Presented at the 39th AIAA/IEEE DASC conference, Virtual event October 11-16, 2020

Building Transparent and Personalized AI Support in Air Traffic Control

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

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

Clark Borst
Control and Simulation
Technical University Delft
Delft, The Netherlands
c.borst@tudelft.nl

Erik-Jan van Kampen
Control and Simulation
Technical University Delft
Delft, The Netherlands
e.vanKampen@tudelft.nl

Magnus Bång
Computer and Information Science
Linköping University
Linköping, Sweden
magnus.bang@liu.se

***Abstract*— Artificial intelligence is considered a key enabler for realizing a more efficient future air traffic management system. As the automation designed to support us grows more sophisticated and complex, our understanding of it tends to suffer. Recent research has addressed this issue in two ways: either through increased automation transparency or increased personalization. This paper overviews recent work in these two areas of strategic conformance (i.e., personalization) and automation transparency (e.g., explainable artificial intelligence and machine learning interpretability). We discuss how to achieve and how to balance conformance and transparency in the context of a machine learning system for conflict detection and resolution in air traffic control. In the MAHALO project, we aim to build, and empirically evaluate, a personalized and transparent decision support system by combining supervised and reinforcement learning approaches. We believe that such a system could strive for optimal performance while accommodating individual differences. By knowing the individual's preferences, the system would be able to afford transparency by explaining both why it suggests another solution (that deviates from the individual's), and why this solution is considered to be better.**

***Keywords*— AI; ATM; machine Learning; personalized; strategic conformance; transparency**

I. Introduction

Despite the recent downturn and turmoil in the economy and airline industry, traffic forecasts retain expectations of an increase in traffic in the medium and long term. For continued development and growth, the Air Traffic Management (ATM) industry depends on new technology and increasingly sophisticated automation [1]. Of particular interest are the possibilities and promises of artificial intelligence (AI) [2]. Machine Learning (ML) technologies offer enhanced performance, efficiency, and utilization of human resources through a system design that can sense, learn, and act autonomously. AI solutions for ATM have also started to emerge, such as using ML and big data techniques to improve the accuracy in trajectory prediction [3] [4], to identify novel route patterns and predict airlines' route selection [5], and for speech recognition in air traffic controller (ATCo)-pilot communication [6]. Moreover, ML solutions appear particularly suitable for the task of conflict detection and resolution (CD&R), assuming much of the ATCo's cognitive work involved in overseeing, separating, and expediting air traffic. For conflict resolution, Pham et al. [7] recently presented an AI agent able to resolve over 81% of all conflicts in high uncertainty and high traffic scenarios.

As automation grows increasingly capable of performing the deeper, "thinking" parts of many jobs, it is essential to consider how to facilitate operators' trust and acceptance of the automation. Unlike traditional AI methods, ML does not rely on explicitly programmed algorithms. Rather than the old 'if then' rules of traditional expert systems, ML rests on models that can self-organize, learn, and improve performance (e.g. classification) over time. ML approaches generally have the implicit benefit of graceful degradation (unlike traditional expert systems that might crash spectacularly). The flip side of such graceful degradation, however, is that the output of ML algorithms can be unintuitive and difficult to interpret. Ironically, ML methods (e.g., 'deep learning' neural network extensions) that show the best learning performance, tend to be 'opaque' and the most challenging to understand [8]. Several human factors constructs have explored the breakdown in human understanding of automation, including automation surprise [9] [10], out-of-the-loop [11], and loss of situation awareness [12].

Research has addressed understanding issues of automation in two different ways: either through increased transparency or increased personalization. For the human operator to retain authority, the system must be able to afford transparency and explain its reasoning and behavior. An alternative approach may be to develop automation that conforms to humans, and even the individual's, problem-solving strategies and preferences. This paper asks two simple but profound questions: in the emerging age of ML, 1) to what extent should we design automation to match individual human behavior (i.e., strategic conformal), and 2) to what extent should the automation be made transparent? Answering these questions is important for guiding the design of future ML systems in ATM and other safety-critical domains.

In relation to ongoing work in the European Horizon 2020 [13] funded *Modern ATM via Human/Automation Learning Optimization* (MAHALO) project, we discuss how to achieve and how to balance conformance and transparency in the context of a ML system for ATM CD&R. This paper provides an overview of recent work in the areas of strategic conformance

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 author's view only.

and automation transparency. A theoretical framework is presented for discussing the balance of conformance and transparency of automation. Finally, the hybrid ML approach explored in the MAHALO is introduced, where the objective is to not only demonstrate a ML CD&R capability but to create an empirically derived framework and guidelines for how to develop advanced future AI, specifically for ATM.

II. Theoretical Constructs

Human understanding of automation grows more challenging as automation becomes increasingly complex. In attempts to support understanding of automation, research has explored two seemingly opposing constructs of *automation transparency* and *strategic conformance*. Fig. 1 illustrates the relationship between the three constructs of conformance, transparency, and understanding.

- Conformance—the *apparent strategy match* between human and machine solutions. This similarity is external, overt, and observable, and is the extent to which cause and effect can be observed.
- Transparency— the extent to which aspects of the *automation's* inner process underlying a solution can be explained in human terms.
- Understanding – the extent to which the *human* understands the automation's reasoning underlying the solution. To make the system understandable to humans, we need to align its explanations to aspects that are of relevance to human decision-making.

