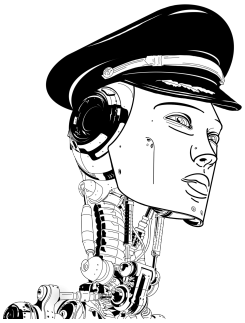
J. Michael, J. Pfeiffer, A. Wortmann (Hrsg.): Software Engineering 2022 Workshops,
Digital Library, Gesellschaft für Informatik 2022 157

An Anthropomorphic Approach to Establish an Additional Layer of Trustworthiness of an AI Pilot

The Concept

Christoph Regli[1] Bjoern Annighoefer[2]



Abstract: AI algorithms promise solutions for situations where conventional, rule-based algorithms reach their limits. They perform in complex problems yet unknown at design time, and highly efficient functions can be implemented without having to develop a precise algorithm for the problem at hand. Well-tried applications show the AI's ability to learn from new data, extrapolate on unseen data, and adapt to a changing environment — a situation encountered in flight operations. In aviation, however, certification regulations impede the implementation of non-deterministic or probabilistic algorithms that adapt their behaviour with increasing experience. Regulatory initiatives aim at defining new development standards in a bottom-up approach, where the suitability and the integrity of the training data shall be addressed during the development process, increasing trustworthiness in effect. Methods to establish explainability and traceability of decisions made by AI algorithms are still under development, intending to reach the required level of trustworthiness. This paper outlines an approach to an independent, anthropomorphic software assurance for AI/ML systems as an additional layer of trustworthiness, encompassing top-down black-box testing while relying on a well-established regulatory framework.

Keywords: AI; artificial intelligence; ML; machine learning; aviation; AI pilot; avionics; cockpit; certification; licencing; trust; trustworthiness; black-box testing; independent software assurance; post-market monitoring; pilot training; flight instructor; pilot checking; flight examiner; anthropomorphism; dehumanization

[1] ZHAW School of Engineering, Technikumstrasse 9, CH-8401 Winterthur, Switzerland, regl@zhaw.ch
[2] University of Stuttgart, Pfaffenwaldring 27, D-70569 Stuttgart, Germany, bjoern.annighoefer@ils.uni-stuttgart.de

1 Introduction

Artificial intelligence (AI) may be defined as *"any technology that appears to emulate the performance of a human"* [**EASA-AIRM**], *"the capacity of an agent to select the appropriate strategy in relation to its goals"* [**Baeldung**], or *"acting humanly, thinking humanly, thinking rationally, and acting rationally"* [**AIModernApproach**].

In the past, AI experienced quite some springs and winters, periods of excitement followed by rather long stretches of disappointments. Recently, however, AI-based software demonstrated remarkable capabilities. The massive increase in computing power, new sophisticated algorithms, and the explosion in the amount of available data drove significant and tangible advances, heralding another AI spring [**AISpring**]. Nowadays, ubiquitous applications, e. g. in neural language processing or image recognition, provide a glimpse of the potential of AI, exciting the imagination of humankind of what else might be possible.

Data-driven machine learning methods, a subset of AI, are algorithms whose performance may improve as they are exposed to data, with the ability to continue self-learning even after deployment [**EASA-AIRM**]. Several experiments that led to media attention showed the potential of *deep reinforcement learning*, a specific class of data-driven machine learning methods [**DeepReinforcementLearning**], e. g. IBM's Deep Blue beating the world chess champion in 1997 [**IBM-DeepBlue**] or DeepMind's AlphaGo defeating a Go champion [**Deepmind-AlphaGo**], algorithms autonomously driving vehicles [**AutonomousDriving**] or playing video games [**VideoGames**].

Throughout this document, the generic term *AI* will be used, implicitly including machine learning with deep neural networks [**DeepLearning**].

The Potential of AI in Aviation

For the aviation sector, [**EASA-AIRM**], [**FlyAI**] and [**AIClassification**] list some potential applications, like flight automation, flight controls, natural language processing, cyber-security, predictive maintenance or development assistance. The aviation industry even offers concrete products or publishes ongoing projects, e. g. [**Airbus**], [**Acubed**], [**Daedalean**], or [**Paladin**].

By envisaging the *holy grail* [**HolyGrail**] of AI in aviation, i. e. the fully autonomous flight, quite some challenges lie ahead (refer to section *The Challenges of Certifying AI* below). Even though the current regulatory framework does not allow for AI applications in aviation as of today, a multitude of applications related to an AI-based pilot with different levels of automation can be identified as intermediate milestones, potentially yielding inspiring insights and results:

Coping with complexity The range of use of *pilot assisting and automation systems* could be expanded by the potential of AI to learn non-linear and complex relationships,

infer new relationships, generalize, predict on unseen data, and perform adaptively in complex and highly dynamic environments with uncertainties.

