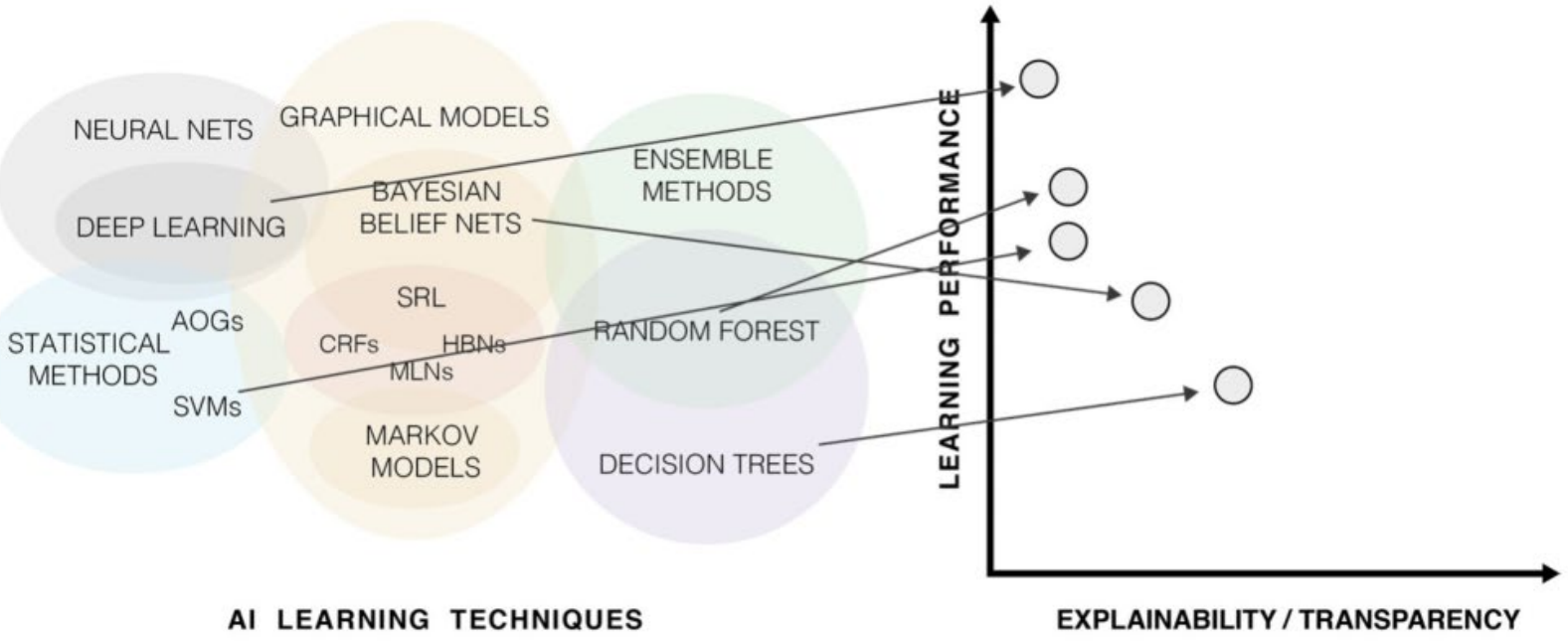
MAHALO AI-based CD&R tools: Conformity vs. Transparency

Brian Hilburn & Tiago Monteiro Nunes
ENGAGE TC2 workshop: AI, ML, and Automation
3 Sep 2021

MAHALO
Modern ATM via Human/Automation Learning Optimisation

Inspiration



AI performance vs explainability (after Gunning, 2017)