A. Automation Transparency

The notion of automation transparency has been researched broadly across many domains, including human-computer interaction, human factors, and numerous AI subdomains. Differences in the type of automation explored and requirements for understanding that automation has given rise to different terms across domains. Examples from the human factors and cognitive engineering domains include automation [14] and agent transparency [15], automation visibility [16], understandability [17], observability [18], and comprehensibility [19]. In the AI community, the notion of transparency is captured in terms of ML interpretability [20], Explainable AI (XAI) [8], and intelligibility of context-aware systems [21]. These concepts all strive to make aspects of the inner workings of the automation 'black box' understandable to the human. Transparency has also been advocated as a requirement of automation by politicians and legislators, such as the EU general data protection regulation that contains a controversial 'right to an explanation' criterion [22].

Automation transparency has been defined as: "*the ability of the automation to afford understanding and predictions about its behavior*" ([14] p. 202). In this regard, automation transparency is a property of the automation. The human reaction to this property can be assessed in terms of understanding and predicting the automation's behavior. Transparency is a multifaceted construct that ultimately is shaped by what is sought to be understood: what the human is trying to understand governs what needs to be explained. Consequently, transparency has been considered for many different reasons, such as to explain abnormal automation behavior [23], why the automation might err [24], the behavior of intelligent agents and autonomous robots [25] [26], as an indication of the automation's reliability [27], and the automation's proximity to its performance envelope [28].

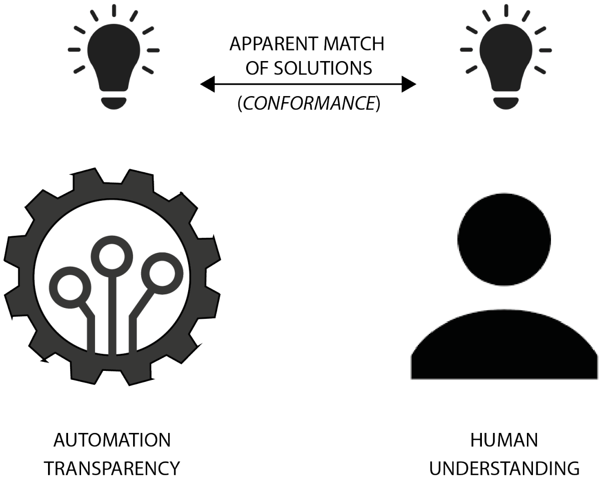

Fig. 1. Relationship between conformance, transparency, and understanding.

Two models of agent transparency have been proposed for autonomous systems; Lyon's [29] model for human-robot teaming approach transparency from the requirement of establishing a shared intent and awareness between robot/autonomous systems and humans, and Chen et al.'s [15] situation awareness-based agent transparency (SAT) model for human-agent teaming. The SAT model specifies three levels of transparency, paralleling the SA levels of perception, comprehension, and predictions. In a literature review on agent transparency, Bhaskara et al. [30], investigated the effects of transparency on performance, response time, subjective workload, situation awareness, trust, and usability. Five studies were reviewed, all of which explored military applications and most of which operationalized transparency on the SAT model. Bhaskara et al. distinguished between four levels of transparency: low (supporting perception in terms of basic information or advice), medium (supporting comprehension in terms of adding information about the agent's reasoning), high (supporting prediction in terms of adding information on either expected outcome/consequence or added reasoning for a recommendation), and very high (supported prediction by adding more information in the high transparency level (e.g., both an outcome prediction and uncertainty information). While findings show some benefits of increased (levels of) agent transparency to acceptance, trust, situation awareness, workload, and response time, Bhaskara et al. note that results so far are inconclusive and that there are simply not enough results to form stable conclusions [30].

Researchers have explored transparency of intelligent agents (e.g., robots) by providing explanations based on the agent's decision-making processes. Frameworks such as the Belief-Desire-Intention (BDI) [31], Partially Observable Markov Decision Process (POMDP) [32], and Parallel-rooted, Ordered, Slip-stack Hierarchical (POSH) frameworks [33] have been used for modeling the decision-making processes and actions of agents. The framework used to model the mental process of the agent can also be used to provide explanations of the agent's behavior. Generalized, all three frameworks model the agent's behavior against three components: *goals* for which *actions* are accomplished based on an understanding of the current *state*. For increasing the user's understanding of the agent and its conduct, transparency of each component can be afforded. In experiments with an agent, built on the POMDP framework, explanations of the agent's inferred state (based on its sensors) and associated uncertainty were shown to benefit understanding, trust, mission success, and percentage correct decisions made, particularly when the agent's reliability was low. Noteworthy is that research within this domain has not explored AI algorithms for providing explanations.

B. *Explainable and Interpretable AI*

In the context of AI and ML systems, 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 timely make sense of the data. Echoing transparency, ML interpretability has been defined as the "*ability to explain or to present in understandable terms to a human*" ([34] p. 2).