Informed monitoring/assisting layer Statistical safety evidence may be required for the final certification of AI systems. While acting in shadow mode, an upcoming AI system could acquire experience and, later on, provide hints and suggestions to human pilots in the sense of an additional monitoring/assisting layer. Unlike a human pilot, an AI system is neither subject to emotions nor imperilled by startle effects [**StartleEffect**].

Dehumanizing If there is eventually a way to certify an AI-based pilot, an AI flight instructor or flight examiner could as well be feasible. Such systems could train and check human pilots, allowing an unbiased comparison amongst the pilots' corps, or train and check AI pilots. Taking this idea even further, a flight examiner routine could continuously run in the background, flagging any decrease in the pilot's performance — or an AI pilot could have to pass several check flights even concurrently to an actual flight mission.

Sophisticated training devices Certified FNPTs[3] or desktop-based flight simulators could profit from AI algorithms. Such algorithms could assist the flight instructor, profit from the flight instructor's experience, or even coach training sessions or parts thereof autonomously. In effect, such training devices could lead to cost reductions in training or more sophisticated training assistance for (non-certified) desktop-based flight simulators for the interested public.

Sharing experience Upon the retirement of a human pilot, the experience acquired during the aviation career is lost, at most partially divulged to trainees if still engaged as a ground or flight instructor. There is ongoing research about conserving and disseminating the *experience* of individual AI systems, with federated or collaborative learning methods [**FederatedLearning**]. Driving this idea to its ultimate end, an AI pilot could look back to an experience of millions of flight hours, compared to maybe 20k hours in a career of a full-time airline pilot — a proliferation of experience gained by a large number of AI pilots to other AI pilots, multiplying the accumulated flight experience.

New approaches Depending on the definition of the term *creativity*, AI could lead to insights on creative yet unknown approaches [**Creativity**]. Indeed, the board game experiments mentioned above ([**IBM-DeepBlue**], [**Deepmind-AlphaGo**]) revealed new or unusually focused strategies for winning the game. Research papers about artificial curiosity and creativity have been published, e. g. [**Creativity1**] or [**Creativity2**].

[3] Flight and Navigation Procedures Training devices

The Challenges of Certifying AI

While exploring that potential of AI, several challenges with respect to certification have to be tackled, e. g. in the development process, in the requirements engineering, in predictability, explainability, robustness, and validation [**EASA-AIRM-Trustworthiness**].

Predictability Adaptive systems may change their behaviour depending on their experience. AI algorithms with enabled self-learning (on-the-fly learning) may act differently under the same preconditions, manifesting a *non-deterministic* or a *probabilistic* behaviour [**EASA-AIRM**], depending on what the system has learned so far. This acquired experience may improve the system performance or worsen it.

Explainability Accidents and incidents in aviation are subject to investigation for the purpose of prevention of accidents and incidents [**ICAO-13**]. Consequently, AI systems involved in accidents or incidents as well would have to be examined so that an upgrade will hopefully prevent future occurrences. But AI systems are rather black boxes with an enormous number of parameters, making it difficult to retrace the proceedings in hindsight. — [**EASA-AIRM**] features *Explainability of AI* as one of the *AI Trustworthiness Building Blocks* and lists research initiatives in that context, e. g. [**DARPA**]. In addition, explainable AI is a wide-spread research topic, e. g. [**ExplainableAI**] or [**Google**].

Insurance Going even beyond the certification requirements, operational systems must be insured, as failures may lead to cost-intensive efforts. Insurance companies may impose additional requirements for their insurance cover, e. g., in manned aviation, postulate a particular flight experience of pilots, exceeding the licencing requirements stipulated by the regulatory framework. Such additional requirements have to be anticipated when dealing with AI systems. [**AIInsurance**] elaborates on that issue.

Social acceptance The operational use of advanced automation systems, e. g. a fully autonomous flight, will have to be publicly accepted. Non-representative surveys conducted during the writing of this paper show divergent acceptance, depending on the age or the subject knowledge of the surveyed persons, among other things. Furthermore, social acceptance is influenced by the fear of job losses, the fear for AGI[4] — research topics in economy (e. g. [**Forbes**]) and psychology (e. g. [**Frontiers**]), amongst others.

The current certification standards stipulate deterministic behaviour and explainable systems [**RTCA-DO178C**], encompassing some of these challenges. Moreover, certification, insurance and social acceptance rely on a certain **trustworthiness** of the system in question [**EASA-AIRM**], rendering *trust* as the central issue for AI in safety-critical areas like aviation.

