Pilots' decision-making during unstabilized approaches: Group-level policy capturing for cognitive assistance

Laura Salvan[1, a], Alexandre Marois[1], Bénédicte Chatelais[1],
Daniel Lafond[1] and Amandine Audouy[2]
[1]Thales Research and Technology, Québec, Canada
[2]Thales, Toulouse, France
[a]laura.salvan@thalesgroup.com

INTRODUCTION

Among all international flight fatal accidents between 2011 and 2020, 54% happened during the final approach, or in the landing phase (Boeing, 2021). Moreover, Reynal et al. (2017) found that half of the pilots persisted in an erroneous landing decision rather than performing a go-around to retry the approach. Accident analyses show that go-around procedures are often improperly performed due to their complexity, high time stress, and scarcity, which complicate training (Dehais et al., 2017). Besides, a study revealed that only about 3% of unstable approaches result in a go-around (Shah & Campbell, 2018). In light of this evidence, better support of pilots' decision-making during landing approaches is warranted. In previous decades, pilots were used to manually control the aircraft compared to today which is now more about global mission management and automation. According to Myers and Starr (2021), existing automation solutions can perform some tasks better than humans (e.g., monitoring a tremendous amount of parameters simultaneously) but lack one thing that humans have: judgment. Recent trends focus on the design of more ecological human-machine interfaces to support decision-making for pilots and enhance safety level. One strategy for developing decision aids that are tailored to user judgments is called policy capturing (e.g., Cooksey, 1996). This technique relies on statistical or machine learning to infer a decision model based on previous decisions (e.g., a medical diagnosis) as well as on an ensemble of features used to explain the decision (e.g., a list of symptoms; see Backlund et al., 2009). Unstable approaches and the decision to either land or go-around seems an ideal use case for investigating policy capturing, particularly because context-based human expert judgments remain critical and have not readily been converted into tractable procedural rules. Indeed, policy capturing is intended as a method for deriving cognitive aids for individuals having to make decisions in real-life situations (Goldberg, 1970).

The Cognitive Shadow is an online modeling and decision tool developed by Thales Research and Technology Canada that automatically learns a user's decision pattern and provides recommendations based on past decisions with similar attributes. Expert's decisions are collected and sampled to train linear and nonlinear machine learning models concurrently (namely logistic regression, decision trees, k-nearest neighbors [KNN], neural networks, naïve Bayes, support vector classifiers and random forest) to provide a prediction originating from a specific model or an aggregate of the models. Marois et al. (under review) demonstrated that the judgement policies of three pilots facing an unstable approach on static and low-fidelity settings could be predicted with an accuracy of approximately 88% (and with an accuracy of 100% with all experts combined). This suggests that the model was perfectly representative of the decision-making patterns of the experts when combined using a majority rule.

The study reported herein aimed at extending Marois et al.'s (under review) work by further evaluating the land/go-around group modeling capacity of the Cognitive Shadow. Group modeling may lead to more or less accurate predictive modeling when involving different individual patterns (Lafond et al., 2009; Lee & Webb, 2005). Indeed, the effectiveness at

capturing a common decisional pattern may vary according to pilots' judgment homogeneity/heterogeneity. Such group-based approach thus needs to be further evaluated. The second objective involves improving ecological validity of the data collection settings compared with the capture interface of the previous study (i.e., reading a list of contextual factors instead of looking at cockpit quadrants). In fact, Marois et al.'s approach might have impacted the pilots' habits to screen the situation and reduce mental resources availability to take appropriate decision. For that reason, in the current study, a Primary Flight Display (PFD) instrument was added to the interface in order to improve the ecological validity of the experiment.

METHOD

Participants

Four aircraft pilots took part in the study. They had an average of 18.8 years of experience in piloting including on a typical single aisle aircraft.

