual en automatic flight
- Zijn de events geschikt voor toekomstig onderzoek?
- Hoe ervaren vliegers de events? Hoe worden ze gediagnostiseerd?

Schedule

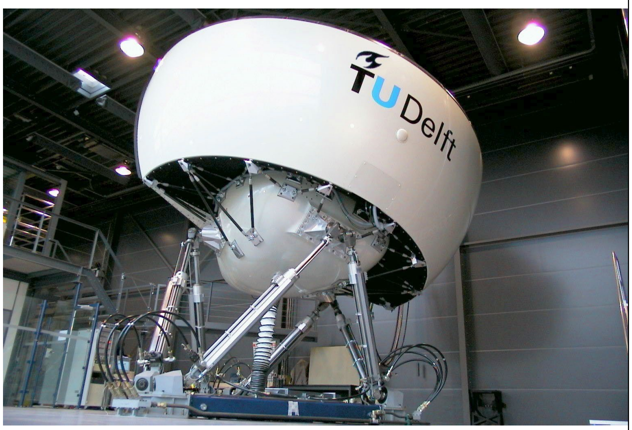
- 8:30 hrs: Walk-in
- 9:00 hrs: Briefing and documentation
- 9:30 hrs: Familiarisation and first experiment runs
- 10:30 hrs: Coffee break
- 10:50 hrs: Experiment runs
- 11:50 hrs: Debrief

Schedule

- 12:00 hrs: Walk-in
- 12:30 hrs: Briefing and documentation
- 13:00 hrs: Familiarisation and first experiment runs
- 14:00 hrs: Coffee break
- 14:20 hrs: Experiment runs
- 15:20 hrs: Debrief

SIMONA

- 6 DOF hydraulic motion system
- Kan op vele manieren worden geconfigureerd



Piper Seneca II

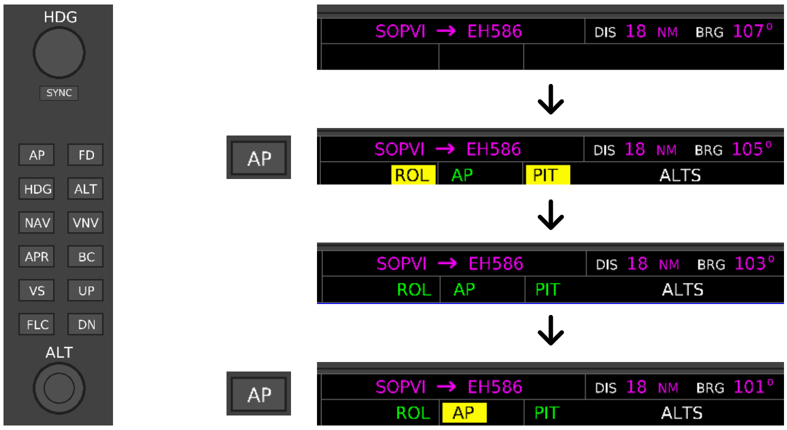
- Twin Engine propellor
- GFC700 Autopilot
- G1000 Avionics