[4] Artificial General Intelligence; the hypothetical human-like general intelligence, not narrowed to a specific task

The Anthropomorphic Approach

Seen from a distance, however, the overarching issue of trust crystallizes as well-known and daily encountered, as there *is* already another element in the aviation system that is self-learning, non-deterministic or probabilistic, occasionally even lacking its own explainability: The human being.

Well-defined and well-proven standards and processes aim to ensure the required level of safety and trustworthiness of systems where humans are involved, 'certification' is replaced by 'licencing', insurance as well is based on a certain trustworthiness, and social acceptance can be regarded as given in most cases.

In short, the general idea of the anthropomorphic approach is to apply the regulatory framework for (human) pilot training and checking [**EASA-FCL**] to AI pilots. In future, the certification of a machine might be replaced, but could for sure be *extended* by a licencing of a machine — the issuing of a pilot licence in analogy to manned aviation, as an additional layer of trustworthiness.

The fictitious scenario starts with a given black box — of unknown origin — claiming to be able to fly an aircraft[5]. This leads to the central research question:

How is it possible to gain trust in an AI pilot?

2 State of the Art

According to the EASA AI Roadmap, first approvals of AI/ML algorithms are expected in 2025, and roadmaps of major players foresee single-pilot commercial air transport operations in 2030 and autonomous commercial air transport operations in 2035 [**EASA-AIRM**]. The EASA AI Task Force currently is, together with experts from the industry, working on CoDANN[6] [**EASA-CoDANN**]. In addition, guidance material for level 1 ML applications (assistance to human) has been published [**EASA-CoDANN**].

One of the main contributions of the CoDANN approach is to integrate learning assurance into the development process, extending the traditional development assurance framework, ensuring the integrity of the training data. This approach may be regarded as bottom-up as it concerns the development process.

In addition, and on a more strategic level, the European Commission (EC) produces a coordinated plan on artificial intelligence, a proposal for a regulation on artificial intelligence, a strategy and a white paper on artificial intelligence [**EC**], as well as ethic

[5] as an autonomous level 3 system according to the EASA classification [**EASA-AIRM**], setting the investigation focus on the AI system, excluding the human element

[6] Concepts of Design Assurance for Neural Networks

guidelines for trustworthy AI [**EC-Guidelines**]. These guidelines call for lawful, ethical, and robust AI systems and list corresponding key requirements. The proposed *Regulation on Harmonized Rules on AI* covers, inter alia, AI systems that continue to learn in the productive environment, stipulating a post-marked monitoring system (title VIII), embedded as a test and validation process into the model governance framework, according to [**KI-Regulierung**] and [**AIRequirements**].

Industry-driven initiatives try to follow the standard certification framework by disabling the post-deployment self-learning feature of AI algorithms in order to achieve a deterministic system that is testable under the current regulations. The intention is to train the AI model in a standard way, then to freeze the model and to test it conventionally [**Acubed**]. However, this approach prevents the system from improving its performance after deployment — one of the key features of AI systems is sacrificed.

Academic papers report on early successes with AI trained to fly simulated aircraft with behavioural cloning and reinforcement learning [**LearningToFly**]. However, no concrete approach to certify such systems under the current regulation is mentioned.

In a short paragraph of an early NASA paper [**NASA-Adaptive**], the idea of a paradigm shift from certification to licencing has been mentioned as one option to gain trust in adaptive systems. However, no evidence is available whether this approach has been followed up.

This is where the present paper steps in, not intending to illustrate a complete shift from certification to licencing, but in an effort to try to combine the two pillars of certification *and* licencing, with the noble goal to achieve, if possible at all, a trustworthy system — pending a formal and applicable definition of the term *trustworthiness*.

3 The Anthropomorphic Method under Investigation

The approach discussed in this paper encompasses a *top-down, black-box testing/verification approach*, intending to provide an *additional level of trustworthiness*. It shall not replace but complement any bottom-up initiative like [**EASA-CoDANN**] e. g., in order to increase trust and — in effect — social acceptance of AI systems in aviation.

The approach may be compared to the *Turing Test* [**ComputingMachineryAndIntelligence**]. Originally called the Imitation Game, the Turing Test is a test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human, by a human evaluator.

In addition, the method contains elements of the *performance-based environment* promoted by [**EASA-PBE**]. In contrast to prescriptive requirements specifying required methods of compliance, performance-based regulations focus on the desired, measurable outcome.

Introducing Trustworthiness Anchors

Trustworthiness lacks a generic, mathematical definition that can be used to test a system. Trust seems to be more a subjective impression that builds up over time, depending on experience and a minimized divergence between expectation or hope and the system's output. As a corollary, a diverse compilation of trustworthiness anchors seems to be required to rationalize trust. Applied to aviation systems, this compilation e. g. includes software certification, hardware certification, systems certification, formal methods, FMEA[7], statistical evidence, training and checking (and licencing) of humans, amongst many others.

