Personalized and transparent AI support for ATC conflict detection and resolution: an empirical study

Carl Westin
Science and Technology
Linköping University
Linköping, Sweden
carl.westin@liu.se

Clark Borst, Erik-Jan van Kampen, Tiago M.M.Nunes
Control and Simulation
Delft University of Technology
Delft, The Netherlands
c.borst@tudelft.nl

Supathida Boonsong
Research, Innovation and Digitalisation
LFV Swedish Air Navigation Service
Norrköping, Sweden
supathida.boonsong@lfv.se

Brian Hilburn
Center for Human Performance Research
Voorburg, The Netherlands
brian@chpr.nl

Matteo Cocchioni, Stefano Bonelli
Deep Blue SRL
Rome, Italy
matteo.cocchioni@dblue.it

***Abstract*— Artificial Intelligence provides both opportunities and considerable challenges to the continued growth of Air Traffic Control (ATC) services. This paper presents a study where a personalized and transparent machine learning decision aid for ATC conflict resolution was built and empirically evaluated with air traffic controllers. Multi-site simulations were conducted with 34 controllers working together with an AI agent to solve conflicts between aircraft in enroute traffic scenarios. Resolution advisories varied in conformance (degree of personalization) and transparency. Main effects of conformance were found on controllers' resolution performance and response to advisories in terms of acceptance and ratings of agreement and similarity to own solution. The separation distance aimed for by the advised solution was found to be particularly important for the response to optimal advisories. More positive responses were measured for controllers whose separation margin preferences was closer aligned with the advisory. The study provides the aviation community with knowledge on how conformal and transparent AI support systems affect operators' responses to system-generated resolution advisories.**

***Keywords*—Machine learning; Artificial intelligence; Air Traffic Control; Conflict detection and resolution; Personalization; Strategic conformance; Transparency; Explainability; Decision support systems**

I. Introduction

Advances in the Artificial Intelligence (AI) subdomain of Machine Learning (ML) provide opportunities to the continued growth of Air Traffic Control (ATC) services. Particularly for the conflict detection and resolution (CD&R) process, ML can offer potential workload, efficiency, and safety benefits. However, mismatches in AI and human problem solving strategies can negatively influence human acceptance and willingness to interact with these systems. ML applications tend to be difficult to understand for the operators it intends to support. A core human factors challenge in safety-critical systems is how to design systems so that humans can understand its output and behavior. Overcoming this hurdle is considered essential for achieving meaningful teamwork between humans and autonomous agents [1]–[3].

The constructs of agent transparency and Explainable AI (XAI) have been suggested for overcoming understanding issues and support trust development, acceptance, and situation awareness. Here, transparency refers to the degree to which the system makes its internal processes apparent to the operator. Achieving the potential of AI decision aids implies solutions to be different from those preferred by the human, i.e., in a nonconformal way. Transparency is therefore envisioned to be most beneficial in situations where there is a mismatch between the human and automation solution.

AI systems are often designed to achieve optimal performance in a given domain, with little regard to individual preferences and needs. Yet, ML systems offer a unique ability to adapt to individual preferences in decision making strategies. Research has shown benefits from systems that adapt their behavior and performance to the individual user. Here, system personalization, or its strategic conformance, refers to the apparent decision making strategy match between human and automation/AI systems [4]. Conformal (personal) automation may reduce the need for transparency, at least in situations when the automation's solutions match those of the human.

This study focuses on two constructs expected to underlie human-AI interaction: the conformance of ML models and the transparency of their output (i.e. solution). This paper presents original research from experimentally manipulating these two constructs in simulations with air traffic controllers (ATCOs) to explore their main and interaction effects on a broad number of ATCO perceptions and performance measurements. This article builds on the theoretical framework of conformance and transparency developed in the MAHALO project and presented in Westin et al. [5].

Next, Section II overviews previous work in relation to conformance, transparency, and AI applications to ATC CD&R.

This project has received funding from the SESAR Joint Undertaking (JU) under grant agreement No 892970. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the SESAR JU members other than the Union. The opinions expressed herein reflect the authors' view only.

Section III presents the study method and a brief description of the developed ML models for resolving conflicts. Section IV presents the effects of ML model conformance and advisory transparency on ATCO's response to resolution advisories. The implications of the findings to the ATC community is discussed in Section V.

II. Theoretical foundation

A. Personalized decision support

Researchers have advocated for *personalized* systems that conform to an individual's needs and preferences [4], [6]–[8]. *Conformal* systems supports understanding by providing a solution that, in appearance, matches the strategy or solution preferred by the individual. Because conformance is an attribute of the system, it requires that the system knows the individual's preferred decision making strategy or solution.

In ATC, Hilburn et al. [9] (the MUFASA project) simulated conformal automation via unrecognisable replays of ATCOs' previous CD&R performance. Results from sixteen ATCOs showed that when given conformal resolution advisories (in this case replays of their own solutions), they accepted and agreed more with advisories and responded to them faster compared to non-conformal advisories. For solving two-aircraft conflict situations, Regtuit et al. [10] developed a group conformal Reinforcement Learning (RL) agent that replicated ATCO-like CD&R strategies based on 'best practices'. Tran et al. [11] had a RL system learn from ATCO's conflict resolutions during training, approaching conformance on a group level. Van Rooijen et al. [12], [13] developed and trained a Supervised Learning (SL) model using Convolutional Neural Networks (CNN) on individual ATCOs' CD&R performance to output conformal (personal) conflict resolution advisories. As input, the SL model used pixel data from interface images of the Solution Space Diagram (SSD) CD&R support tool that captured the ATCO's conflict solutions. The model was tested in an experiment with twelve novices (no ATC experience, but some ATC familiarity). The model was trained with data from each participant, resulting in twelve personal models. Results showed that the conformal model was able to capture differences in individuals' CD&R strategies. The model's prediction performance was better for participants who were more consistent in solving conflicts over time (significant positive correlation). The conformal models were found to perform significantly better than the reference group models (trained on combined participant data).