Automation transparency

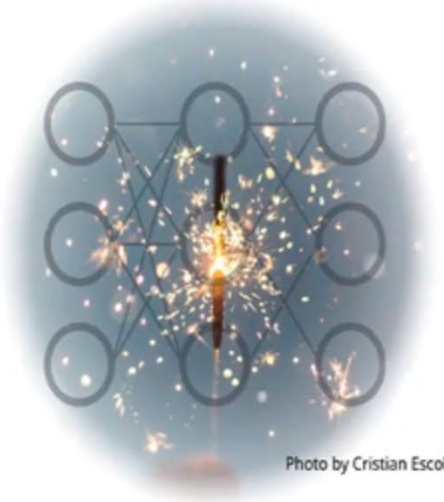
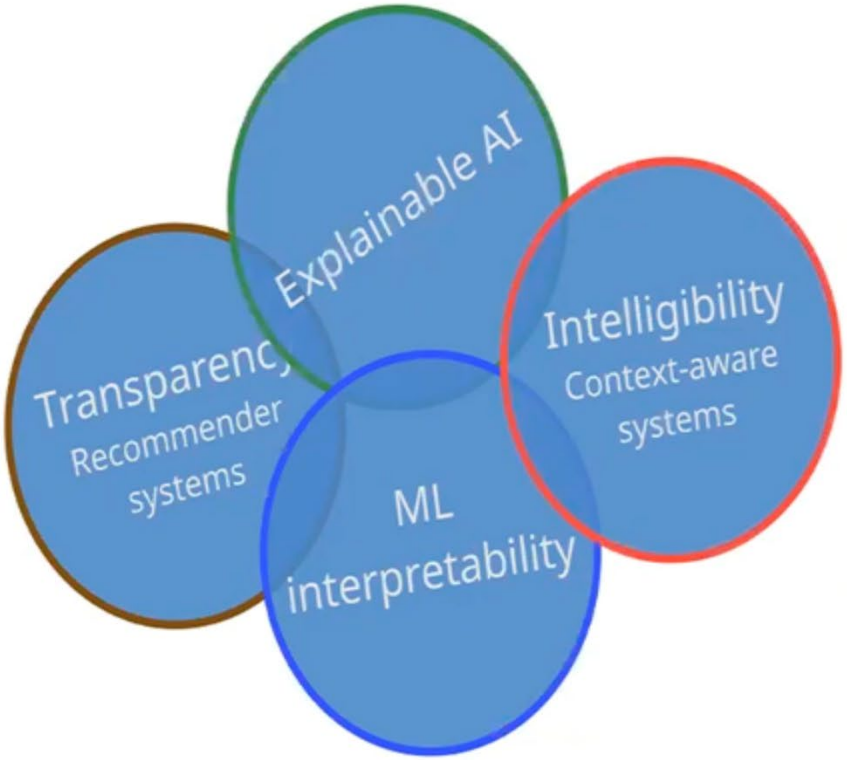


Photo by Cristian Escobar

"the automation's ability to afford understanding and predictions about its behaviour"



Should we build automation that is TRANSPARENT or CONFORMAL?

How to build ML?

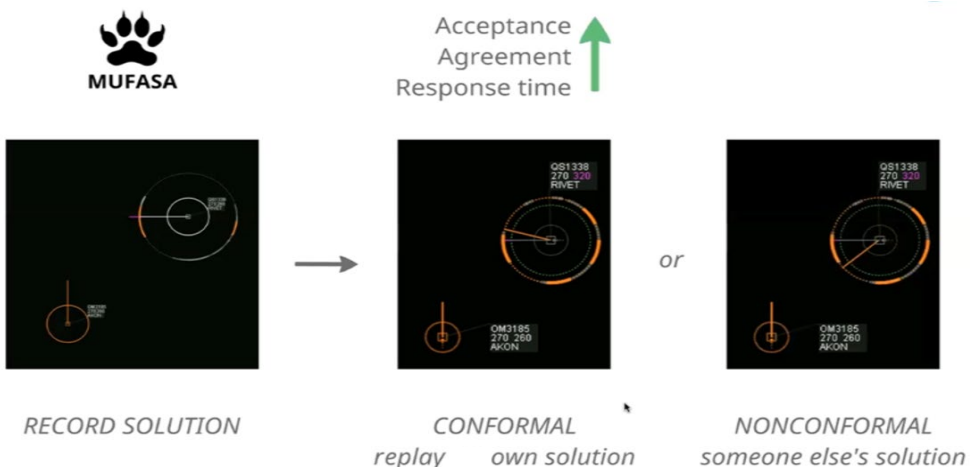
Transparency— is automation's inner process explainable to human?