No formal evidence has been found on whether the trust is a transitive relationship. If Alice (a) trusts Bob (b) and Bob trusts Charlie (c) — does Alice trust Charlie then, or, formally, for the trust relationship T in an arbitrary set X:

$$\forall a, b, c \in X : (aTb \wedge bTc) \Rightarrow aTc?$$

There is subjective evidence, however, that trust is not a Boolean relationship but rather nuanced and that Alice might credit *some* of Charlie's trust. A certain transitivity factor might, whereas not directly quantifiable, be attributed to trust. Even if binary transitivity is not fully given, combining several diverse and accepted trust anchors can eventually increase the overall level of trust in a system.

For human pilots, trust is built up during the training and checking, under the existing regulatory framework, inter alia, [**ICAO-01**], and [**EASA-FCL**], with well-defined training syllabi, training and checking items and pass standards for check flights. — In analogy to Tombstone Diagrams used in compiler construction and bootstrapping [**CompilerGenerator**], figure 1 summarizes an extract of the pilot training and checking scenarios, simplified from [**EASA-FCL**]:

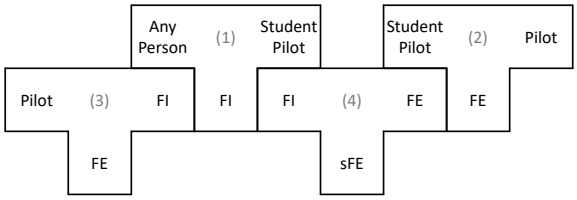


Fig. 1: Trustworthiness Propagation in Manned Aviation

(1) The promotion of any person to a student pilot is done by a flight instructor (FI).

(2) A student pilot may become a licenced pilot through a flight examiner (FE).

(3) A pilot can become a flight instructor through a flight examiner.

(4) It requires a senior flight examiner (sFE) to promote a flight instructor to a flight examiner.

[7] Failure Mode and Effects Analysis

Anthropomorphism vs. Dehumanization

The obvious approach within the anthropomorphic context would be to train AI software to fly a specific aircraft by human flight instructors and check out the software through a human flight examiner. This would stipulate a transparent interface[8], e. g. via speech recognition, so that the AI pilot may be instructed like a human pilot — a setup comparable to the Turing Test [**ComputingMachineryAndIntelligence**].

Compared to humans, however, AI systems have more extensive training needs; they require more training data [**TrainingData**]. Therefore, an aspiring AI pilot will need more training hours than human pilot aspirants, leading to the demand for automated training. In the anthropomorphic context, this demand for a so-to-say *dehumanizing element* rises widdershins: A *flight instructor/flight examiner* software component. This component would guide the AI pilot's reinforcement learning process and evaluate its performance automatically. It would highlight areas of improvement and flag tendencies in the wrong direction. Should the AI performance degrade and/or safe flight envelope parameters be exceeded, it would trigger adequate contingency measures — preferably in a simulated environment. Eventually, the flight instructor/flight examiner modules could confirm the AI student pilot having reached the required standards to pilot an aircraft in the framework of continuous training and checking. Once in operation, it continues to monitor the AI pilot's performance and its experience-based evolution perpetually.

The experience acquired during such monitoring could finally be fed back to the continuing training and checking — as well in analogy to manned aviation, where the regulatory framework mandates that operational experiences are fed back into the training design.

The noteworthy side-effect: In the context of an anthropomorphic approach, where machines shall, in short, be tested like humans (anthropomorphism), the call for automated training stipulates an antidromic element, the function of a flight instructor/flight examiner taken over by a machine (dehumanizing).

Figure 2 superimposes this dehumanizing element to AI software:

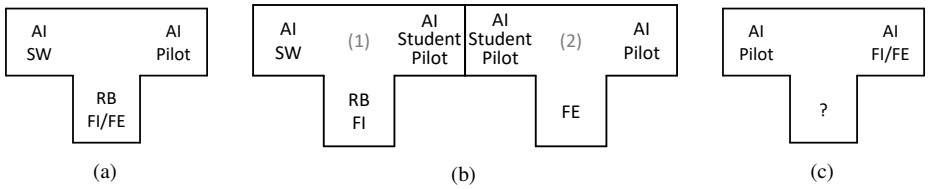


Fig. 2: Trustworthiness Anchors with AI Elements

[8] An interface that allows the connection and operation of a system with another, without modification of system characteristics or operational procedures on either side of the interface.

(a) The flight instructor/flight examiner is supposed to be a rule-based (RB), conventionally certifiable piece of software in order to have an accepted trust anchor — an AI-based flight instructor/flight examiner would a priori miss such an anchor. Given that, an AI software system could then be trained and checked to become an AI pilot.