B. Transparent decision support

Transparency is considered a key attribute of automation and autonomous system, for making its behavior understandable to the human [2], [3]. Research on *automation and agent transparency* aims to increase human understanding, trust, and acceptance of the system [14], [15]. Transparency has been used to explain the behavior of agents and robots [16], why a decision support system might err [17], or how close the automation is to its performance envelope [18].

Approaches for achieving transparency are known in the AI community as *explainable AI* (XAI) [19] and machine learning *interpretability* [20]. In the context ML, transparency becomes intrinsically difficult to achieve due to the amount of data processed and complexity of the systems (e.g., multiple deep layers, number of rules) that greatly exceed human abilities to make timely sense of the data. A common explainability approach for deep neural networks, which can be used in either (un)supervised learning or reinforcement learning, is to visualize which parts of the input (features) are most important in generating an output [21]. Example methods in image classification using convolutional neural networks (CNNs) are the Pixel-Wise Decomposition (PWD), which uses heatmaps to visualize individual pixels of the input image that determine the output [22]. Another approach is the Visual Back Prop (VBP) method, which uses masks to visualize the set of pixels in the input image that determine the output [23].

There is a shortage of empirical findings to form stable conclusions on the benefits of transparency and how to visualize and apply transparency in interface design [24], [25]. Research on AI explainability has focused on building ML models and deriving novel explanation methods [26] while neglecting the underlying psychology of the end user - What do users need? How are explanations presented and perceived? How can users interact in a dialogue with the system?

C. AI applications to CD&R

Previous studies on ML methods for CD&R have largely focused on the conflict resolution problem using RL approaches [11], [27]–[30], and only a few have focused on conflict detection [31], [32]. Some studies have developed conformal ML models [10]–[12], but none have explored transparency and explainability issues. The ML method explored by Brittain and Wei [28] is among the more advanced, relying on deep RL, Proximal Policy Optimization (PPO), and a Deep Distributed Multi-Agent Variable framework with attention networks. Tran et al. [11] and Pham et al. [27] explored an advanced RL approach making use of a Deep Deterministic Policy Gradient (DDPG) algorithm. Most previous approaches have restricted CD&R to a 2D representation of the environment (aircraft fixed to one altitude) and limited resolution maneuvers to heading changes. As an exception, Mollinga et al. [29] considered all three resolution types (heading, speed, altitude) for solving conflicts.

III. Method

The experimental design was a 3x3 within participant design varying model conformance and advisory transparency. Simulations were conducted in Italy (SIM1) and Sweden (SIM2). The research question explored how model conformance and advisory transparency of a ML CD&R decision support system affected ATCOs' performance and advisory response in solving conflicts. Each simulation comprised three steps: First a training pre-test was conducted that recorded ATCOs' resolution strategies as they solved two-aircraft closing conflicts. Second, the resolution data was used to train a

ML group model and synthetically derive personal models. A third optimized ML model was developed, *nonconformal* to ATCO preferences (i.e., not making use of ATCO resolution data as input) that tried to find an optimal solution according to objectively defined parameters. The three models were used to generate resolution advisories for use in the main experiment. Third, in the main experiment the same ATCOs interacted with an AI agent to solve conflicts, where the agent's resolution advisories varied in conformance and transparency.

A. Participants

In total, 34 ATCOs took part in the study. Erroneous data was observed for two participants, resulting in a final sample of eighteen ATCOs in SIM1 and fourteen ATCOs in SIM2. In SIM1, age varied between 35-59 years (M = 45.9, SD = 7.6) and experience between 9-36 years (M = 21.3, SD = 7.3). In SIM2, age varied between 32-55 years (M = 43.0, SD = 7.2) and experience between 4-30 years (M = 16.5, SD = 7.6).

B. Independent variables and dependent measures

ML model conformance reflected the similarity of the ML model with ATCOs' resolution preferences for specific scenarios. Three conditions were used: 1) personal models (matching the ATCO's preference), 2) a group model (matching the preference of the group of ATCOs in the study sample), 3) and a nonconformal optimized model (disregarding ATCO preferences). The models generated advisories that differed in timing, aircraft selected, heading direction and value, and target closest point of approach (CPA). Advisory transparency had three conditions that varied the amount of information provided about the underlying rationale for an advisory, reflecting the target CPA of the advisory. Figure 1 depicts the advisory transparency conditions: 1) low: vector line (T0); medium: vector line and SSD (T1); and 3) high: vector line, SSD, and text-based explanation (T2). The T0 condition corresponded to a baseline level of transparency where no additional information is provided beyond the actual advised heading. The T1 condition was based on the SSD that integrates speed (by concentric green rings) and heading (via red no-go zones) in visualising the relative position of another aircraft. The T2 condition added a text explanation of the chosen solution including target CPA. The seven dependent measures collected are shown in Table I.

C. Conformal machine learning models

Two ML models were created: a SL model and a RL model. The SL model created for generating personal and group advisories was based on previous work [13]. The model uses CNN to process pixel (128x64) SSD graphical data as input. For a given conflict, the SL model strove to find the best match of the conflict situation (input data in terms of images of conflict situation) with a label. For a two aircraft conflict situation, the SL model comprised different labels reflecting alternative solutions with varying heading values. The sixteen labels were made up by five degree heading intervals ranging from -45 to +45 degrees deviation from the current heading (except for heading values around zero degrees that were labeled -10 to -1 and +1 to +10). During training of the SL model, a participant's conflict resolutions were classified according to these labels. E.g., a person's solution that consisted of turning aircraft A 15 degrees to the right was assigned to the label: aircraft A, 11-15 degrees. Because of limited data, the learning stability and performance of the personal ML models was inadequate. Personal models (one for each participant) were therefore synthetically created from analysing consistent patterns in individual ATCO's resolutions to repeated conflict situation in the training pre-test simulation. This resulted in 34 unique personal models. For the group conformal model, sufficient data were collected to allow adequate model performance. The SIM1 group model was trained on SIM1 data only (720 resolution samples), while the SIM2 group model was trained on both SIM1 and SIM2 data (1296 resolution samples).