Flight Deck

Flight Deck

Flight Deck

Flight Deck

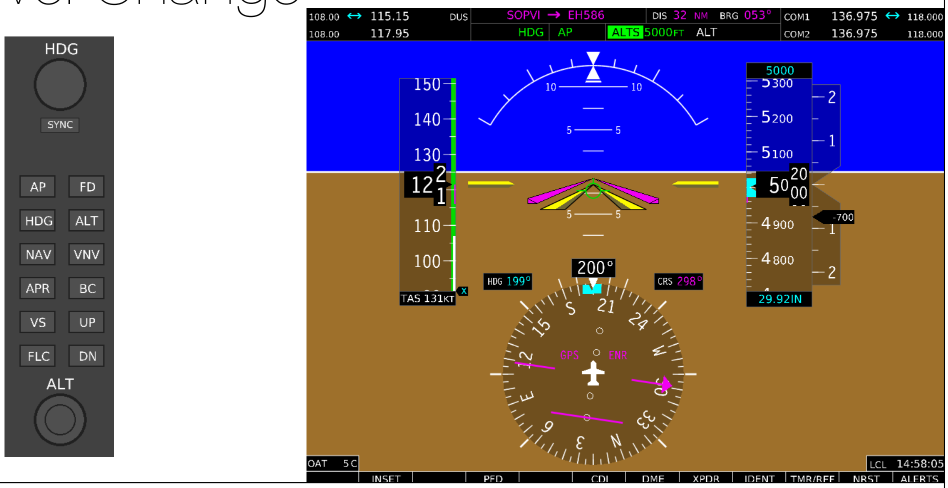


Level Change

Level Change

Level Change

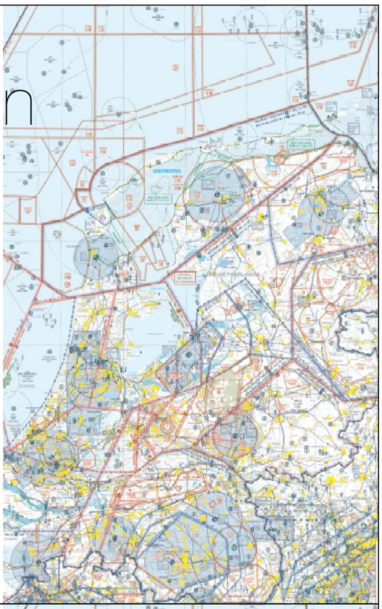
Level Change



Level Change

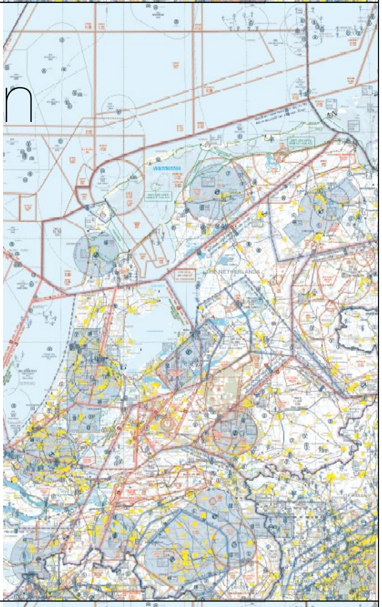
Example Experiment Run

- Initialisatie
- Korte briefing
- Instructie Manual of Autopilot vliegen
- ATC commands zullen begeleiden
- Korte debrief



Example Experiment Run

- Communiceer uw acties en relevante zaken die u opmerkt, zoals u dat zou doen met een Pilot Monitoring.
- ‘Switching to HDG’ ‘Snelheid is te laag.’
  ‘Hee, dit is vreemd.’
- De audio wordt opgenomen en geanalyseerd, om te onderzoeken hoe u de events ervaart.
- Na elk scenario stel ik een aantal vragen.

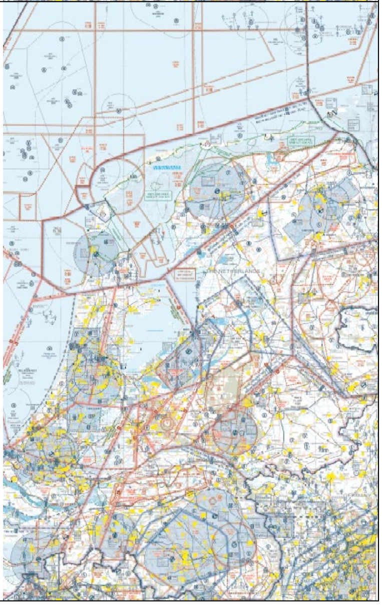


Familiarisation Flights

3 in total

- Eerste vlucht: Manual flight, landing.
- Tweede vlucht: Manual Take-off at Rotterdam followed by Autopilot practice.
- Derde vlucht: AP & Manual, Circuit and Touch-and-go.