(b) As an intermediate step, a rule-based flight instructor module could train AI systems (or even aspiring human pilots), and a human flight examiner could check them out. — However, the passed check flight must not be the only trust anchor to 'licencing', as the training phase is explicitly a part of the whole process: In manned aviation, an ATO[9] must confirm that a candidate has satisfactorily completed the required training before a flight examiner will take a skill check. Therefore, it is essential to note that this approach is not just mere check-driven trust propagation.

(c) A further extension would be to train and check an AI pilot to become an AI flight instructor/flight examiner with a yet-to-be-defined trust anchor.

The Objectives of the Investigation

Concluding the preceding section, verifying the suitability of the anthropomorphic approach boils down to implementing and validating a rule-based flight instructor/flight examiner module (RB FI/FE). This module shall be embedded in an environment supporting the test case scenarios listed in table 1:

Instructor / Supervisor	Test Cases			
FI (a)	HUP (1)	RBP (2)	AIP (3)	
FE (b)	HUP (1)	RBP (2)	AIP (3)	FDR (4)

Tab. 1: Test Case Scenarios

(a) FI: Training according to the items for licence training.

(b) FE: Assessing the pilot's performance, with regard to the pass standards of a skill test.

(1) HUP: Humans, both aspiring pilots without a licence and licenced pilots, both with and without instructor and/or examiner privileges.

(2) RBP: A rule-based pilot to generate test data.

(3) AIP: The (future) AI-based pilot.

(4) FDR: Flight data feed from an FDR[10] from actual flight missions (refer to section *Test Data* below). As this data is static, it may not be used in an instructional scenario.

[9] Approved Training Organisation
[10] Flight Data Recorder

Given that, testing will be possible even if the final test candidate, the future AI pilot, the black box in the fictitious scenario from chapter 1, is not yet available.

The additional constraint for a transparent interface will be fulfilled by the FI/FE module able for speech generation, so that humans can be trained and checked interchangeably with the rule-based pilot. Furthermore, a rule-based pilot supporting speech recognition would allow for human flight instructors/flight examiners to be involved without having to get accustomed to the interface.

The Research Setup for the Investigation

The research setup comprises the required components to evaluate and validate the anthropomorphic approach, supporting the objectives from the previous section. It encompasses a central flight deck module (FD), an abstraction layer for the X-Plane desktop-based COTS[11] flight simulator. This module makes the flight situation data available and accepts inputs to control the simulated aircraft. Apart from the X-Plane support, future extensions could include connectivity to time-lapsed flight simulation modules for accelerated training, or interfaces to other flight simulation software or devices — while maintaining the layout of the FD abstraction layer.

The architecture of the research setup (figure 3) shows the three piloting options:

1. A human pilot (HUP), interacting with controls and instruments of the flight deck layer.
2. A rule-based pilot (RBP), programmable to fly a particular pattern.
3. An AI pilot (AIP), potentially available in the future.

Three instructional and/or supervising elements are connected to the pilot:

1. The flight instructor (FI) and
2. the flight examiner (FE) modules to be developed.
3. An aeromedical examiner (AME) in analogy to manned aviation, yet to be defined. The AME could, inter alia, implement a watchdog, listening for 'heartbeats' of the pilot in charge.

The flight instructor/flight examiner modules will implement an assessment and evaluation of the time series of flight parameters provided by the flight data recorder (FDR) within the flight deck layer (FD). They monitor the pilot's performance and give feedback/rewards to an AI pilot's possible deep reinforcement learning system. — The flight instructor/flight examiner must recognize if the autopilot is engaged or if the pilot is flying manually.