Conformance— does automation seem to match human strategies?

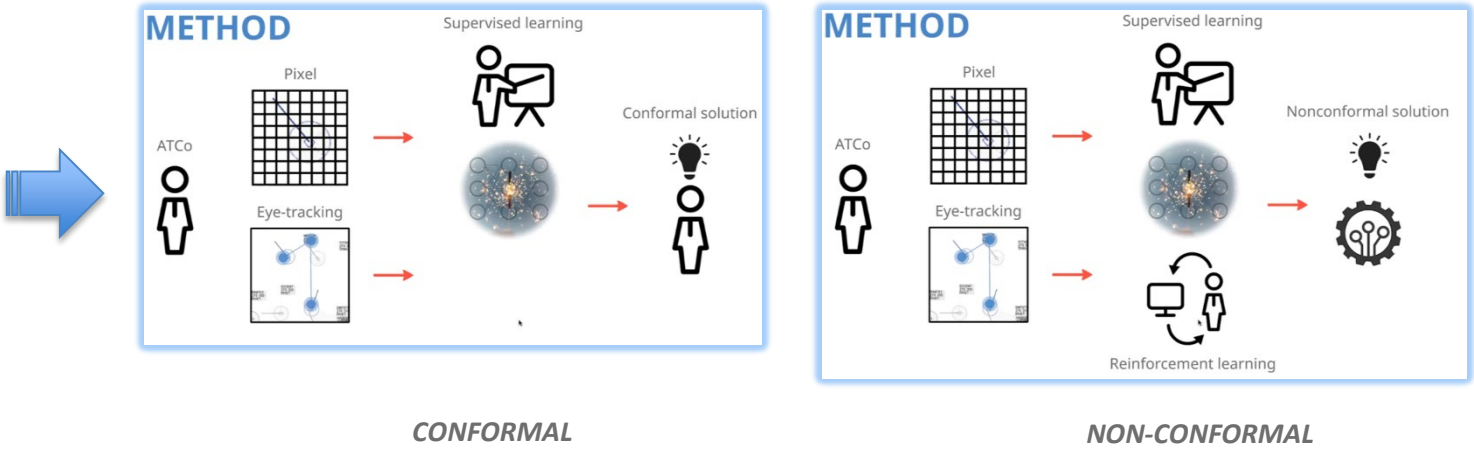
		TRANSPARENCY	
		Low	High
CONFORMANCE	Low	**Stupid automation:** *"It's doing a strange thing, and I don't understand why…"*	**Peculiar automation:** *"It's doing a strange thing, but I understand why…"*
	High	**Confusing automation:** *"It's doing the right thing, but I don't understand why…"*	**Perfect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

Conformal ML: Fake it or make it!

MUFASA (2011)

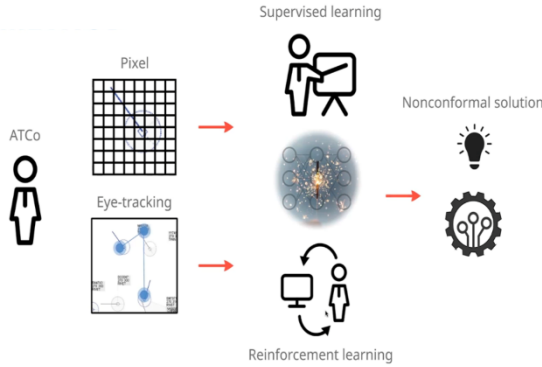


MAHALO (2020)