In the field of interpretable ML, Guidotti et al. [35] found that graphical decision tree representations and textual decision rules (i.e., 'if-then') are among the most commonly used methods for explaining both the ML model (i.e., global explanation) and its specific output (i.e., local explanation) [35]. The output of linear models is often explained by highlighting key input parameters and their relative importance. Similarly, explanations of deep neural networks (DNNs) used for image recognition often make use of either saliency mask, which visualizes key areas/features in the input image, or activation maximization, which determines key neurons in certain layers activated by the input image. 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 [36], and the Visual Back Prop (VBP) method, which uses masks to visualize the set of pixels in the input image that determine the output [37].

Other recent developments include, for example, Local Interpretable Model-Agnostic Explanations (LIME) [38], Contextual Explanation Networks [39], and Contextual Decomposition [40]. However, some have been shown to be unstable (i.e., LIME) by providing different explanations for similar inputs that prevent their use in high stake domains [41]. Contextual Explanation Networks is an interesting approach to XAI that combines ML methods and probabilistic models [39]. Contextual Explanation Networks processes a subset of input features and generates parameters for a sparse linear model which can be assessed by domain experts. Subsequently, the generated model is applied to another subset of inputs and produces a prediction [39]. According to the developers is this approach robust and a candidate for high stake domains. Contextual Decomposition [40] provides explanations by decomposing the Long short-term memory (LSTM) output. Although primarily used for natural language processing, this approach should be able to provide importance scores for individual features and feature interactions also for other LSTM-based models and domains such as aviation. Moreover, reward decomposition approaches have been used to explain decisions, in particular action selection, of Reinforcement Learning (RL) agents [42].

Despite the amount of research on interpretability, there is a shortage of empirical research exploring the effects of transparency on acceptance and trust [32]. Research on transparency in the AI domains has been criticized for focusing primarily on *how to build* explanations while neglecting the underlying psychology and human interpretation of them [43].

C. *Personalised Automation and Strategic Conformance*

Automation is typically developed to optimize the performance of a task or solution to a problem, beyond that of human capabilities and preferences. In contrast, humans tend to apply heuristics that satisfice rather than optimize performance [44][45]. Westin et al. [46] argued that in contexts where humans and automation are expected to work together, the divergence in decision-making processes can negatively affect human acceptance and trust of the automation. Thereto, researchers have advocated automation that can adapt to an individual's needs and preferences [46] [47][48][49]. In the automobile domain, for instance, individual driving styles have been explored to increase the human acceptance and comfort of self-driving cars [50].

While AI systems are often designed with little regard to individual preferences and needs, they have the ability to adapt to the individual by selectively processing data. Ideally, the best way for achieving acceptance and trust may be to combine the strengths of a generalized AI model with that of a personalized AI model. Such a system could strive for optimal performance while accommodating individual differences. Moreover, by knowing the individual's preferences, the system would be able to afford transparency by explaining both why it suggests another solution (that deviates from the individual's), and why this solution is considered to be better.

The term strategic conformance was introduced to describe the match between human and automation solutions [46]. The concept of strategic conformance took inspiration from the concept of cognitive tools proposed in the European Commission sponsored "Role of the Human in the Evolution of ATM" (RHEA) project and later explored in the Conflict Resolution Assistant (CORA) project [51]. Hilburn et al. [52] demonstrated in the MUFASA project that strategic conformance can play a critical role in fostering acceptance and trust in an advanced ATM CD&R decision support system. ATCos' acceptance of, and agreement with, tactical CD&R automation both benefitted when solutions appeared to match the underlying strategies of the individual controller. However, the MUFASA project did not actually build a conformal system. Instead, conformance was achieved by replaying participants' previously captured (i.e., recorded) solutions and present them as the decision aid's recommended solutions. Since then, research has explored ways in which conformal systems can be designed to support ATCos on CD&R.

For example, in the study by Regtuit et al. [53] a RL agent was developed that could replicate human-like CD&R strategies based on ATC 'best practices' in simple two-aircraft conflict situations. Although the study showed promising results, the personalization of solutions and controller acceptance were not considered. A follow-on study by Van Rooijen et al. [54] aimed to achieve personalization by developing an individual prediction model of conflict resolutions based on pixel data of conflict situations. That study used a visual representation of velocity obstacles, in combination with a tailored Convolutional Neural Network (CNN), to predict controller solutions based on observed controller data collected in an ATC simulation. Results indicated that controller consistency and the selected (visual) feature(s) in conflict resolution play important roles in prediction accuracy.

III. Balancing Conformance and Transparency

The concepts of strategic conformance and automation transparency are partly at odds. Achieving the potential of AI decision aids implies solutions 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. In such instances, transparency may be essential for affording an understanding of automation behavior and reasoning, which drives acceptance and trust in the system.

Results from the MUFASA project, however, suggest that transparency may not be needed if acceptance and trust can be achieved by developing automation that solves problems using strategies conformal to the operator. Based on the *apparent* conformance of solutions, the human is likely to infer that the reasoning underlying the automation's proposed solutions was derived in a similar way to that of the human. But what is the benefit of having automation producing human solutions?