[11] Commercial Off-The-Shelf

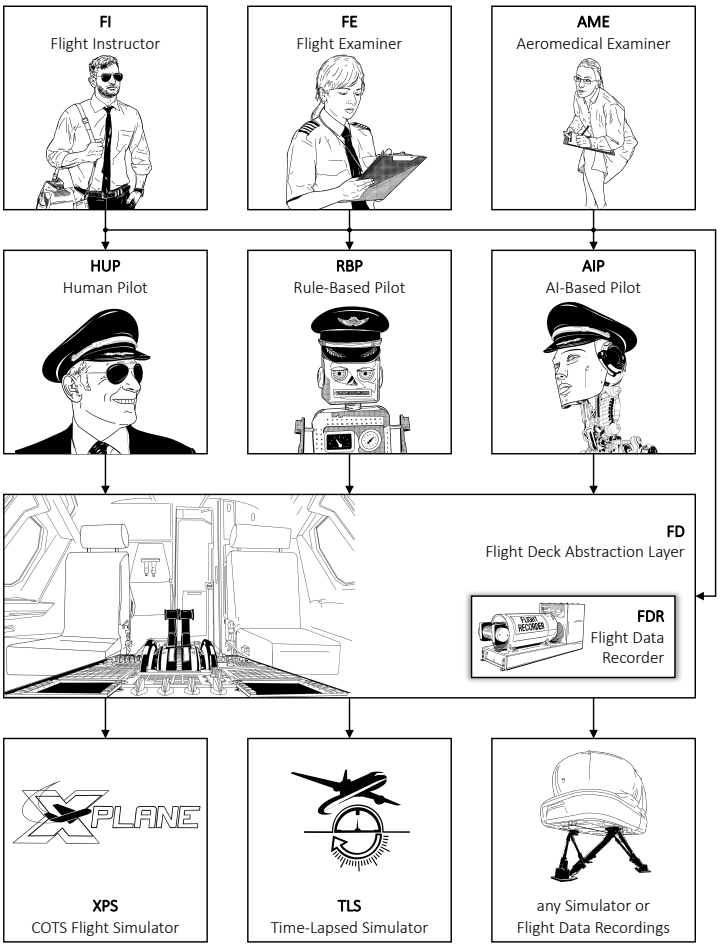


Fig. 3: Architecture of the Research Setup

Thitherto, the modules depicted in figure 4 are implemented:

FD features GUI[12] elements like the PFD[13] as well as an autopilot, consisting of a set of PID[14] controllers, each with low-pass filters both for the setpoint value and for the controlled variable. The integrated flight data recorder provides a set of external views on the inner data for tuning the PID parameters, plots of flight parameters, and a map display.

[12] Graphical User Interface
[13] Primary Flight Display
[14] Proportional–Integral–Derivative

HUP routes joystick inputs and other interactions with control elements (flaps, gear, brakes) to the flight deck.

RBP incorporates some aircraft-specific constants (currently for Cessna 172 Skyhawk and Cirrus Vision Jet SF50 aircraft types), navigational computation functions and is prepared to accept commands via speech recognition, in an attempt to provide a transparent interface to a human flight instructor/flight examiner, comparable to the 'interface' of a human student pilot.

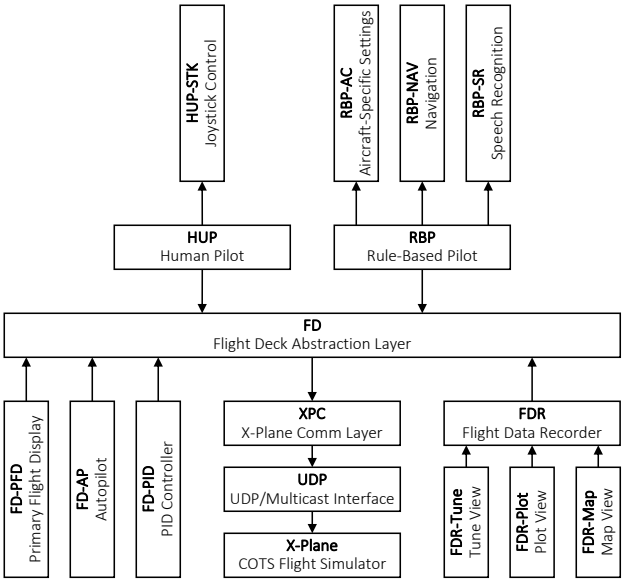


Fig. 4: Modules of the Research Setup

The GUI of the research setup is shown in figure 6.

Test Data

The trustworthiness propagation in manned aviation (figure 1) relies on a variety of operational scenarios encountered during the training and checking phases. The more extensive training needs of AI systems as mentioned in section *Anthropomorphism vs. Dehumanization* further emphasize this requirement.

Both the human pilot and the rule-based pilot options will generate test data for the development and evaluation of the flight instructor/flight examiner modules. Complementarily, instead of using a flight simulator and any of these two piloting options, flight data from actual flight missions could be fed into the system (figure 5).

(a) (b)

Fig. 5: (a) Regular Case with a Pilot and a Flight Simulator. — (b) Flight Data Feed.