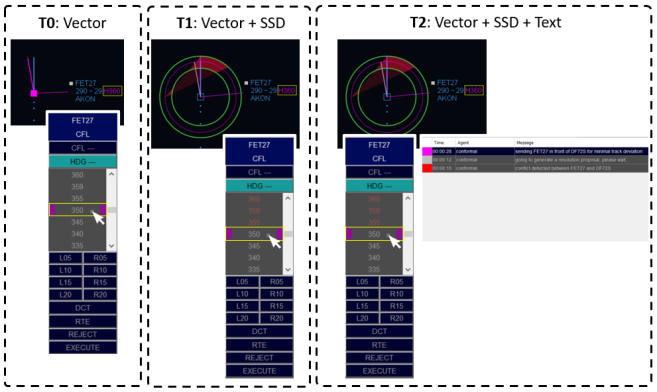


Figure 1. Transparency conditions in SectorX.

TABLE I. DEPENDENT MEASURES

Measure	Description
Advisory acceptance	Ordinal categorical SIM1: four-point scale; SIM2: five-point scale). **Accept** (accept as given); **Nudge** (preserve aircraft choice, heading type, and direction but nudge heading by e.g., adding 5 degrees); **Adjust** (preserve aircraft choice and heading type but change heading direction by e.g., turn left instead of right); **Change** (preserve aircraft choice but change clearance type to vertical); and **Reject** (only in SIM2). Rejecting the advisory allowed for interaction with the other aircraft involved in the conflict (not possible earlier). See Figure 2 for how interactions with an advisory was implemented in the SectorX simulator.
Agreement	0-100 rating scale.
Advisory conformance	Six point likert rating scale (1 = disagree highly, 6 = agree highly) with statement: "The system solved the conflict the same way I would have."
Advisory understanding	Six point likert rating scale (1 = disagree highly, 6 = agree highly) with statement: "I can understand why the system suggested that solution."
Workload	0-100 rating scale.
Response time	Measured in seconds from onset of advisory to response (execute/reject button pressed).
Delta CPA distance	Difference in nautical miles (nm) between proposed separation distance of advisory and achieved separation distance in solution (as modified by participant).

For generating optimized advisories, a RL model was created that did not use ATCO derived training data. The RL model used sector and traffic information (e.g., aircraft location, velocity, heading) as input, as well as information on the given traffic advisory (pixel data). A Q-Learning agent was coupled with the modified voltage potential (MVP) model of CD&R, which achieves separation assurance by representing other aircraft and destination as similar- and opposite potentials, respectively [33]. In both SIM1 and SIM2, the optimized RL model corresponded to this Q-learning algorithm adjusting the parameters that determine the behavior of the MVP algorithm. The main parameters the Q-learning agent adjusted were the separation distance and lookahead time at which the MVP algorithm would provide the heading change commands. The RL model was optimized for avoiding loss of separation and minimizing flight path deviation. Because of differences in options given to the agent for SIM1 and SIM2, e.g., for possible lookahead times and CPA, the advisories output by the RL model were different between simulations. There is no explicit model of the environment provided to the agent and the vectors are selected to minimize the cost function.

D. Simulator and scenarios

The SectorX simulation platform was used as a radar interface. SectorX is a Java-based medium-fidelity ATC research simulator. The interface design and support tools (e.g., the Verification of Separation and Resolution Advisory - VERA tool) was based on actual radar displays at the Eurocontrol operated Maastricht Upper Area Control Centre (MUAC). Flight dynamics conform to BADA flight performance models. SectorX was run on a Windows laptop connected to an external 28” display with a resolution of 1920 x 1080. Participants interacted via mouse and keyboard and could solve conflicts using heading and altitude via a clearance menu (see Figure 2).

For the training pre-test, six reference scenarios were created with a maximum sector occupancy of 22 aircraft, each 2.5 minutes long running at twice the real-time speed (representing five minutes real time movements). We decided to use short traffic vignettes as a trade-off between maximizing the input data to the conformance ML models (i.e., number of scenarios and conflict resolutions generated) and retaining a realistic simulation exercise. The approach was found acceptable by fourteen ATM experts and ATCOs participating in a workshop prior to the experiments. Scenarios were based on a 100 x 100 nm generic sector in an approximately octagonal shape (see Figure 3) that allowed creation of unrecognisable scenario variants (via rotation). Altitudes were within reduced vertical separation minimum (RVSM) airspace, FL290 to FL410. All scenarios were scripted to present a single same-altitude two-aircraft closing conflict with a CPA of 0 nm to reduce solution bias. For the main experiment, two of the six reference scenarios were used: Scenario A (Scen.A) had a 68 angle conflict between two aircraft and Scenario B (Scen.B) a 134 angle conflict between two aircraft. The scenarios are shown in Figure 3. While identical to the ones used in the training pre-test in terms of airspace, traffic, and conflicts, the scenarios

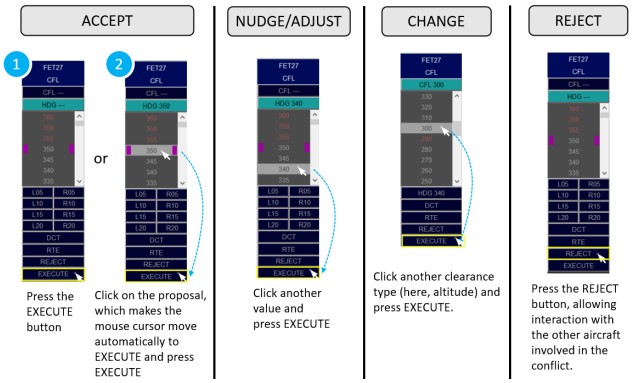
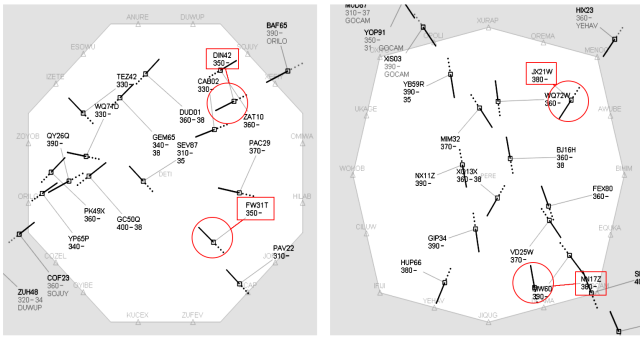


Figure 2. Advisory interactions in SectorX.