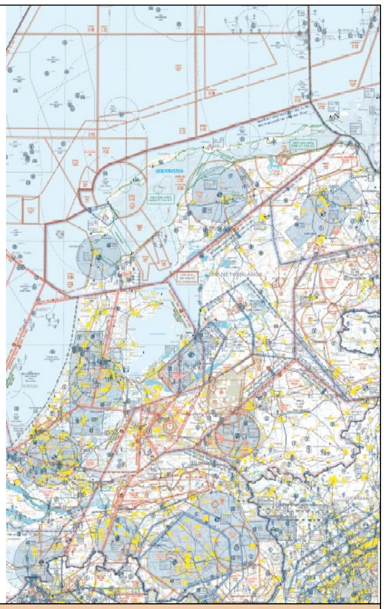
Focus op
1. Het vliegtuig leren kennen
2. Praten gedurende de vlucht, alsof met PM



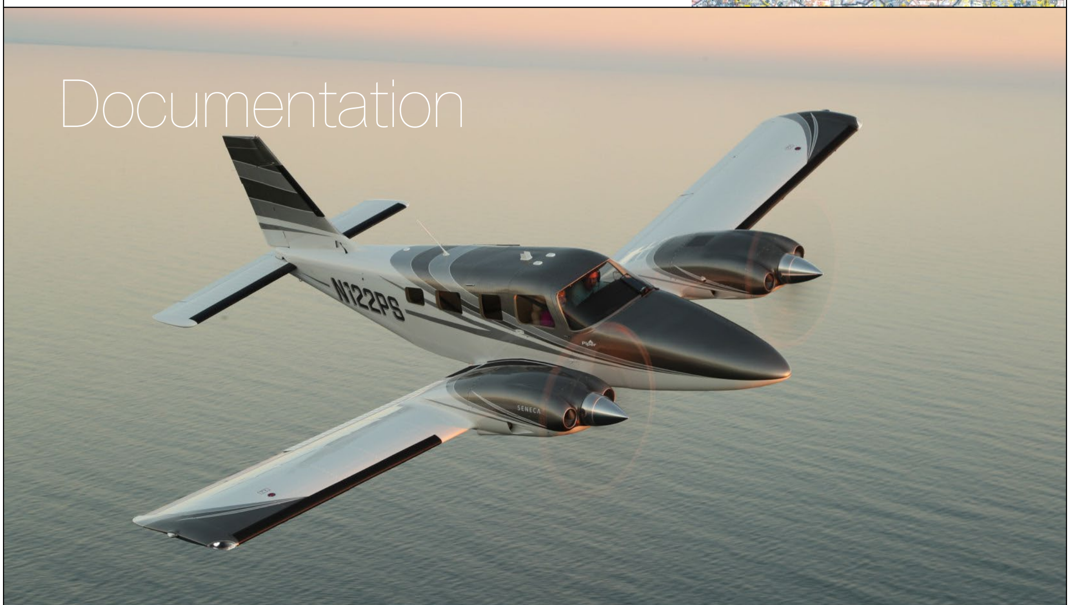
Flying Instructions

Takeoff:		**Landing:**	
Flaps	UP	Flaps	Land
V_r	80 kts	V_app	83 kts
V_2	92 kts	Gear	DOWN
Gear	UP		

- Configuratie en throttle lever moeten altijd voor de start van een scenario ingesteld worden.
- Speeds worden herhaald over de radio.
- Relatieve vrijheid tijdens runs, ook voor checklists
- Bij een event zal het scenario enige tijd erna stoppen, zodat je even de tijd hebt om het te analyseren en je directe reactie erop uit te voeren.

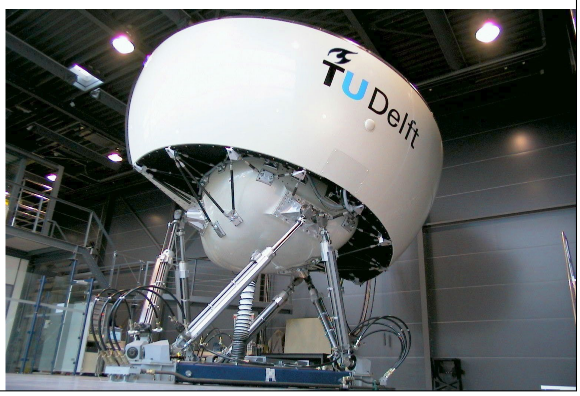


Documentation



SIMONA Safety Video

- https://www.youtube.com/watch?v=PXijsyJ3hro



C

Changes to the DUECA Project

To fit the specific needs of this project the DUECA Project has been altered. This appendix will describe these changes.