First, conformal automation can be particularly beneficial for fostering acceptance and trust in automation during the initial introduction phase [46]. Second, automation would be able to provide conformal solutions faster and more reliably than a human. Third, and most important, by knowing individual solution preferences, the system will be in a better position to afford transparency and explain its reasoning for recommending another (nonconformal) solution. The system will be able to provide an argument for why the proposed solution is better than that of the individual.

Given recent growth in ML methods and theory, the issues of conformance and transparency have taken on enormous urgency, and have given rise to three fundamental questions (Q). To what extent does automation allow us to understand:

Q1. What it will do?

Q2. How it will do it?

Q3. Why it will do it in a certain way?

While strategic conformance and transparency can support understanding of all three questions, they do so in different ways. Strategic conformance is hypothesized to foster acceptance based on how similar the system's solution (i.e., output) is with the human's preferred solution. Since the conformance of a system is judged based on its apparent behavior, the human will have difficulties knowing what the system does when solving a problem (Q1), *how* it solves a particular problem (Q2), and *why* a particular behavior/output/solution is chosen (Q3). If an ML system proposes a solution in line with the controller's strategy, the human may assume that the system will use a similar method (Q 1 and Q2) as the human and apply a similar underlying strategy (Q3). In an ATC CD&R task, the strategic conformance of the ML system can relate to the conflicts detected by the system or the conflict resolutions proposed by the system. Automation transparency, however, strives to explain to the ATCo how the automation will solve the problem and why. In a CD&R task, ML system transparency relates to the system's underlying reasoning when detecting conflicts or determining the most optimal conflict resolution.

Automation transparency is hypothesized to foster acceptance by providing an explanation for what the system does when solving a problem (e.g., explain the ML model or automation algorithm; Q1), how the automation will solve the specific problem (e.g., explain the relationship between input and output; Q2), and why this particular solution is chosen (e.g., because the following factors are considered/weighed most important for determining the output: a, b, x, and z; Q3).

Automation conformance and transparency can vary independently as shown in Table I. This can lead, at the extreme, to one of four outcomes. The impact of these four outcomes on human/system performance can vary with contextual factors, such as task complexity, time pressure, etc. For example, time pressure can change the tendency to accept automation output.

If both conformance and transparency are low, the human is likely to find it challenging to understand the automation. In reaction, the human perceives the automation to be *stupid*, weird, or even malfunctioning. Although the automation's solution may be optimal (given certain criteria) for the problem at hand, the human may reject it. Moreover, the human will be at greater risk of automation surprise, becoming out-of-the-loop, and losing situation awareness. If conformance instead is high, while transparency remains low, the human may perceive the automation to do the 'right thing' but be *confused* as to why. The solution derived by the automation, although similar to that of the human, can be achieved using very different methods (e.g., ML methods or human heuristic reasoning).

TABLE I. Automation Conformance and Transparency Matrix

		Transparency	
		Low	***High***
Conformance	***Low***	Stupid automation: *"It's doing a strange thing, and I don't understand why…"*	Peculiar automation: *"It's doing a strange thing, but I understand why…"*
	High	Confusing automation: *"It's doing the right thing, but I don't understand why…"*	Perfect automation: *"It's doing the right thing, and I understand why…"*

In situations when transparency is high and conformance low, the human may instead find the automation peculiar or interesting despite its strange behavior. This is perhaps the most interesting interaction between conformance and transparency as it reflects systems that surpass human performance while accommodating understanding of what the system is doing, how, and why. Assuming that automation is most beneficial when able to solve problems in ways exceeding human capabilities (e.g., considering big data), it will be intrinsically nonconformal and challenging for the human to understand. By affording transparency, however, the automation can provide explanations that foster acceptance, understanding, and potentially educate humans towards more optimal solutions. Finally, the human is likely to perceive the automation to be ’perfect’ if it provides a conformal solution that also is explained (provided the automation derives the solution differently). Contrary to the description, however, the solution in itself may not be perfect or optimal to the problem at hand. As such, perfect automation may not advance performance beyond that of human control alone.

IV. MAHALO ML Approach to CD&R

MAHALO is a European Union Horizon 2020 research project that aims to develop a prototype ML CD&R system, coupled to an enhanced Ecological User Interface (E-UI). MAHALO proposes to develop an ML system that learns from the individual ATCo, but also provides insight into what the system is learning. Simulations will be conducted to empirically explore the impact and balance between ML conformance and ML transparency on ATCos’ trust, acceptance, system understanding, and performance. Supervised Learning (SL) techniques will be explored to analyze controller data and develop generic (“one-size-fits-all”) and individualized prediction models (high conformance). RL will be used to derive more optimized solutions (low conformance). The challenge here is to discover the “features” that capture the breadth of human decision-making in the dynamic CD&R task. Any form of automation, whether based on AI techniques or a set of standardized rules and logic, tends to be designed as a “black box.” In MAHALO, the outputs of the ML models will be made transparent and explainable by adopting the Ecological Interface Design (EID) framework.