Material & Procedure

Model parameters used for the experiment comprised 21 features: Headwind speed (Knots), Lateral wind speed (Knots), Runway headwind speed (Knots), Runway lateral wind speed (Knots), Distance with nearest aircraft (Nautical miles), Lateral deviation (Dot tenths), Δ Lateral deviation (Dot tenths), Vertical deviation (Dot tenths), Δ Vertical deviation (Dot tenths), Engine power N1 (%), Δ Engine power N1 (%), Pitch (degrees), Δ Pitch (degrees), Roll (degrees), Δ Roll (degrees), Vertical speed (feet per minute), Aircraft speed (IAS; Knots), Δ IAS (Knots), Theoretical approach speed (Vapp; Knots), IAS – Vapp difference (Knots), Precipitation on aircraft ratio (ratio [0,1]). The ranges of features were refined to consider expert inputs in order to optimize the representativeness of the cases presented to the pilots. The delta (Δ) features aimed at helping pilots to take their decision while accounting for the recent trend (change in the last five seconds). The pilots were shown a custom web interface illustrating the main features through the Primary Flight Display (PFD) instrument on the left and the others on the right (see Figure 1). The dashboard displayed a series of different approach situations whose attributes were generated by the Cognitive Shadow.
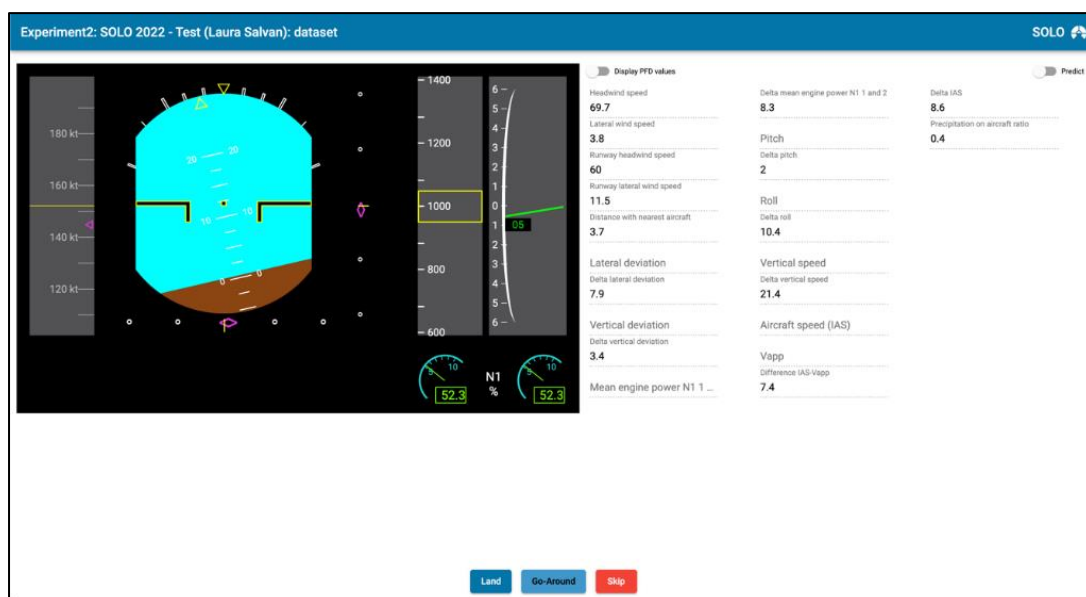

Figure 1: Cognitive Shadow web interface illustrating some features
through PFD instruments and other features on a side-list.

Participants were asked to decide whether they would land or go-around according to a series of features representative of different unstable approach situations using one of the three buttons (Figure 1) for the classification (i.e., Go-around, Landing, Skip). The *Skip* option allowed pilots to discard any cases deemed unrealistic. The experiment comprised two sessions. The first session was a knowledge capture phase aiming at collecting the participants' decisions to automatically train the decision-making models. Each pilot was shown 50 different cases from a predefined dataset designed with realistic value ranges. The second session represented the test phase using the majority response as the ground-truth (i.e., the correct decision). Each pilot was shown the same 30 cases and the Cognitive Shadow provided recommendations and decision explanations when their decision differed from the group model predictions. The group model was trained from the 200 (50 decisions × 4 pilots) decisions of the first session. While 50 decisions per expert seems low, the group modeling strategy allowed to sum decisions from all the pilots and thus increased the number of training data. It also helped to demonstrate the frugal learning capacities of the Cognitive Shadow to learn from small amounts of data.