(a) (b)

Figure 3. Scenarios in main experiment with conflict aircraft annotated in red: a) Scen.A (left), and b) Scen.B (right)

were adapted during the interim ML training phase to incorporated automated CD&R solutions and support supervisory control procedures. Because every scenario and conflict can be considered unique, we chose to use two different scenarios. The same scenario was used for all experimental conditions (i.e. reflecting the two independent variables with three levels) to avoid scenario and conflict design as confounding variables. Availability of ATCOs (i.e., three hours) restricted the use of more scenarios.

E. Procedures

Identical procedures were used in SIM1 and SIM2. In the training-pre test, ATCOs solved 36 conflicts (six scenarios repeated six times) in manual control. Air–ground communication was carried out via datalink with no radiotelephony (RT) was required. Participants were instructed to *ensure separation between aircraft and make sure aircraft leave the sector through their assigned exit waypoint at the correct flight level as indicated by their flight plan. Each task is equally important*. Before the test started, consent forms and a demographics questionnaire was administered. Participation lasted for three hours. The training pre-test data was used to train the group ML model and synthetically derive the personal models. The duration between the pre-test and main experiment was six weeks for SIM1 and four weeks for SIM2. For the main experiment, the same ATCOs interacted with ML solved analogues of Scen.A and B. Their task was to supervise

an automated agent managing the routine ATC tasks including assuming flights, handover flights, issue climb and descent clearances, and direct to clearances when conflict free. The agent detected conflicts and provided resolution advisories to which the ATCO had to respond. ATCOs played eighteen runs where Scen.A and B was repeated nine times each to achieve a complete variation of the conformance and transparency variables. Figure 4 shows the simulator environment for the main experiment. Run order varied according to a Latin Square to mitigate effects of learning, boredom, and complacency.

In each run, a heading advisory was issued with respect to a conflict pair. The ATCO responded by either accepting with or without revising the advisory by clicking "EXECUTE" (directly, or followed by a nudge, adjustment, or change), or by clicking "REJECT" (only SIM2, see Figure 2). At this point the simulator paused and the ATCO was prompted to provide ratings for advisory agreement, conformance, and understanding. The simulation resumed after all ratings were collected. At the end of the run, the ATCO provided a workload rating.

Figure 4. Simulator setup showing participant responding to an advisory at the highest transparency level (T2) in SectorX. Animations of a typical experiment session can be seen here: http://mahaloproject.eu/?page_id=133.

F. Data analysis

Data from SIM1 and SIM2 were analysed separately due to differences in the generation of ML advisories (SL and RL models) and SectorX interface (i.e., acceptance interaction). Statistics were calculated in SPSS v.28. Repeated measures ANOVAs were applied for analysing continuous data, using pairwise post-hoc tests with Bonferroni adjustments for multiple comparisons. Ordinal data was analysed using non-parametric Friedman's ANOVAs, using Wilcoxon signed-rank post-hoc tests with a Bonferroni corrected significance level of $p < 0.006$. Mann Whitney U tests were used for comparing participant groups. Given inter-respondent variability in agreement ratings, workload ratings, and response time, data were normalized into within-participant calculated Z scores.

IV. Results

Results are divided in two parts. First, the main and interaction effects for SIM1 (n = 18) and SIM2 (n = 14) are presented. Second, comparisons between groups of ATCOs' responses to optimal advisories are presented. Table II details the groups created. Because the SL and RL models were developed in isolation (independent variables), for some ATCOs and scenarios the different advisories were found to be similar (e.g., an ATCO's personal advisory was similar to the optimized advisory or group advisory. Moreover, it was observed that ATCOs tended to form a bimodal distribution in terms of their average separation margin during the training pre-test. To explore this, ATCOs were divided in groups using a binary split based on the deviation of their average separation distance from the CPA aimed for by the optimal advisory.

TABLE II. Separation distance deviation (NM) from optimal advisory CPA for ATCO groups.

	SIM1 (n = 9 / group)		SIM2 (n = 7 / group)	
	Scen.A	Scen.B	Scen.A	Scen.B
Group / Optimal CPA	6.6nm	7.7nm	10.7nm	10.5nm
Closer to optimal	$< 2.3nm$	$< 1.0nm$	$< 2.5nm$	$< 2.7nm$
Farther from optimal	$> 2.5nm$	$> 1.1nm$	$> 2.5nm$	$> 2.9nm$

A. Effects of conformance and transparency

Statistical results from the repeated measures ANOVAs are shown in Table III, and from the non-parametric Friedman's ANOVAs in Table IV. Only conditions and measures where significant findings were found are shown.

TABLE III. Significant main and interaction effects in SIM1 and SIM2 scenarios. Repeated measures ANOVA with $p < .05$.

	SIM1		SIM2	
Measure	Scen.A	Scen.B	Scen.A	Scen.B
Conformance effects				
Agreement	p = .003		$p < .001$	$p < .001$
Workload		p = .044		
Delta CPA	p = .007		p = .046	$p < .001$
Response time			p = .005	
Interaction effects between transparency and conformance				
Agreement		p = .011		

TABLE IV. Significant effects across conformance and transparency conditions in SIM1 and SIM2 scenarios. Friedman's ANOVA with $p < .05$.