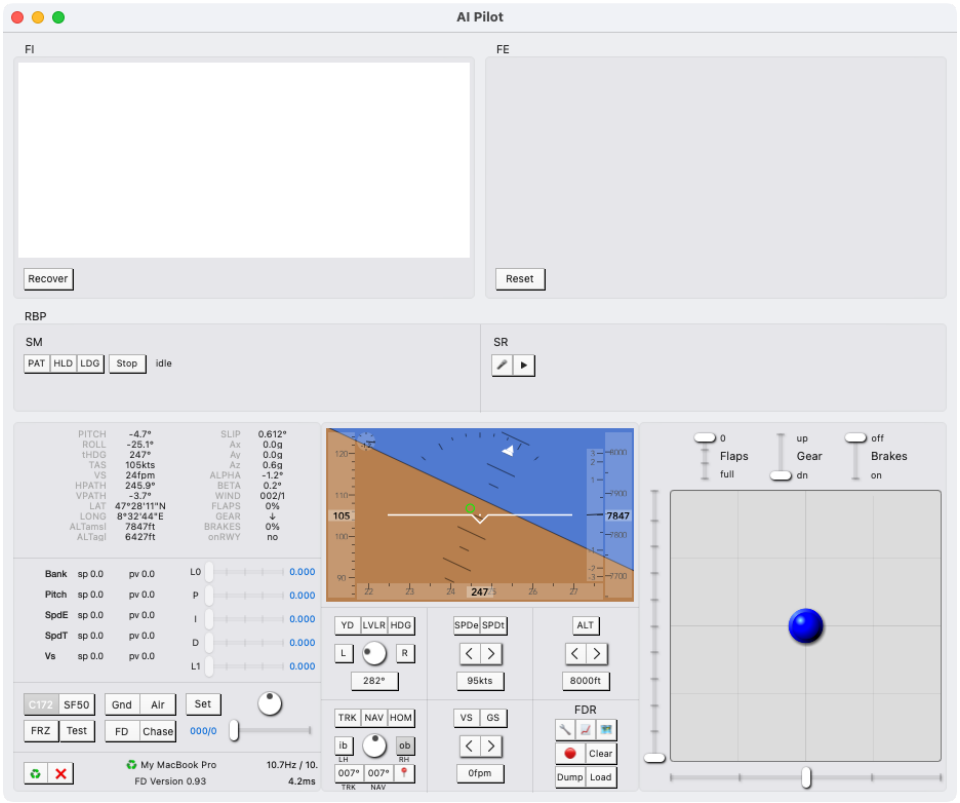


Fig. 6: Graphical User Interface of the Research Setup

4 Intermediate Results and Next Steps

This paper concludes the conceptual description of the anthropomorphic approach and sets the ground for the upcoming detailed investigation and evaluation of the concept. The preparation of the research setup led to a development environment supporting these next steps and revealed several findings that will co-determine the further proceeding.

Findings

The following points are noteworthy in particular:

Anthropomorphism and dehumanization Having consequently applied the anthropomorphic approach, the most surprising insight is the unexpected necessity for a dehumanized element, the rule-based flight instructor/flight examiner modules certifiable under the current regulations, stipulated both by the call for automated training and by the requirement for a trustworthiness anchor.

AI flight instructor/flight examiner An instructional situation distinguishes by many unforeseen events, unanticipated actions of the student pilot and uncertainties. Following the arguments listed in chapter 1, this seems to be an ideal starting scenario for AI — the implementation of the flight instructor/flight examiner modules as AI-based components would be a tempting idea. However, as the aim is to produce an additional level of trustworthiness, a flight instructor/flight examiner lacking trustworthiness does not seem to be an opportune starting point.

Research setup The available research setup with the integration of the X-Plane COTS flight simulator and the two piloting options HUP and RBP seems to be an adequate playground for further investigation and to start the implementation of the flight instructor/flight examiner modules in focus.

Test data The already implemented rule-based pilot is able to produce some first-hand test flight data, supporting the design and development of the flight instructor/flight examiner modules.

Validation The anthropomorphic context reveals an additional opportunity for the validation phase. At the present date, no AI pilots are yet available that could be engaged to validate the flight instructor/flight examiner modules. But these modules can just as well be validated with humans — flight instructors/flight examiners themselves, licenced pilots, student pilots or persons without any previous flying experience. Consequently, the interface to the pilot shall include synthetic voice generation to convey the flight instructor's orders.

Potential overfitting issues A machine learning model that has become too attuned to training data may exhibit overfitting issues [**Overfitting**]. A certain share of noise in the training data e. g. could become too determinant, impairing the system's ability to generalize. As a consequence, the system performs perfectly correct on the training data but will be much less accurate on new data. In the case of a deep reinforcement learning system in training to fly an aircraft, the automated flight instructor/flight examiner will have to ensure that the training scenario is not static but varies within the training epochs. One approach to tackle this issue could be changing environmental conditions (e. g. wind, turbulences, temperature/air density) and diverse flight tasks on different altitudes, locations, and aircraft types — in analogy to manned aviation, where the continuing training and checking of pilots encompasses ever-changing scenarios.