RESULTS

The aggregation method selected to discriminate the output of the multiple algorithms is called *Best Model*. This method defines the recommendation according to the model's accuracy that was better at predicting the classes (i.e., Go-around, Landing) using a standard 10-fold cross-validation procedure. The best model identified was a KNN with an accuracy of 91.19%. Then, the group-of-experts model created from all the participants' decisions in the first session was evaluated with the second session's ground-truth responses (i.e., majority response). If no majority could be defined between pilots' decisions (e.g., two go-around vs. two land), the prediction of the Cognitive Shadow was chosen to break the tie (this occurred only once). The ground-truth on the test set ($N = 30$) established 27 landing decisions and 3 go-around decisions. Table 1 displays the confusion matrix of the decisions on the test phase for the four pilots and for the group-of-experts model as well as the accuracy, sensitivity, and specificity measures (while considering landing as the "positive" action and go-around as the "negative" action). On the test phase, Pilot 1 reached perfect accuracy (100%), Pilots 2 and 3 both reached a 90-% accuracy, and Pilot 4 reached an accuracy of 86.21%. The group-of-experts model correctly predicted 96.67% of the decisions on the test phase according to the ground-truth established by the majority decision. One pilot skipped a case considering having unrealistic values; consequently, the test phase comprised 119 rather than 120 (30 × 4 pilots) decisions.

	Pilot(s) decision	Ground-truth decision			Accuracy (%)	Sensitivity (%)	Specificity (%)
		Landing	*G-A*	*Total*			
Pilot 1	*Landing*	27	0	27	100.00	100.00	100.00
	Go-around	0	3	3			
Pilot 2	*Landing*	25	1	26	90.00	92.59	66.67
	Go-around	2	2	4			
Pilot 3	*Landing*	24	0	24	90.00	88.89	100.00
	Go-around	3	3	6			
Pilot 4	*Landing*	23	0	23	86.21	85.19	100.00
	Go-around	4	2	6			
Group predictions	*Landing*	27	1	28	96.67	100.00	66.67
	Go-around	0	2	2			

Table 1: Confusion matrix and measures of accuracy (%) for the four pilots and for the group-of-experts in the test phase. Ground truth represents the majority decision.

Interrater agreement for the test phase led to a Fleiss' kappa of $\kappa = .42$ ($p < .001$, 95% CI [0.27, 0.56]), representative of a moderate agreement between the four raters (cf. Landis & Koch, 1977). The unique performance of each pilot on the ground-truth (i.e., before the Cognitive Shadow provided retroaction when the system considered a mismatch between the pilot decision and the group-of-experts model prediction) was also evaluated and compared with the joint cognitive system (decision following feedback from Cognitive Shadow). Performance according to the ground truth revealed a mean increase of 3.1% ($SD = 2.1$) when pilots considered the decision-aid recommendation (i.e., joint cognitive system) compared to the human-only system, and this increase was systematically positive. Because the joint cognitive system was anticipated to be better than the human-only system, a Wilcoxon signed-rank test was carried out with a unilateral critical alpha of 0.05, supporting that the joint cognitive system produced a significant increase in performance accuracy on the ground-truth, $Med_{human} = 86.67\%$, $Med_{joint} = 90.00\%$, $Z = 2.04$, $p = .0415$. Finally, the "land" class was better predicted (~86%) compared to the go-around class (~77.5%). This can be explained by the large amount of land cases on which the Cognitive Shadow could be trained.

DISCUSSION

The approach and landing phases represent a major proportion of all commercial aircraft accidents (IATA, 2016). One way to support humans in high-risk environments such as aviation is through decision-support systems (Sarter & Schroeder, 2001). Nevertheless, predicting pilot decision-making is difficult because they have their own perceptions of risk during the approach phase which impact their go-around decision (Shah & Campbell, 2018). Herein, the Cognitive Shadow decision-aid tool created a group-of-experts model with the decisions of the four pilots that was able to predict 96.67% of the classifications on the test phase. This experiment, consistent with Marois et al. (under review), suggests that this model was judiciously representative of the experts' decision-making patterns. These results imply the potential of creating useful decision aids by extracting pilots' knowledge using policy capturing. Individual-level modeling was not performed herein since it would have required a lengthier data collection. Yet, group-level modeling provided the advantage of reassuring pilots who appreciated the idea that the expert's community endorsed or challenged their decisions through the group's model predictions. Moreover, the increase of performance, due to pilots considering the recommendation of the Cognitive Shadow, demonstrates the potential benefit of the joint cognitive system approach (as a distinct strategy from automation).