A. ML Approaches

Many ML methods make use of Artificial Neural Networks (ANNs, or just ‘neural nets’), which are often said to mimic the functioning of the human brain. ANNs consist of a group of nodes or neurons, each of which is a simple processor that can operate in parallel. A mapping from input to output space is created and strengthened by modifying connection weights (either positive or negative) between neurons, and connection weights are generally represented as a matrix. The principle underlying the behavior of these neurons is that knowledge can be represented through the cooperation of relatively simple neurons, with each one comparing a threshold value to the sum of weighted inputs and producing in response a nonlinear output. At the most basic level, an ANN simply sums the weighted (excitation or inhibition) activations from inputs. This net summed activation is then passed through a nonlinear activation function over the net input.

Neural network modeling has progressed over the last few decades for example through the introduction of multiple (deep) hidden layers (DNNs with intermediate neuron layers that combine weighted inputs to produce output via an activation function, thereby allowing the network to solve the XOR problem), non-binary methods for weight adaptation (specifically, the least mean squares approach), and relaxation algorithms for pacing error correction steps. The most common and simple current-day architecture incorporates feed-forward, backpropagation, and hidden layers. Recent architectural refinements to the neural network approach include CNN (for enhanced pattern recognition), recurrent neural networks (RNNs, which incorporate enhanced memory, and better handle time series data) and LSTMs (an extension of RNN approach that adds a ‘forget’ function, thereby allow a sliding window approach to time series processing).

The field of ML generally distinguishes three approaches:

- Supervised Learning (SL): labeled datasets of predictor (input) and criterion (output) pairs are used to train the SL model in classification (or, less commonly, regression)—that is, to approximate the relationship between input and output pairs;
- Unsupervised Learning (USL): does not use labeled data, which means the algorithm has to infer the natural structure present (i.e., find relevant relations) in the input data by itself;
- Reinforcement Learning (RL): a rule-based agent exploration of the environment, to maximize a reward function. RL is particularly powerful in cases of large solution space, in which clear predictor/criterion classifications are not practical. ANNs are used to train an internal model to the agent of the value of the states in its environment. From this internal model, the agent can derive a policy that will maximize rewards (or minimize penalties) from interaction with the environment. The agent-based approach of RL makes it especially useful in the field of autonomous control, where an agent has to learn how to control an unknown system by interacting with it.

B. Methodology: the Hybrid ML Approach

To realize a highly efficient ATM system, we expect that ATCos will have to closely collaborate with AI agents. Thereto, understanding the system will be of utmost importance for accepting and trusting it. Acceptance, trust, and understanding issues in human-automation interaction are some of the most difficult problems to solve. There is neither an easy-to-follow recipe nor a mathematical formula that captures and optimizes the dynamics of interaction. A hybrid ML approach is proposed to achieve a balance between high conformance and more optimal CD&R solutions. SL will be used to achieve high conformance by creating individualized CD&R prediction models from recorded controller data. As mentioned earlier, achieving high conformance may also imply that ‘better’ (and perhaps more optimal) solutions are not explored. By linking RL to the SL model, the ML system can explore more optimal solutions derived from CD&R 'best practices' and airspace performance criteria.

SL, the first part of the hybrid ML approach, involves the classification of input examples with output conditions. An SL model can come to associate a certain input pattern with an output criterion, even if the underlying relationship is nonlinear. SL models make use of multiple layers of hidden nodes to refine feature extraction. The automation aims to learn a relationship between an input set and a target set, based on the input/output data labels. A well-known example is to classify objects in a scenery based on labeled training data. Large nets can be created by combining and interconnecting both SL (e.g., LSTM), together with logic/code computations for real-time inference. Specifically, for SL the use of CNN / LSTM deep learning, an extension of general neural network modeling, appears promising. Computations in these inference pipelines are time-controlled and support feedback loops to maximize performance – at run time – of individual nets (i.e., SL and RL). Since nets and logic are combined, it is also possible to get partial explanations on internal computations within the pipeline.

Research has demonstrated the utility of SL in RPAS conflict detection – given enough training examples, an SL model can predict quite early and accurately when a traffic pattern will result in conflict [55]. However, there are some problems in which there are so many combinations that it is practically impossible to create a 'supervisor.' In the case of CD&R steps, SL might work well for conflict detection (i.e. predicting a conflict), but there are generally several ways to resolve a predicted conflict. SL has proven effective at conflict detection in dynamic ATC scenarios [54], however, SL by itself is not enough, since any mistakes in the training data will also be learned by the SL algorithm. It is in that sense not intelligent but will try to replicate the strategy underlying the training data, even if this strategy is not optimal. An SL system could be trained on specific resolutions, ending up with a model that mimics the controller. Although such a system could facilitate high trust and acceptance by being highly conformal to the individual controller, it would be limited to human performance. As such, SL alone is likely insufficient for CD&R ML.

The second part of the hybrid ML approach rests on rule-based RL. In RL, which is based on how humans learn, a policy is trained that will maximize the sum of expected future rewards, which are generated by applying actions in an environment. Some actions lead to high rewards, while others lead to low rewards. The RL agent builds an internal model of expected future rewards for a given state-action combination and uses this internal value-model to create a policy that will maximize the future rewards. An RL based layer can be added to the policy trained by SL. This RL layer makes the automation intelligent since it can detect obvious flaws in the strategy, based on a predefined cost or reward function, and will eliminate these flaws. Although some conformity will be lost in this step, the performance will be improved, and, by making the whole process transparent, it is expected that the acceptance of the automation will be improved as well. An important task is to design the rewards function of the RL layer in such a way that a good trade-off between conformance and performance is obtained. The internal model of expected future rewards that the agent builds can be used to explain its actions to the human operator. For example, it is possible to 'explain' why a certain action was selected, based on the individual contributions in the reward function, such as conformity to the strategy learned by SL, or the addition rule-based conditions.