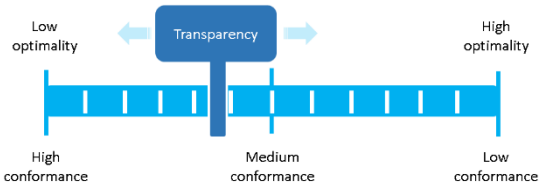


MAHALO Objectives

Develop a Hybrid ML model for CD&R via
- **Supervised Learning** (SL)– deep learning– to detect / classify conflicts
- **Reinforcement Learning** (RL)– rule based– to resolve conflicts



Evaluate (via HITL sims) how conformance and transparency impact: Acceptance, Understanding, Trust, Workload, Performance



Derive general design lessons

Conformance Variable

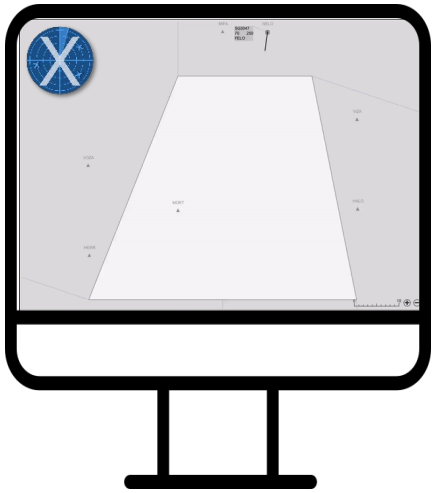
Conformal (SL) model

Optimal (RL) model

Personalized solutions

Hybrid solutions

Optimized solutions



Varies with individuals and cohort

Conformance Pre-test

Same for all simulations

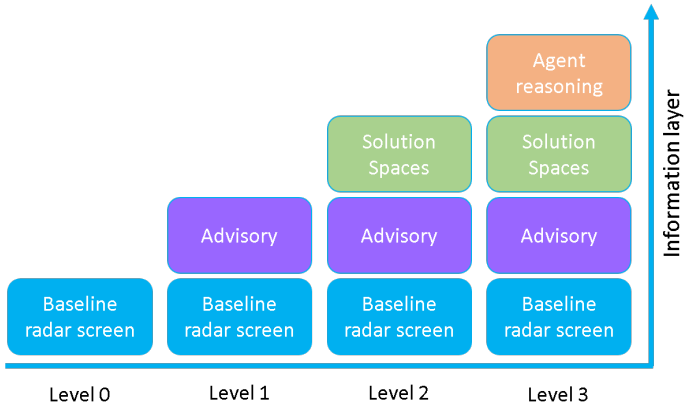
Participant A's preference

Personalized solution (C)

Hybrid Group average solution (GC)

Optimal solution (NC)

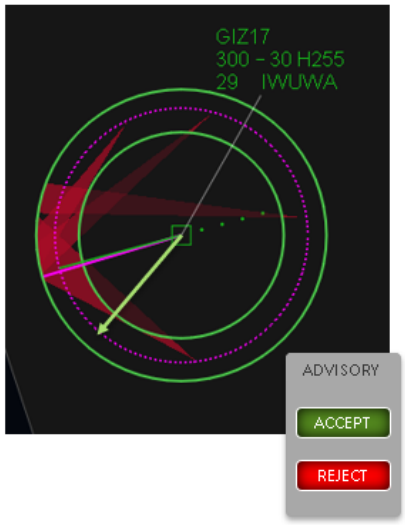
Transparency Variable



No transparency (T1)

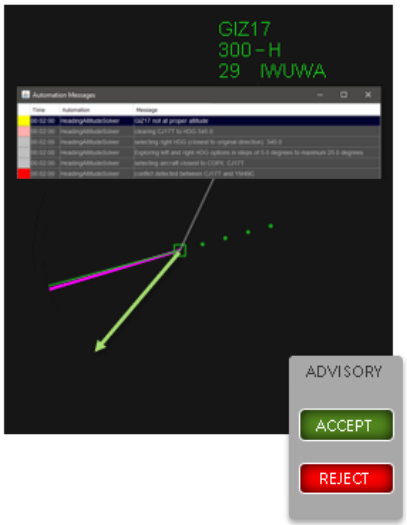
No underlying rationale

Domain transparency (T2)