	SIM1		SIM2	
Measure	Scen.A	Scen.B	Scen.A	Scen.B
Acceptance	p = .001		$p < .001$	p = .006
Advisory conformance	$p < .001$		$p < .001$	p = .001
Advisory understanding			$p < .001$	p = .002

1) Acceptance: A significant effect of conformance and transparency on participants' acceptance response (see Table IV) as found in SIM1 Scen.A ($X^2(8) = 25.49$, $p = .001$), SIM2 Scen.A ($X^2(8) = 30.17$, $p < .001$), and SIM2 Scen.B ($X^2(8) = 21.67$, $p = .006$). In SIM1 Scen.A, optimal advisories were generally accepted as given compared with group and personal advisories that generally were nudged (i.e., slightly adjusting the heading before accepting the advisory, see Figure 2). While post-hoc tests were not significant, trends were observed favouring optimal advisories over group advisories (Z = -2.74, p = .006) in the low transparency (T0) condition. Similarly, optimal advisories were favoured over group advisories (Z = -2.65, p = .008) and personal advisories (Z = -2.71, p = .007) in the medium transparency (T1) condition. In SIM2 Scen.A, personal advisories were generally accepted as given compared with optimal advisories that often were nudged, and group advisories that often were nudged or adjusted. Post-hoc tests revealed a significant difference in the low transparency (T0) condition, where optimal advisories were favoured over group advisories (Z = -2.84, p = .004). In SIM2 Scen.B, personal and group advisories were generally accepted as given compared with optimal advisories that often were nudged or adjusted. The post-hoc analysis showed a significant difference in acceptance response for the high transparency (T2) condition, where group advisories were favoured over optimal advisories (Z = -2.79, p = .005).

2) Agreement ratings: Significant effects of conformance on agreement ratings were measured in SIM1 Scen.A ($F(2, 34) = 8.68$, p = .003), SIM2 Scen.A ($F(1.82, 23.62) = 14.50$, $p < .001$), and SIM2 Scen.B ($F(2, 26) = 20.02$, $p < .001$). For SIM1 Scen.A, post-hoc analysis (pairwise comparisons) showed that optimal advisories had significantly higher agreement than personal (p = .028) and group advisories (p = .021). For Scen.B, a significant interaction effect was found between conformance and transparency ($F(4, 68) = 3.57$, p = .011). A subsequent contrasts analysis revealed that agreement increased for group advisories as the level of transparency increased from T0 to T2, while they decreased for personal advisories ($F(1,17) = 6.45$, p = .021) and optimal advisories ($F(1,17) = 10.96$, p = .004). In SIM2, Scen.A, a significant effect was observed for conformance level on agreement ratings ($F(1.82, 23.62) = 14.50$, $p < .001$). Pairwise comparisons showed that personal (p = .003) and optimal (p = .002) advisories received significantly higher agreement ratings than group advisories. In Scen.B, a contrasting significant effect of conformance level on agreement was found ($F(2, 26) = 20.02$, $p < .001$). Personal ($p < .001$) and group ($p < .001$) advisories received significantly higher agreement ratings compared to optimal advisories.

3) Workload ratings: For workload, the only difference was observed for the conformance variable in SIM1 Scen.B ($F(2,34) = 3.43$, p = .044). Pairwise comparisons showed that workload increased as conformance decreased, with the largest difference between personal and optimal advisories, although not significant (p = .072).

4) Delta CPA: Model conformance was found to significantly affect Delta CPA distance in SIM1 Scen.A ($F(2, 34) = 5.82$, p = .007), SIM2 Scen.A ($F(2,26) = 3.46$, p = .046), and SIM2 Scen.B ($F(2,26) = 14.63$, $p < .001$). Pairwise comparisons in SIM1 Scen.A showed that optimal advisories had significantly lower delta CPA distances than personal advisories (p = .014) and group advisories (p = .037). To illustrate the differences, participants' deviations were smaller for optimal advisories (M = 0.54 nm) compared to personal advisories (M = 1.62 nm) and group advisories (M = 1.52 nm). In SIM2 Scen.A, pairwise comparisons were not significant, but the conformance graph revealed that delta CPA distances increased from personal (M = .70 nm), to group (M = 1.45 nm), and optimal advisories (M = 2.02 nm). In SIM2 Scen.B, pairwise comparisons showed that optimal advisories had significantly higher delta CPA distances compared to personal advisories (p = .006) and group advisories (p = .002).

5) Response time: A significant effect of conformance on response time was measured in SIM2 Scen.A ($F(2,26) = 6.43$, p = .005). Post-hoc pairwise comparisons showed that participants reacted significantly faster to personal advisories compared to group advisories (p = .003).

6) Advisory conformance ratings: Significant differences across conditions in ratings of the advisory's similarity to the ATCO's preferred solution was observed in SIM1 Scen.A ($X^2(8) = 26.63$, $p < .001$), SIM2 Scen.A ($X^2(8) = 45.96$, $p < .001$), and in SIM2 Scen.B ($X^2(8) = 25.66$, p = .001). In SIM1 Scen A, optimal advisories were consistently rated higher than both group and personal advisories. Post-hoc tests revealed a significant difference in the low transparency (T0) condition with personal advisories being rated less similar to own preference than optimal advisories (Z = -2.81, p = 0.005). In SIM2 Scen.A, personal and optimal advisories were consistently rated higher than group advisories. Post-hoc tests found a significant differences in the low transparency (T0) condition with personal advisories being rated more similar to own preference than group advisories (Z = -2.82, p = 0.005). A significant difference was also noted in the medium transparency (T1) condition with personal advisories being rated higher than group advisories (Z = -2.95, p = 0.003). In SIM2 Scen.B, personal and group advisories were consistently rated higher than optimal advisories but not reaching significance.

7) Advisory understanding ratings: Significant effects of experimental condition were found in SIM2 Scen.A ($X^2(8) = 36.72$, $p < .001$) and SIM2 Scen.B ($X^2(8) = 24.53$, p = .002). Post-hoc tests did not reach significance in either Scen.A or Scen.B. However, in Scen.A the noticeable larger IQR for group advisories (*IQR* for T0 = 3-6, T1 = 3-5.3, T2 = 3-5.3), compared to personal (*IQR* for T0 = 5-6, T1 = 6-6, T2 = 4.8-6) and optimal advisories (*IQR* for T0 = 5-6, T1 = 3.8-6, T2 = 5-6), indicates that group advisories were less understandable. For Scen.B, a similar pattern was found between optimal (*IQR* for T0 = 3.8-6, T1 = 4-6, T2 = 3-5.3), personal (*IQR* for T0 = 5-6, T1 = 5-6, T2 = 5-6), and group advisories (*IQR* for T0 = 5-6, T1 = 5-6, T2 = 5-6), indicating that optimal advisories were less understandable.

B. Group analysis of conformance and transparency effects

Group effects were only analysed for optimal advisories. Table V shows the significant results from the statistical analysis comparing the groups in II.