On another note, in order to balance land and go-around cases, a general rule derived from interviews with the expert pilots was used. It however seems that the application of the rule changed when pilots were being shown the cases. This experiment indeed revealed 90% of the cases were classified "land" which corroborates the propensity of pilots to land instead of go-around. Notably, the Airbus research program aims at preventing flight crews from the perseveration syndrome which characterizes the gathering of pilots' mental efforts toward a unique objective even if it is unsafe (Dehais et al., 2010). Future work will make use of the models generated in this study to build datasets balanced between lands/go-arounds and avoid any bias toward landing. The decision-aid recommendations in case of flawed reading of events by the pilot could represent cognitive countermeasures to enhance pilots' attention shifting capabilities. Highlighting critical situation variables would thus be of importance as a complement to alerts and decision recommendations. Thus, the human propensity of perseveration and landing at all costs (Causse et al., 2013; Curtis & Smith, 2013) could likely be mitigated by such new technologies. Nonetheless, more extensive human factors studies, including in more dynamic settings, are required to investigate longer term impacts of interaction with such decision aids. Indeed, a stronger relationship between human decision

makers and automated decision aids could be built with additional insights into introducing artificial cognitive agents into a decision-making context (Mosier & Skitka, 1996).

REFERENCES

Boeing. (2021). *Statistical Summary of Commercial Jet Airplane Accidents: Worldwide Operations 1959 – 2020*. Retained from the Internet under: https://www.boeing.com/resources/boeingdotcom/company/about_bca/pdf/statsum.pdf.

Backlund, L. G., Bring, J., Skånér, Y., Strender, L. E., & Montgomery, H. (2009). Improving fast and frugal modeling in relation to regression analysis: Test of 3 models for medical decision making. *Medical Decision Making*, *29*, 140–148.

Causse, M., Péran, P., Dehais, F., Caravasso, C. F., Zeffiro, T., Sabatini, U., et al. (2013). Affective decision making under uncertainty during a plausible aviation task: an fMRI study. *NeuroImage*, *71*, 19–29.

Curtis, B., & Smith, M. (2013). *Why are Go Around Policies Ineffective*. Mississauga, Canada: Presage Group.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. San Diego, CA: Academic Press.

Dehais F., Tessier C., Christophe L., & Reuzeau F. (2010). The perseveration syndrome in the pilot's activity: Guidelines and cognitive countermeasures. In: P. Palanque, J. Vanderdonckt, & M. Winckler (Eds.), *Human Error, Safety and Systems Development. Lecture Notes in Computer Science, vol. 5962* (pp. 68–80). Berlin, Germany: Springer.

Dehais, F., Behrend, J., Peysakhovich V., Causse, M., & Wickens, C. D. (2017). Pilot flying and pilot monitoring's aircraft state awareness during go-around execution in aviation: A behavioral and eye tracking study. *The International Journal of Aerospace Psychology*, *27*, 15–28.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological bulletin*, 73(6), 422.

International Air Transport Association (IATA). (2016). *Unstable Approaches, Risk Mitigation Policies, Procedures and Best Practices. Final report.*

Lafond, D., Lacouture, Y., & Cohen, A. L. (2009). Decision-tree models of categorization response times, choice proportions, and typicality judgments. *Psychological Review, 116*, 833–855.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*, 605–621.

Marois, A., Lafond, D., Audouy, A., Boronat, H., & Mazoyer, P. (under review). Policy capturing to support pilot decision-making: A proof of concept study. *Aviation Psychology and Applied Human Factors.*

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

Myers, P. L., & Starr, A. W. (2021). Single pilot operations in commercial cockpits: Background, challenges, and options. *Journal of Intelligent & Robotic Systems*, *102*, 19.

Reynal, M., Rister, F., Scannella, S., Wickens, C., & Dehais, F. (2017). Investigating pilot's decision making when facing an unstabilized approach: An eye-tracking study. *Proceedings of the 19th International Symposium on Aviation Psychology,* 335–340.

Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, *43*, 573–583.

Shah, S., & Campbell, A. (2018, June). Analyzing pilot decision-making using predictive modeling. In *Proceedings of the ICRAT 2018, 8th International Conference on Research in Air Transportation*, Castelldefels, Catalonia, Spain, 26–29 June 2018.