Constraints affecting solution option

Agent transparency (T3)

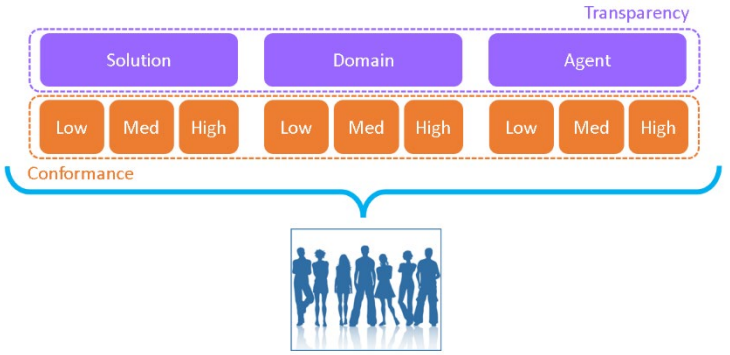


Rationale underlying advisory (according to ML model)

Experiment

Design

- 3x3 within subjects



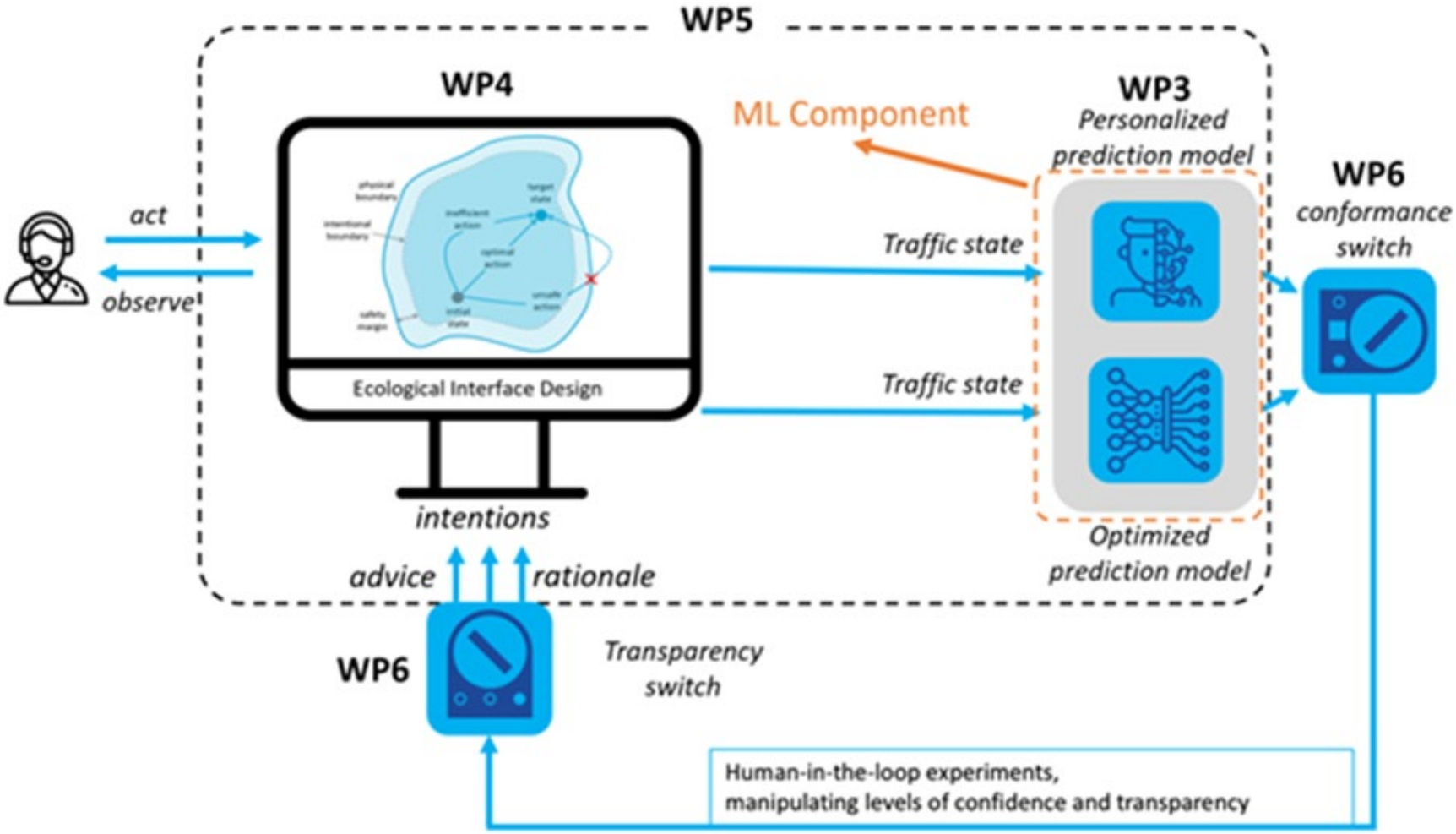
Dependent variables

- Advisory agreement rating
- Self-reported workload (ISA, RSME)
- Agent understanding (survey re confidence, understanding…)
- Trust in automation (SATI)
- Supervisory control performance:
  - Attention allocation (eye tracking?)
  - Spotting and reporting anomalies (?)
  - Activity (mouse clicks, label drags, …)
  - …