TABLE V. SIGNIFICANT DIFFERENCES BETWEEN ATCO GROUPS DEPENDING ON THEIR SEPARATION DISTANCE PREFERENCES IN RELATION TO THE TARGET CPA IN THE OPTIMAL ADVISORY IN SIM1 AND SIM2 SCENARIOS. MANN WHITNEY U TESTS WITH $p < .05$. THE SUBSCRIPT SHOWS THE TRANSPARENCY CONDITION.

	SIM1		SIM2	
Measure	Scen.A	Scen.B	Scen.A	Scen.B
Agreement	$= .031_{T0}$	$= .031_{T2}$		
Workload			$= .004_{T1}$	
Delta CPA		$< .001_{T2}$		$= .038_{T2}$
Response time			$= .026_{T0}$	
Acceptance		$= .014_{T2}$		
Advisory confor.		$= .050_{T2}$		$= .017_{T0}$
Advisory underst.	$= .031_{T2}$			$= .038_{T0}$

1) Acceptance: A significant difference between groups on advisory acceptance of optimal advisories was found in SIM1 Scen.B for the high transparency (T2) condition (U = 13.00, z = -2.73, p = .014). The group whose preferences were closer to the optimal advisory CPA had a higher acceptance rank than the group farther away. Differences in the same direction was found between groups in the T0 and T1 conditions, but not reaching significance. In SIM2 Scen.B a trend was observed for the medium transparency (T1) condition (U = 39.5, z = 1.99, p = .053). The "closer" group had a higher acceptance rank than the group "farther" from the optimal advisory. Differences in the same direction was found between groups in the T0 and T2 conditions but not reaching significance.

2) Agreement ratings: Groups differed significantly in their agreement ratings of optimal advisories in SIM1, Scen.A for the low transparency (T0) condition (U = 65.00, z = 2.16, p = .031). The "farther" from group had a higher agreement ratings than the group closer to the optimal advisory CPA. In the high transparency condition (T2), the relationship between groups had changed so that the "closer" group had higher agreement ratings compared to the "farther" from group, although not significant (U = 21.00, z = -1.72,p = .094). A significant difference was also found in SIM2 Scen.B for the high transparency (T2) condition (U = 16.00, z = -2.16, p = .031). The group closer to the optimal advisory CPA had higher agreement ratings than the group farther away.

3) Workload ratings: In SIM2 Scen.A, a significant difference was found between groups on workload ratings in the medium transparency (T1) condition (U = 3.00, z = -2.75, p = .004). Workload ratings were lower for the group closer to the optimal advisory CPA compared to the group farther away.

4) Delta CPA: In SIM1 Scen.B, groups differed significantly on delta CPA distance in the high transparency (T2) condition (U = 76.50, z = 3.50, $p < .001$). The group closer to the optimal advisory CPA had smaller delta CPA values compared to the other group. The "closer" group had a *Mdn* 0 nm deviation, in contrast to a *Mdn* 2.57 nm deviation for the "farther" group. Differences between groups were in the same direction for the T0 and T1 conditions but not reaching significance. A significant difference was also found in SIM2 Scen.B for the T2 condition (U = 8.5, z = -2.06, p = .038). The "closer" group had smaller Delta CPA values compared to the "farther" group. While both groups acted to reduced separation, the "farther" group made bigger changes (*Mdn* = 3.24 nm) compared to the "closer" group (*Mdn* = 1.05 nm). Similar patterns were observed for the T0 and T1 conditions, although not reaching significance.

5) Response time: In SIM2 Scen.A, groups differed significantly in the low transparency (T0) condition (U = 7.00, z = -2.24, p = .026), with the "closer" group responding faster than the "farther" group.

6) Advisory conformance ratings: A significant difference was found between groups on advisory conformance ratings in SIM1 Scen.B for the high transparency (T2) condition (U = 18.50, z = -2.08, p = .050). The group closer to the optimal advisory CPA had rated advisories as more similar to their own solution compared to other group. Similar patterns were observed for the T0 and T1 conditions, but not reaching significance. In SIM2 Scen.B, groups differed significantly for the low transparency (T0) condition (U = 43.00, z = 2.44, p = .017). The "closer" group had higher ratings compared to the "farther" group. Similar patterns were observed for conformance ratings in the T1 and T2 conditions, although not reaching statistical significance.

7) Advisory understanding ratings: In SIM1 Scen.A, groups differed significantly on their understanding of optimal advisories in the high transparency (T2) condition (U = 16.50, z = -2.43, p = .031). The group closer to the optimal advisory CPA rated their understanding of the advisory higher compared to other group. In SIM2 Scen.B, a significant difference was found in the low transparency (T0) condition (U = 41.00, z = 2.18, p = .038). The "closer" group had higher ratings than the "farther" group. A similar difference were observed in the T1 condition, although not significant.

V. DISCUSSION

A. Conformance effects

The effects of ML model conformance were not consistent across simulations and scenarios. Personal advisories generated higher ratings of agreement and conformance and lower delta CPA distances in SIM2 Scen.A and B. In SIM1 Scen.A, however, optimal advisories received higher ratings of agreement and conformance and lower delta CPA distances. In SIM2 Scen.A, ATCOs rated group advisories as less agreeable, conformal, and understandable. In contrast, responses to group advisories were more positive in SIM2 Scen.B. Here, optimal advisories received the lowest ratings for agreement, conformance, and understanding, while also resulting in higher delta CPA distances. Why are we seeing these different effects of conformance condition on dependent measures?

A possible explanation can be found in how the conformance models were defined. The personal models were the least stable given the limited sample size (six resolution

samples) for their definition. While a few ATCOs in each Simulation were found highly consistent in the training pre-test (e.g., turning the same aircraft to the right behind the other in all repetitions, achieving a similar CPA), the majority showed variations in choice of aircraft, turn direction and achieved CPA. Moreover, the optimal models were qualitatively very different in both simulations (see Table II. In SIM1, the optimal advisory, compared to most personal models, had a tighter target CPA (6-7 nm), smaller heading value (around 15 degrees), and was provided earlier (20 seconds into scenario). In SIM2, however, the optimal advisory was provided later than most personal models (96-114 seconds into scenario) and with a larger CPA (around 10.5 nm). As a reference, the group model generated advisories that in both simulations and scenarios better reflected the ”average” personal model.