The RL ruleset is best seen as a set of strategies and principles used by controllers. Examples can be found in EUROCONTROL's CORA work [51], as part of their effort to develop a smart medium-term CD&R advisory system for en-route controllers:

"Never put converging aircraft at the same altitude"

"Never put a faster aircraft in an overtake situation"

C. Experimenting with Conformance and Transparency

In MAHALO, a series of real-time, human-in-the-loop experiments will provide empirical insights into the impact of conformance and transparency on ATCo's trust, acceptance, system understanding, and performance. Should we develop automation that is conformal to the human, or should we develop (more optimal) automation that is transparent to the human? By manipulating levels of conformance and transparency, MAHALO aims to answer the high-level research questions.

The interaction between conformance and transparency is illustrated in Fig. 2. The rationale for developing a hybrid ML model is that it allows for manipulating levels of conformance. At one end, we propose to develop a conformal system based on SL where the AI learns to 'mimic' the human based on ATCo-derived data (e.g., by relating ATCo clearances to radar screen pixel data). We expect acceptance and trust to increase with a highly conformal system. While fully conformal, the ML system will be susceptible to the same mistakes as a human and thus can lead to inefficient solutions. At the other end, we allow the ML system to disregard the human and optimize solutions to CD&R problems. We expect this to benefit performance but also rejections and distrust. The system may generate solutions that ATCos may never come up with, or fully understand. Thereto, it will be challenging for the human to intervene in case of inadvertent 'stupid' or bad solutions, e.g., derived from spurious correlations or biased data in the training. We expect to find a middle ground that offers the best tradeoff between human and AI solutions. Using RL, the ML model can learn to solve problems more optimally while adhering to human-like best practices. Thus, less conformal to the individual, but more efficient from an airspace perspective.

Conformance is intended to be a naturally-occurring, non-manipulated variable. Earlier research using "Wizard of Oz"

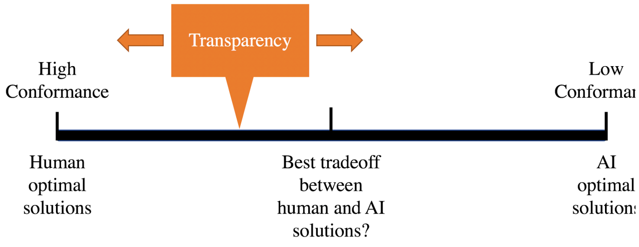

Fig. 2. Interaction between conformance and transparency.

techniques simulated conformance by using replaying self- or other generated solutions [47]. Part of the aim is to assess the performance of automation using a non-manipulative approach. An ATCO's conformance rating for a given ML solution defines the conformance level of that solution for that controller.

Transparency can be used to migrate closer to more optimal solutions while preserving human acceptance and trust. In relation to the solution conformance, the automation must be able to explain

- how the ML system has derived at a particular solution (i.e., process and relationship between input and output), and
- why the ML system has derived that a particular conflict resolution matches the individual's preferences (i.e., why it is strategic conformal).

To make the AI understandable we intend to create a shared mental model, using an EID interface that provides transparency by explaining the machine solution to the human. With respect to the time- and safety-critical aspects of CD&R, a difficult balance to achieve is the comprehensivity and simplicity of explanations. In situations where decisions have to made fast, time for detailed and exhaustive explanations, which best afford an understanding of the ML system, is limited. Explanations will address the three fundamental questions pertaining to automation behavior: what will it do, how, and why.

V. Conclusion and Future Work

The greatest strength of ML systems lies in their ability to learn, re-organize, and improve performance over time. When ML systems are required to work with humans, this ability may also be its greatest weakness. While ML systems approximate some unique human abilities, such as flexibility and adaptability, it also makes ML systems unpredictable and difficult to interpret. This is a major concern when considering introducing ML systems to safety-critical domains such as CD&R in ATM, which strive for control, predictability, and stability. Therefore, it is of utmost importance that the operators expected to work with these systems can understand them. This paper argues that a promising way forward for facilitating an understanding of ML systems is for the system to adapt to the individual and/or explain its behavior. We intend to build such a system in the MAHALO project and empirically explore the benefits and limitations of strategic conformance and automation transparency in real-time simulations. While fully automated systems may be an alternative in the long-term, the short-term and more realistic development of future ATM systems will to a large extent incorporate human-automation teams. A challenge is to design AI systems that humans accept, trust, and are willing to work with, and delegate tasks. Only when the automation is accepted can true collaboration be achieved and performance ambitions realized.

References

[1] SESAR, "European ATM Master Plan: Digitalising Europe's Aviation Infrastructure. Executive View," SESAR Joint Undertaking, Luxembourg, 2020. Retrieved from: https://www.sesarju.eu/sites/default /files/documents/reports/European%20ATM%20Master%20Plan%2020 20%20Exec%20View.pdf [Online].