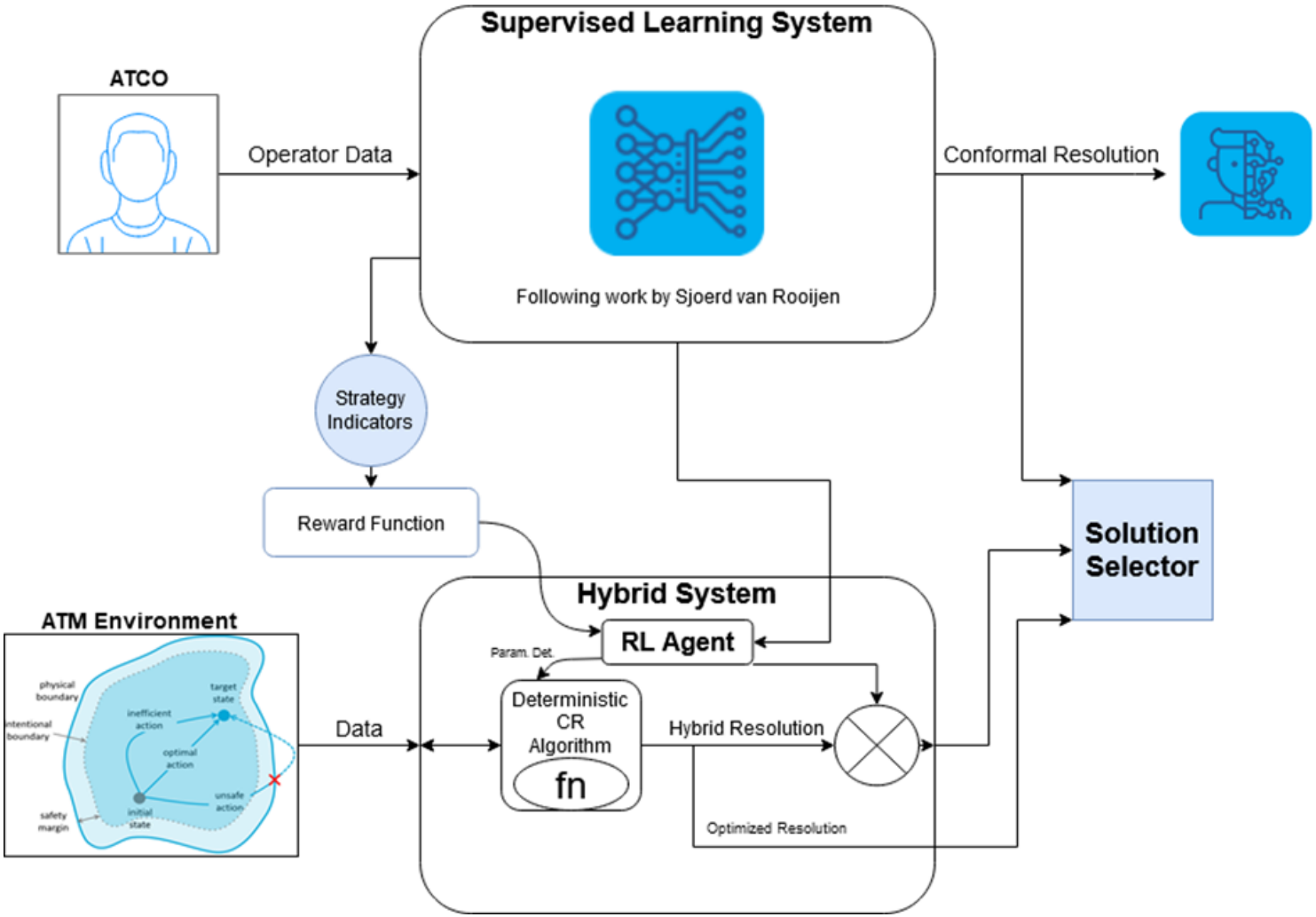
Simulator: SectorX

- ✓ Present day / futuristic control
- ✓ Single executive ATCO
- ✓ En-route (3D)
- ✓ Tactical CD&R
- ✓ Larger sectors
- ✓ Various traffic densities
- ✓ Realistic acft behaviour (BADA)
- ✓ Simplifying assumptions (e.g. no wind)