B. Transparency effects

It was hypothesised that ATCOs' reactions and perceptions of advisories would be higher if advisories were presented in higher transparency display formats. While no significant main effects of transparency were found, the data points in the opposite direction - that increased transparency negatively affected ATCOs' responses to advisories. The transparency variables provided qualitatively different information about the CPA that the model was aiming for. With more detailed information about the target CPA, it becomes easier for ATCOs to compare the proposed CPA with their separation margin preference. If the difference is large it is more likely that the advisory is interfered with (e.g., by adjusting the heading to reduce or increase the separation distance to get closer to the preferred CPA). Transparency and explainability should increase acceptance and agreement for an optimal algorithm, but it should also decrease acceptance and agreement for a sub optimal algorithm. But because what is ”optimal” partly depends on the user's preferences, this general objective of transparency is misleading. The more suitable transparency objective is to increase the user's understanding of the automation, to support an informed decision on whether the system's solution or behavior is suitable for the given situation. Importantly, and increased understanding does not not necessarily increase agreement and acceptance.

C. The influence of separation distance preferences

Advisories target CPA between conflicting aircraft was found to be an important factor reflecting ATCO's individual resolution preferences. Results showed that ATCOs' response to optimal advisories depended on the advisory's conformance with personal CPA preferences. The closer the match, the more positive was the response to the advisory. In the group analysis, a reoccurring pattern emerged where the ”closer” group (to optimal CPA) showed unchanged or more positive responses to optimal advisories with increasing transparency. That is, their acceptance and ratings for agreement, conformance, and understanding of advisories was higher compared with the other group. Also, their delta CPA tended to be smaller. The ”farther” group generally had a less positive response as transparency increased. Here, increased transparency tended to result in more changes to the advisory (i.e., reduced acceptance response), lower ratings of agreement, conformance, and understanding, and larger Delta CPA values (an exception was SIM2 Scen.B).

The traditional system design approach is to design a system that all operators have to conform to. This is sensible for many standard operating procedures and tasks or decision that easily can be optimized. This makes less sense, however, for time and safety critical situations where the definition of what is an optimal solution is subjective and performance depends more on the operator's capacity to cope with the tasks (and avoid peaks of workload and stress). In environments such as tactical CD&R, we want to avoid the human and automation/AI disagreeing (or ”arguing”) on how to solve a particular problem.

D. Personal machine learning models

The study revealed ATCO variability where differences both within and between ATCOs presented challenges to training the SL conformance models. Regardless of ML approach, the robustness of a personal model is highly dependent on the internal consistency of that ATCO in choosing solutions (i.e., the extent to which a given ATCO solves the same conflict the same way every time). For example, low within-ATCO consistency reduces the robustness of a personal model. Similarly, the robustness of a group model is highly dependent on the extent to which ATCO's agree on solutions. If between-ATCO consistency is low, the group model becomes less representative of the group. Further, both SL and RL models add the additional challenge that there is still a great deal of artistry required in designing/configuring the model, e.g., in terms of defining labels or the reward structure. It can not be expected that all decision making tasks are suitable for personalized applications. For instance, for problems where optimal objectives are unambiguous and where subjective preferences are less relevant, there is little incentive to personalize. The same applies to tasks or problems where there is consensus between operators on how to act.

E. Future research

First, an avenue to explore is the potential utility of personalisation, or tuneable parameters that might allow for a hybrid of the optimal and personal model view. Such that ATCOs could tune certain parameters within the confines of an advisory system that strives to optimize performance. This study suggests that separation margin appears to be the most prominent tuneable parameter to explore. Second, research on ML approaches is needed for determining how to best derive robust personal models given accessibility limitations to large amounts of individual data. An associated challenge regards how to train these models, what data to consider, and how to preprocess the data (e.g., label or feature definition, reward functions etc) to best capture end users' preferences.

Finally, future research is required to explore and better understand the impact of system transparency on acceptance

and system trust. In contrast to expectations, increased transparency had in some cases an inverse effect with acceptance and agreement reaching higher values in the lowest transparency condition. The probability for an explanation to achieve its goal depends to a large extent on what the user is trying to understand. Therefore, research is welcomed on personalized transparency functions that allows the system to determine what the operator is trying to understand and what the operator's preferences are. As a next step, systems should be able explain why a proposed action/solution is better than that preferred by the operator (if different).

VI. Conclusion

This study has demonstrated that personalization helps in making an AI-based advisory system more acceptable and easier to work with (e.g., reduce interventions). Personalization can therefore be seen as a way to overcome some of the challenges associated with integrating ML/AI techniques in ATC from the perspective of human-machine collaboration. Furthermore, a person's response to a resolution advisory partly depends on how close that advisory is to the person's solution preferences. Systems that are more personalized may also lessen the need for system transparency. Future ATC systems should acknowledge and embrace in the design that controllers differ in their conflict resolution preferences and that ATCOs can be grouped according to similarities in their decision making strategies.

Acknowledgment

The authors would like to thank all participating ATCOs in Italy and Sweden for their invaluable support and engagement in the study.

References

[1] J. Y. C. Chen and M. J. Barnes, "Human-agent teaming for multirobot control: A review of human factors issues," *IEEE Trans. Human–Mach. Syst.*, vol. 44, no. 1, pp. 13–29, 2014.

[2] J. B. Lyons, K. Sycara, M. Lewis, and C. A, "Human–autonomy teaming: Definitions, debates, and directions," *Front. Psychol.*, vol. 12, no. 589585, 2021.

[3] M. R. Endsley, "From here to autonomy: Lessons learned from human–automation research," *Hum. Factors*, vol. 59, no. 1, pp. 5–27, 2017.

[4] C. Westin, C. Borst, and B. Hilburn, "Strategic conformance: Overcoming acceptance issues of decision aiding automation?" *IEEE Trans. Human–Mach. Syst.*, vol. 46, no. 1, pp. 41–52, 2016.