Computing power Currently running on a desktop computer, the available computing power restricts sophisticated real-time time series evaluations and their graphical representation. Consequently, the X-Plane flight simulator has been outsourced to another desktop computer connected via the local network.

Next Steps

The forthcoming phase of the investigation will include the conceptual design and the implementation of the rule-based flight instructor/flight examiner modules, followed by a respective validation.

Flight instructor/flight examiner The current focus is on these modules: Evaluation of the time series of flight parameters provided by HUP and RBP. These parameters shall be assessed, inter alia, regarding the stability of flight (e. g. by evaluating the angle of attack), the adequacy of the control inputs (e. g. with regard to oscillations), the reaction to external events and to orders from the flight instructor, the successful completion of the flight mission, and other aspects.

Validation The flight instructor/flight examiner modules shall be developed and tested by licenced pilots and licenced flight instructors/flight examiners.

Risks Assessment and mitigation of the risks associated with the anthropomorphic approach in general and the overfitting issues in particular.

Benefits Investigation of the anticipated key advantages identified in [**NASA**], like performance focus, reduced costs, reduced stagnation, and reduced manufacturer liability.

Special attention has to be payed to the amount and variety of operational scenarios, the test data mentioned in section *Test Data* in chapter *The Anthropomorphic Method under Investigation*, and considering the *potential overfitting issues* listed above.

5 Summary and Conclusions

In summary, the concept of the anthropomorphic approach depicts as follows:

Goal The goal of the ongoing research is a deepened investigation of whether an *anthropomorphic apporach* could prove to be *one element to verify or to leverage the trustworthiness of AI systems* in cockpit applications. The intention is by no means to replace bottom-up approaches currently under investigation and/or elaboration, but to establish it as a complementary, independent element of assurance to *increase the level of trustworthiness*. — Trustworthiness is considered *the* key element for any future certification/licencing efforts, for insurance covers, and for social acceptance, which in effect again could trigger regulatory initiatives for the development of a corresponding regulatory framework.

Approach to the goal Relying on a well-established and well-proven regulatory framework that is continuously enhanced and refined primarily based on findings out of investigations of occurrences, incidents and accidents, the general idea is to apply the flight crew licencing framework to AI systems. In analogy to the Turing Test [**ComputingMachineryAndIntelligence**] as well as incorporating the general idea behind the performance-based environment [**EASA-PBE**], a machine intended to execute a particular function, i. e. to fly an aeroplane, shall be tested the very same way as a human intended to execute the very same function. This is the so-called anthropomorphic approach. — Furthermore, when compared to humans, AI systems have elevated training needs. It should be possible to conduct the training and checking processes autonomously, or at least to the possible extent. Therefore, it shall be investigated whether the roles of a flight instructor and flight examiner can be partially built up in software. This would then be the so-called dehumanizing but central element within the investigation. — It has to be noted that the term trustworthiness lacks a formal and applicable definition.

Incremental implementation It is paramount to envisage an incremental introduction of such disruptive, game-changing new methods, especially in safety-critical environments like aviation. The rather out-of-the-box approach further emphasizes the importance of gradual progress. For example, an AI software acting in shadow mode could lead to the first statistical evidence. The next phase with AI software at the controls should include human supervision and take place in a simulated environment as long as possible.

Non-goals It is *not* the goal to contribute to the actual development of an AI-based piloting system — this is well beyond the scope. The idea begins in the fictitious scenario where a black-box system, claiming to have piloting capabilities, shall be tested and attributed to a certain level of trustworthiness. — The envisaged flight instructor/examiner will probably not cover the entire airline transport pilot training and checking range. In a first step, specific private pilot training and checking elements

shall be incorporated and examined for the suitability of such an anthropomorphic approach. — Furthermore, the envisaged validation of the anthropomorphic approach assumes a transparent interface to the different piloting options (refer to chapter 3), i. e. the same way of communication between the flight instructor/flight examiner and the pilot at the controls. No distinction is made whether the AI-based pilot will rely on additional sensors like e. g. LIDAR[15] or LRF[16], or AI subsystems for image recognition, or not.

Relevance The relevance of such independent software assurance systems is not limited to the cockpit environment. Similar applications can be found e. g. in air traffic control or flight dispatch, or generally, in most situations involving licenced human staff, not limited to aviation.

Several questions arose during the elaboration of the anthropomorphic approach. Interestingly, it was possible to project most of these uncertainties into the field of manned aviation, where similar challenges are found and where answers transferable to AI pilots are available, further emphasizing the potential of the general idea.

6 Acknowledgement

The first author would like to thank the supervisor for all the helpful discussions as well as for the constructive comments and ideas, and the University of Stuttgart for the opportunity to pursue this exciting approach.

[15] Light Detection And Ranging
[16] Laser Range Finder