MAHALO ConOps

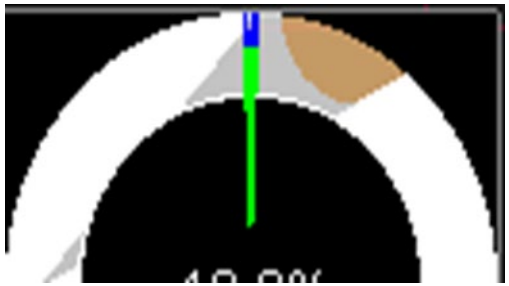


Hybrid SL-RL Model



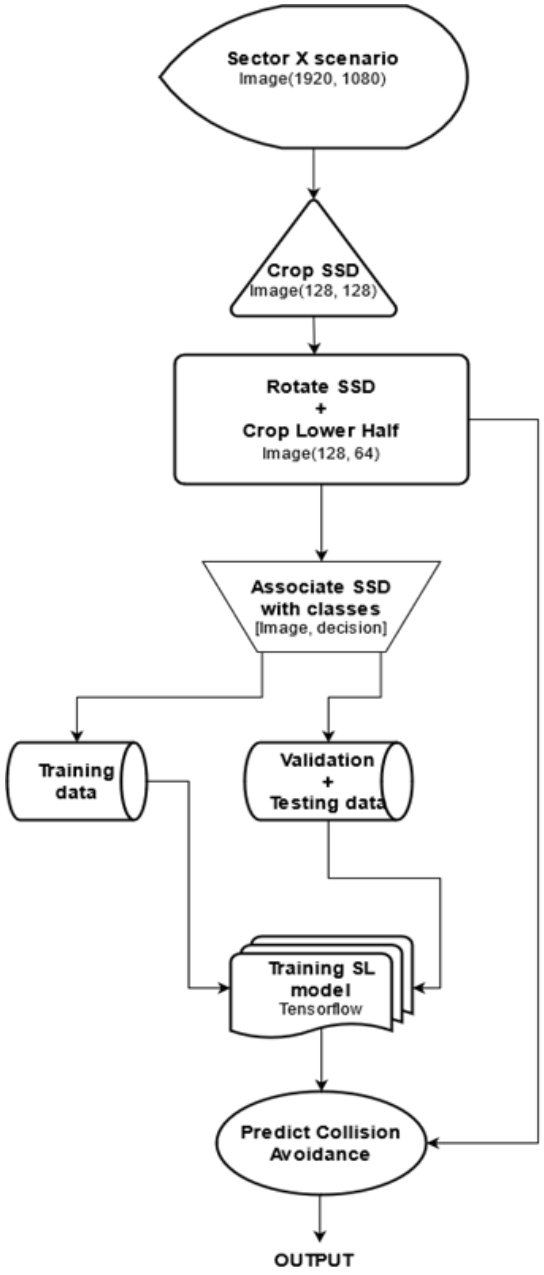
ML approaches

- SL
  - Provides conformal solutions
  - 128x64 images; pixel data
  - Using CNN (van Rooijen, 2019)



- RL
  - Provides optimal solutions (per cost *f*s)
  - Using DDPG framework (continuous action spaces; flexible)
  - Also DQfD
- Hybrid
  - SL → RL
  - Considers conformance and optimality

SL model flow chart



DQfD Pseudo Code

Algorithm 1. The pseudo-code of Deep Q-Learning from Demonstrations (DQfD) [37]. The behaviour policy $\pi^{\epsilon Q_\theta}$ is ϵ-greedy with respect to Q_θ.

Require: $\mathbb{D}^{replay}$: initialised with demonstration data set; θ: weights for initial behaviour network (random); θ': weights for target network (random); τ: frequency at which to update target net; k: number of pre-training gradient updates; α: learning rate; $N_{\text{training epochs}}$: number of epochs for training

1: **for** steps t $\in \{1, 2, ...k\}$ {pre-training phase} **do**
2: Sample a mini-batch of n transitions from D_{replay} with prioritisation
3: Calculate loss L(Q) using target network
4: Perform a gradient descent step to update θ
5: **if** t **mod** τ = 0 **then**
6: $\theta' \longleftarrow \theta$ {update target network}
7: **end if**
8: $s \longleftarrow s'$
9: **end for**
10: **for** steps t $\in \{1, 2, ..., N_{\text{training epochs}}\}$ {*normal training phase*} **do**
11: Sample action from behaviour policy $a \sim \pi^{\epsilon Q_\theta}$
12: Play action a and observe (s', r)
13: Store (s, a, r, s') into D^{replay}, overwriting oldest self-generated transition if over capacity occurs
14: Sample a mini-batch of n transitions from D^{replay} with prioritisation
15: Calculate loss L(Q) using target network
16: Perform a gradient descent step to update θ (Adam optimiser)