[2] EAAI-HLG. "The FLY AI Report: Demystifying and Accelerating AI in Aviation/ATM," 2020. Retrieved from: https://www.eurocontrol.int/ publication/ fly-ai-report [Online].

[3] E. Koyuncu, and B. Başpınar, "Demand and capacity balancing through probabilisticqQueuing theory and ground holding program for european air transportation network," Anadolu Uni. J. Sci. Technol. A – Appl. Sci. Engi., 18(2), 2017, pp. 360-374.

[4] G. Vouros, J. M. Cordero, P. Costas, and G. Fuchs, "Data-driven aircraft trajectory prediction research (DART): Final project results report," D4.52018. Retrieved from: https://www.sesarju.eu/projects/dart [Online].

[5] R. Marcos, O. G. Cantú Ros, and R. Herranz, "Combining visual analytics and machine learning for route choice prediction: Application to pre-tactical traffic forecast," SID, Belgrade, Serbia, Nov. 28-30, 2017.

[6] H. Helmke, "Machine learning of speech recognition models for controller assistance (MALORCA): Final project results report. D5-3," 2018, Retrieved from: https://www.sesarju.eu/projects/malorca [Online].

[7] 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," 13th USA/Europe ATM R&D Seminar, Vienne, Austria. Jun. 18-19, 2019.

[8] D. Gunning, "Explainable Artificial Intelligence (XAI)," Defense Advanced Research Projects Agency (DARPA), 2017. Retrieved from https://www.darpa.mil/attachments/XAIProgramUpdate.pdf [Online].

[9] N. B. Sarter, D. D. Woods, and C. E. Billings, "Automation surprises," In G. Salvendy (Ed.), Handbook of Human Factors and Ergonomics (2nd ed.). New York, NY: Wiley, 1997.

[10] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: A systematic review of frequency, effect mediators, and mitigators," J. Am. Med. Inform. Assn., 19(1), 2011, pp. 121-127.

[11] D. A. Norman, "The 'problem' with automation: inappropriate feedback and interaction, not 'over-automation'," Philos. Tran.s R. Soc. Lond. B. Biol. Sci., 327(1241), 1990, pp. 585-93.

[12] M.R. Endsley, "Toward a theory of situation awareness in dynamic systems," Hum. Factors, 37(1), 1995, pp. 2-64.

[13] European Commission, Horizon 2020. 2020, Retrieved from: https://ec.europa.eu/programmes/horizon2020/en [Online].

[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-Papers Online, Kyoto, Japan, 2016 pp. 201-206.

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

[16] M. C. Dorneich, R. Dudley, E. Letsu-Dake, W. Rogers, S. D. Whitlow, M. C. Dillard, and E. Nelson, "Interaction of automation visibility and information quality in flight deck information automation," IEEE Trans. Human-Mach. Syst., vol. 47, 2017, pp. 915-926.

[17] T. B. Sheridan, Telerobotics, Automation, and Human Supervisory Control, Cambridge, MA: MIT Press, 1992.

[18] D. D. Woods, "Decomposing automation: Apparent simplicity, real complexity, in Automation and human performance: Theory and applications," R. Parasuraman and M. Mouloua, Eds., ed Hillsdale, NJ, England: Lawrence Erlbaum Associates, 1996, pp. 3-17.

[19] J. L. Campbell, J. L. Brown, G. J. S, C. M. Richard, M. G. Lichty, T. Sanquist, P. Bacon, R. Woods, H. Li, D. N. Williams, and J. F. Morgan, "Human factors design guidance for driver-vehicle interfaces," National Highway Traffic Safety Administration, Washington, DC DOT HS 812 360, Dec. 2016.

[20] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," PNAS USA, 116(44), 2019, pp. 22071-22080.

[21] V. Bellotti, and K. Edwards, "Intelligibility and accountability: Human considerations in context-aware systems," Hum.-Comput. Interact., 16(2), 2001, pp. 193-212.

[22] L. Edwards, and M. Veale, "Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for," Duke L. & Tech. Rev., 18, 2017

[23] T. Kim, and P. Hinds, “Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction,” 15th IEEE ROMAN, Hatfield, UK, Sep. 6-8, 2006, pp. 80-85.

[24] 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. Hum. Comput. Stud., 58(6), 2003, pp. 697-718.

[25] M .G. Core, H. C. Lane, M. van Lent, D. Gomboc, S. Solomon, and M. Rosenberg, “Building explainable artificial intelligence systems,” 18th IAAA Conference on Innovative Applications of Artificial Intelligence - Volume 2, Boston, MA, Jul. 2016, pp. 1766-1773.

[26] A. R. Selkowitz, S. G. Lakhmani, J. Y. C. Chen, and M. Boyce, “The effects of agent transparency on human interaction with an autonomous robotic agent,” HFES Annual Meeting, 59(1), 2015, pp. 806-810.