Addition of Icing Accumulation Hazard
The project now includes the ability to simulate icing accumulation in the simulation. This can be programmed using $icingEvent$ in the .sce file. Parameters that can be used to tune this hazard are ice_added_mass, $ice\_effectivity$, and ice_grow_time. Once the event is triggered the aircraft will get a pitch-up tendency and the elevator will become less effective, giving the controls a sluggish feeling.

Added trigger functionality
The parameter $altitudeset$ in combination with $eventaltitude$ can now be used to define when an event is triggered whilst climbing or descending through a set altitude. It is a binary variable with 1 representing a climbing trigger and 0 a descending trigger. It was considered to add a heading trigger, but this was deemed not necessary for this project. Future studies might benefit from such a trigger.

Back-driven Controls
The DUECA Project has been altered to include backdriven controls. This means that when the aircraft is controlled by the autopilot, the control column visualizes the inputs of the autopilot.

Added 3D Model Seneca
When flying the Piper Seneca III, the pilot will see the two engines as they are mounted forward. For this experiment an outside model of the Seneca was added for realism. This is a stripped 3D-model with wings and engine pods.

D
Documentation Informed Consent

The following page will show the Informed Consent Document that was signed by each participant before the experiment started. This documents informs the participants of the data that is collected and the associated risks. The experiment has been approved by the Human Research Ethics Committee (HREC) of TU Delft.

Informed Consent Form

This form is to be used to provide informed consent for participation in an experiment conducted in the SIMONA Research Simulator at Delft University of Technology.

Instructions

You are being asked to voluntarily participate in an experiment conducted in a flight simulator. By signing this form, you are giving your consent to participate in the experiment. Please read the following information carefully before signing this form.

Description of the Experiment

The experiment will involve participating in a flight simulation experiment for a maximum of 3 hours. During the experiment, to participant is asked to fly a total of 4 familiarisation runs and 8 experiment runs. The participant is asked to assess the realism of the simulation and anomalies that might appear. After the experiment, the participants will have room to ask any questions.

Potential Risks

Participation in this experiment is voluntary. The participants might get motion sickness caused by the simulator. The participants can refuse to give answers to questions and stop the experiment at any time, without having to give a reason.

Confidentiality and Data collection

Personal information such as your name and email address, will be collected and used only by the experimenter for planning purposes. All other information gathered during the experiment, including age, flight experience, control inputs, audio recordings, comments, and remarks, will be stored and processed anonymously. The anonymized data will be used for analysis and may be published in anonymous or aggregate form in an MSc thesis report and potentially in scientific publications. The anonymized data will be archived and may be used for future studies. You can withdraw your data at any time by sending an email to the SIMONA lab manager, O.Stroosma@tudelft.nl, and your data will be removed from all analyses that have not yet been published. For the sole purpose of data withdrawal requests, a secure record linking your personal information with the identification used for anonymizing the data will be kept. This record is only available to the SIMONA lab manager and no members of the research team or future researchers will have access to it.

Compensation

There will be no compensation for the participants time and participation in the experiment. There will be a compensation for the travel cost to Delft University of Technology at a rate of €0.19/km.

Signatures

I have read and understood the information provided above. I agree to participate in the experiment and I understand the risks involved.

Name: ___________________________________

Date: ___________________________________

Signature: ___________________________________

For further information please contact Simon van den Eijkel, +31 6 40838324, s.vandeneijkel@student.tudelft.nl

E

After Run Questions

The following appendix will contain the form with questions that were asked after each run.

Vragenformulier na scenario

Proefpersoon nummer: ……………………… Scenario ID: ……………………….

Heb je tijdens deze vlucht iets bijzonders gemerkt? Ja / Nee

Als ja:

Wat denk je dat er gebeurde?

………

………

………

Wat was het eerste waaraan je merkte dat er iets gebeurde?

………

………

Wat deed je of waar keek je naar om de gebeurtenis te confirmeren of diagnostiseren?

………

………

Welke acties ondernam je vervolgens (of zou je ondernemen als dit kon in de simulatie)?

………

………

………

Hoe realistisch vond je de gebeurtenis? Kun je dit toelichten?

………

………

………

Heb je direct een callout gegeven? Ja / Nee

Als nee, waarom niet?

………

………

………