17: **if** t **mod** τ = 0 **then**
18: $\theta' \longleftarrow \theta$ {update target network}
19: **end if**
20: $s \longleftarrow s'$
21: **end for**

DDPG Pseudo Code

Algorithm 1 DDPG algorithm

Randomly initialize critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ.
Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$
Initialize replay buffer R
for episode = 1, M **do**
  Initialize a random process $\mathcal{N}$ for action exploration
  Receive initial observation state s_1
  for t = 1, T **do**
    Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise
    Execute action a_t and observe reward r_t and observe new state s_{t+1}
    Store transition (s_t, a_t, r_t, s_{t+1}) in R
    Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from R
    Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$
    Update critic by minimizing the loss: $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$
    Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$$

    Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

  end for
end for

Some challenges

- Data sample size – synthetic generation being explored
- SL fine tuning– currently only L / R / NONE HDG chg; not degree
- RL data inefficient– prime pump w demos (DQfD)– need good training data
- Exploration vs exploitation problem in RL– MAHALO greedy approach
- Compromise between learning stability and replay buffer size
- Choice of acft
- What if no variance between ATCOs (personal=group)
- What if ATCOs choose optimal (conformal=optimal)

http://mahaloproject.eu

Contact:

Stefano Bonelli stefano.bonelli@dblue.it
Tiago Nunes t.m.monteironunes@tudelft.nl
Brian Hilburn brian@chpr.nl

SUPPORTED BY SESAR JOINT UNDERTAKING

deepblue consulting & research

CHPR

TU Delft

LINKÖPING UNIVERSITY

LFV AIR NAVIGATION SERVICES OF SWEDEN