[5] C. Westin, B. Hilburn, C. Borst, E.-J. Van Kampen, and M. Bång, "Building transparent and personalized ai support in air traffic control," in *39th DASC*, San Antonio, TX (Virtual), Oct. 11-16 2020.

[6] Y. Liu, Y. Lee, and A. N. K. Chen, "Evaluating the effects of task-individual-technology fit in multi-DSS models context: A two-phase view," *Decis. Support Syst.*, vol. 51, no. 3, pp. 688–700, 2011.

[7] R. Parasuraman and Y. Jiang, "Individual differences in cognition, affect, and performance: Behavioral, neuroimaging, and molecular genetic approaches," *Neuroimage*, vol. 59, no. 1, pp. 70–82, 2012.

[8] J. L. Szalma, "Individual differences in human–technology interaction: Incorporating variation in human characteristics into human factors and ergonomics research and design," *Theor. Issues Ergon. Sci.*, vol. 10, no. 5, pp. 381–397, 2009.

[9] B. Hilburn, C. Westin, and C. Borst, "Will controllers accept a machine that thinks like they think? The role of strategic conformance in decision aiding automation," *ATC Quart.*, vol. 22, no. 2, pp. 115–136, 2014.

[10] R. Regtuit, C. Borst, E.-J. Van Kampen, and M. M. Van Paassen, "Building Strategic Conformal Automation for Air Traffic Control Using Machine Learning," in *AIAA SciTech*, Kissimmee, FL, Jan. 8-12 2018.

[11] P. N. Tran, D.-T. Pham, S. K. Goh, S. Alam, and V. Duong, "An interactive conflict solver for learning air traffic conflict resolutions," *J. Aerosp. Inf. Syst.*, vol. 17, no. 6, pp. 271–277, 2020.

[12] S. Van Rooijen, J. Ellerbroek, C. Borst, and E.-J. Van Kampen, "Toward individual-sensitive automation for air traffic control using convolutional neural networks," *J. Air Transp.*, vol. 28, no. 3, pp. 1–9, 2020.

[13] ——, "Conformal automation for air traffic control using convolutional neural networks," in *13th USA/Europe ATM R&D Seminar*, Vienna, Austria, Jun. 17-21 2019.

[14] C. Westin, C. Borst, and B. Hilburn, "Automation transparency and personalized decision support: Air traffic controller interaction with a resolution advisory system," in *IFAC*, vol. 49, no. 19, Kyoto, Japan, 2016, pp. 201–206.

[15] J. Y. C. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. J. Barnes, "Situation awareness-based agent transparency," Army Res. Lab., Aberdeen Proving Grounds, MD, ARL-TR-6905, Apr. 2014.

[16] A. R. Selkowitz, S. G. Lakhmani, C. N. Larios, and J. Y. C. Chen, "Agent transparency and the autonomous squad member," *Hum. Factors*, vol. 60, no. 1, pp. 1319–1323, 2016.

[17] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Human Comp. Stud.*, vol. 44, no. 58, pp. 697–718, 2003.

[18] T. Helldin, G. Falkman, M. Riveiro, A. Dahlbom, and M. Lebram, "Transparency of military threat evaluation through visualizing uncertainty and system rationale," in *Proc. 10th EPCE*, vol. 8020 LNAI, Las Vegas, NV, Jul. 22-26 2013, pp. 263–272.

[19] J. Launchbury, "A DARPA perspective on artificial intelligence," 2017. [Online]. Available: https://www.darpa.mil/attachments/AIFull.pdf

[20] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. Natl. Acad. Sci.*, vol. 116, no. 44, p. 22071–22080, Oct 2019.

[21] C. Seifert, A. Aamir, A. Balagopalan, D. Jain, A. Sharma, S. Grottel, and S. Gumhold, *Visualizations of Deep Neural Networks in Computer Vision: A Survey*. Springer, Cham: Studies in Big Data, 2017, vol. 32, pp. 123–144.

[22] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 2015.

[23] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. D. Jackel, U. Muller, and K. Zieba, "Visualbackprop: visualizing cnns for autonomous driving," *CoRR*, vol. abs/1611.05418, 2016. [Online]. Available: http://arxiv.org/abs/1611.05418

[24] A. Bhaskara, M. Skinner, and S. Loft, "Agent transparency: A review of current theory and evidence," *IEEE Trans. Hum.-Machine Syst,*, vol. 50, no. 3, pp. 215–224, 2020.

[25] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human–autonomy teaming: A review and analysis of the empirical literature," *Hum. Factors*, vol. 64, no. 5, pp. 904–938, 2022.

[26] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *proc. CHI*, Montréal, Canada, Apr. 21-26 2018.

[27] D.-T. Pham, N. P. Tran, S. Alam, V. Duong, and D. Delahaye, "A machine learning approach for conflict resolution in dense traffic scenarios with uncertainties," in *13th USA/Europe ATM R&D Seminar*, Vienna, Austria, 2019.

[28] M. W. Brittain and P. Wei, "One to any: Distributed conflict resolution with deep multi-agent reinforcement learning and long short-term memory," in *AIAA Scitech*, Virtual, Jan. 11–15, 19–21 2021.

[29] J. Mollinga and H. H. V, "An autonomous free airspace en-route controller using deep reinforcement learning techniques," *ArXiv*, vol. abs/2007.01599, 2020.

[30] M. Liang, W. Li, D. Delahaye, and P. Notry, "Policy optimization in automated point merge trajectory planning: An artificial intelligence-based approach," in *38th DASC*, San Diego, CA, Sep. 08-12 2019.

[31] J. Xu-Rui, W. Ming-Gong, W. Xiang-Xi, and W. Ze-Kun, "Application of ensemble learning algorithm in aircraft probabilistic conflict detection of free flight," in *ICAIBD*, 2018, pp. 10–14.

[32] Z. Wang, M. Liang, and D. Delahaye, "Data-driven conflict detection enhancement in 3d airspace with machine learning," in *AIDA-AT*, 2020.

[33] M. Eby, "A self-organizational approach for resolving air traffic conflicts," *Lincoln Lab J.*, vol. 7, no. 2, pp. 239–254, 1994.