[27] G. A. Jamieson, L. Wang, and H. F. Neyedli, “Developing human-machine interfaces to support appropriate trust and reliance on automated combat identification systems,” Cognitive Eng. Lab, University of Toronto, Contract Report W7711-068000/001/TOR, Mar. 2008.

[28] T. Helldin, G. Falkman, M. Riveiro, and S. Davidsson, “Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving,” 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Eindhoven, The Netherlands, Oct. 2013.

[29] J. B. Lyons, “Being transparent about transparency: A model for human-robot interaction,” AAAI Spring Symposium on Trust in Autonomous Systems, Palo Alto, CA, Mar. 25-27, 2013, pp. 48-53.

[30] A. Bhaskara, M. Skinner, and S. Loft, “Agent transparency: A review of current theory and evidence,” IEEE Trans. Human–Mach. Syst., 50(3), 2020, pp. 215-224.

[31] A. S. Rao, and M. P. Georgeff, “BDI agents: From theory to practice,” 1st ICMAS, San Fransisco, CA, Jun. 12-14, 1995.

[32] N. Wang, D. V. Pynadath, and S. G. Hill, “Trust calibration within a human-robot team: Comparing automatically generated explanations,” 11th ACM/IEEE International Conference on Human Robot Interaction, Christchurch, New Zealand, Mar. 7-10, 2016, pp. 109-116.

[33] A. Theodorou, R. H. Wortham, and J. J. Bryson, “Designing and implementing transparency for real time inspection of autonomous robots,” Connect. Sci., 29(3), 2017, pp. 230-241.

[34] F. Doshi-Velez, and B. Kim, “Towards a rigorous science of interpretable machine learning,” arXiv e-prints: 1702.08608v2, 2017.

[35] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” ACM Comput. Surv., 51(5), Article 93, 2018.

[36] 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, 10(7), pp. e0130140, 2015.

[37] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, U. Muller, and K. Zieba, “VisualBackProp: Efficient visualization of CNNs, arXiv e-prints: 1611.05418, 2016.

[38] M. T. Ribeiro, S. Singh, C. Guestrin. ”Why should I trust you?: Explaining the predictions of any classifier,” 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, Aug. 2016, pp. 1135–1144.

[39] M. Al-Shedivat, K. A. Dubey, E. P. Xing, “Contextual explanation networks,” arXiv e-prinst: 1705.10301, 2017.

[40] W. J. Murdoch, P. J. Liu, and B. Yu, “Beyond word importance: Contextual decomposition to extract interactions from LSTMs,” arXiv e-prinst: 1801.05453, 2018.

[41] D. Alvarez-Melis, T. S. Jaakkola, ”On the robustness of interpretability methods,” 3rd ICML WHI, Stockholm, Sweden, Jul. 14, 2018

[42] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, and F. Doshi-Velez, "Explainable reinforcement learning via reward decomposition," 28th IJCAI. Workshop on Explainable Artificial Intelligence., Macau, China, Aug. 11, 2019, pp. 47-53.

[43] 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,” CHI Conference on HF in Computing Systems, Montreal QC, Canada, Apr. 21-26, 2018.

[44] H. A. Simon, “Rational choice and the structure of the environment,” Psychol. Rev., 63(2), 1956, pp. 129-138.

[45] G. Gigerenzer, R. Hertwig, and T. Pachur, Heuristics: The Foundations of Adaptive Behavior, New York, NY: Oxford University Press, 2011.

[46] C. Westin, C. Borst, and B. Hilburn, “Strategic conformance: Overcoming acceptance issues of decision aiding automation?” IEEE Trans. Human–Mach. Syst., 46(1), 2016, pp. 41-52.

[47] Y. Liu, Y. Lee, an A. N. K. Chen, “Evaluating the effects of task–individual–technology fit in multi-DSS models context: A two-phase view,” Decis. Support Syst., 51(3), 2011, pp. 688-700.

[48] 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., 10(5), 2009, pp. 381-397.

[49] R. Parasuraman, and Y. Jiang, “Individual differences in cognition, affect, and performance: Behavioral, neuroimaging, and molecular genetic approaches,” Neuroimage, 59(1), 2012, pp. 70-82.

[50] M. Kuderer, S. Gulati, and W. Burgard, “Learning driving styles for autonomous vehicles from demonstration,” IEEE ICRA, Seattle, WA, May 26-30, 2015.

[51] B. Kirwan, and M. Flynn, “Investigating air traffic controller conflict resolution strategies,” Report asa.01.cora.2.del04-b.rs. EUROCONTROL. Brussels, Belgium, Mar. 2002.

[52] 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,” Air Traffic Control Q., 22(2), 2014, pp. 115-136.

[53] R. M. Regtuit, C. Borst, E.-J. van Kampen, and M. M. van Paassen, ”Building strategic conformal automation for air traffic control using machine learning,” AIAA Information Systems-AIAA Infotech at Aerospace, Kissimmee, FL, Jan. 9-12, 2018.

[54] S. J. 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., 28(3), 2020, pp. 1-9.

[55] P. García, O. G. Cantú Ros, C . Ciruelos, and R. Herranz, “Understanding door-to-door travel times from opportunistically collected mobile phone records: A case study of spanish airports, SID, Belgrade, Serbia, Nov. 28-30, 2017.