

ARTIFICIAL INTELLIGENCE AND THE JUST CULTURE PRINCIPLE



The European Commission has proposed a legal framework on AI. In light of some of the risks and opportunities, **Federico Franchina** highlights the importance of reconciling the use of AI with Just Culture, ensuring clarity on decision-making, standards, training, and liability.

KEY POINTS:

- The European Commission has proposed harmonised rules on artificial intelligence (AI) to address its potential benefits and competitive advantages.
- The proposal highlights the need for transparency, resilience, and human oversight in the design and development of high-risk AI systems, particularly in safety-critical environments.
- The use of AI in aviation raises questions about liability and decision-making, requiring a paradigm shift to share responsibility between humans and machines, avoiding placing undue burden solely on human operators.
- The introduction of AI challenges traditional tests of intent and causation, and a sliding scale system for liability is suggested to adapt to the unique characteristics of AI and maintain a fair approach.
- To uphold the Just Culture principle, it is necessary to consider human behaviour, training, and standards in the context of human-machine relations, ensuring a balanced approach between human oversight and AI capabilities.

In April 2021, the European Commission laid out a proposal for harmonised rules on artificial intelligence (AI). The draft, yet to be voted on by the European Parliament, aims to address this new technology, which, according to the proposal itself, can “support socially and environmentally beneficial outcomes and provide key competitive advantages to companies and the European economy.”

AI will be able to achieve these goals by improving prediction, optimising operations and resource allocation, and personalising services.

According to the proposal, AI is defined as software that generates outputs for a given set of human-defined objectives. These outputs can include content, predictions, recommendations, or decisions that have the ability to influence the environments with which they interact.

A Risk-Based Approach

The proposal establishes rules for AI based on a risk-based approach, with specific attention given to systems that serve as safety components of products. The aim is to integrate these rules into the existing sectoral safety legislation to ensure consistency.

Aviation is partially seen as a high-risk environment that is indirectly affected by this EU proposal when AI systems are used or are a part of a “safety component” that fulfils a safety function for a product. The failure or malfunctioning of such systems can endanger the health and safety of individuals or property.

Based on these assumptions, any introduction of AI in the field of aviation should follow some principles laid down by the same proposal. Some of these are of paramount importance for safety.

First, the proposal states that high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system’s output and use it appropriately.

It also states that high-risk AI systems shall be resilient regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems.

Moreover, it is stated in the proposal that the design and development of AI shall also be made through the lens of human-machine interface tools, as well as the oversight by “natural persons” during its use. Within this provision, human oversight is tasked with the specific goal preventing or minimising the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse.

“Rules have been designed with the understanding that operations and activities are performed by humans. However, the proposal on AI regulation seems to shift from a human-centred approach to a human oversight approach.”

The Human Role

Along with this, it is required by human oversight to fully understand the capacities and limitations of the AI system and be able to duly monitor its operation in order to detect and address any signs of anomalies and dysfunctions.

For the purposes of the regulatory draft, to paraphrase, measures should “enable the individuals to whom human oversight is assigned to do the following, as appropriate to the circumstances:

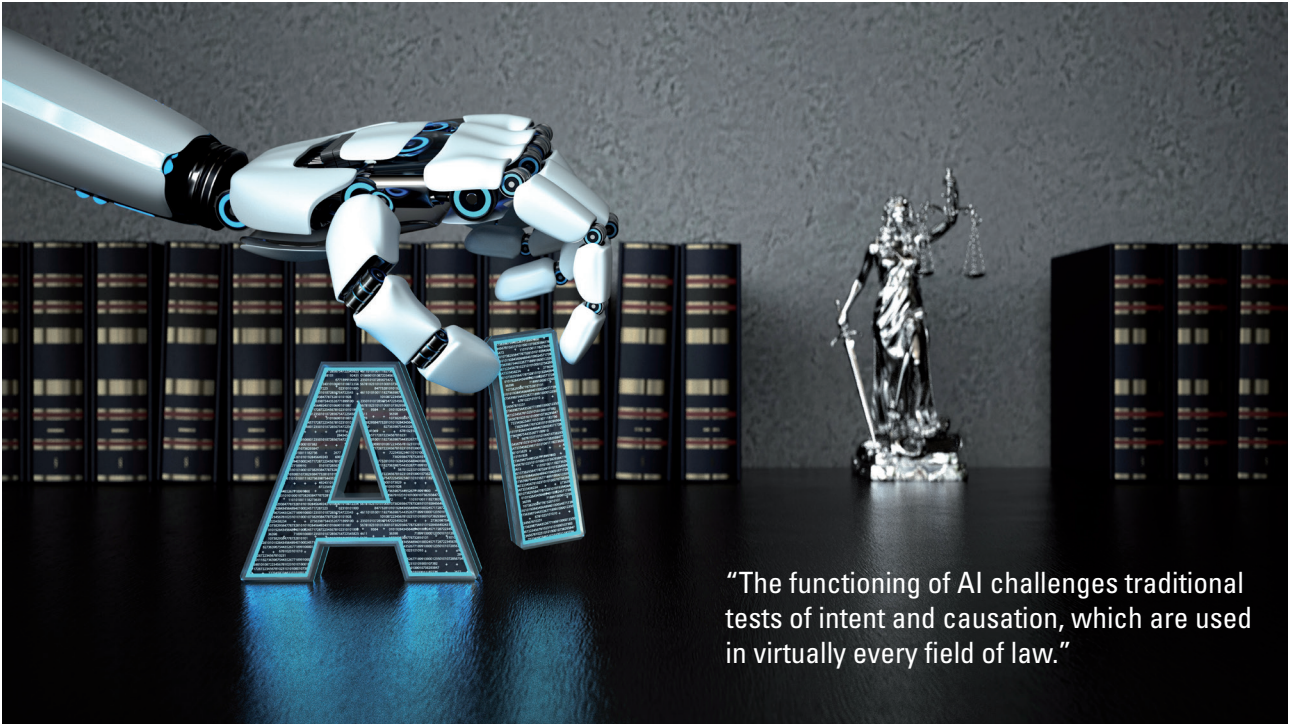
- (a) be aware of and sufficiently understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, so that signs of anomalies, dysfunctions and unexpected performance can be detected and addressed as soon as possible;
- (b) remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (‘automation bias’), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
- (c) be able to correctly interpret the high-risk AI system’s output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;
- (d) be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;
- (e) be able to intervene on the operation of the high-risk AI system or interrupt, the system through a “stop” button or a similar procedure that allows the system to come to a halt in a safe state, except if the human interference increases the risks or would negatively impact the performance in consideration of generally acknowledged state-of-the-art.”

(On 14 June 2023 the European Parliament approved its position (a) and (e) above, which were originally phrased differently.)

Human Oversight and Human Liability

Institutional documents and papers on the topic of aviation AI share a common element: a ‘human-centred approach’. These include the ICAO (2019) working paper on artificial intelligence and digitalisation in aviation, the European Aviation/ATM AI High Level Group FLY AI

report (EUROCONTROL, 2020), the EASA Artificial Intelligence Roadmap (2020), and the SESAR European ATM Masterplan (SESAR Joint Undertaking, 2020). Rules have been designed with the understanding that operations and activities are performed by humans.



However, the proposal on AI regulation seems to shift from a human-centred approach to a human oversight approach. This raises different questions.

The introduction of AI in the aviation environment could involve several actors, including physical persons, air carriers, air navigation service providers (ANSPs), states, and manufacturers. Existing regulations, such as ICAO Annex 11 (also Doc 9426 and Doc 4444) and the EU SES package (Reg. 1139/2018), and certification and personnel licensing regulations, already consider the perspective of air traffic controllers (ATCOs).

From the perspective of liability, the use of AI in aviation (as well as in other sectors) involves various types of liabilities, including criminal, civil (contractual and extra-contractual), state/administrative, product, organisational, and vicarious liabilities.

The 'Black Box Problem'

The proposed framework and definition of AI, as well as the responsibilities placed on humans (in terms of oversight and 'duty of care'), should be understood in the context of AI's functioning through neural networks that break problems down into millions or even billions of pieces and solve them step by step in a linear fashion. We do not know exactly what the algorithm is doing or what methods it is using. This has been referred to as the 'black box problem' because AI can seem like a black box with no visibility into its inner workings.

"AI can seem like a black box with no visibility into its inner workings."

The human decides on the inputs and objectives, and allows the AI to work (in a 'black box' manner), but must oversee its functioning and interrupt the process if necessary. However, ethical questions arise in retrospect: on what basis did the human decide to interrupt the process? Does AI establish a standard or benchmark for evaluating human actions? Two situations can occur:

1. AI suggests a correct action, but the ATCO does not follow the suggestion, leading to an occurrence:
 - Is the ATCO liable for breaching the duty of professional negligence?
 - On what basis does AI suggest a 'correct action'? Does it follow a different standard or benchmark than the one followed by the ATCO?
 - Does the ATCO have a duty to follow AI's suggestions?
 - Can AI suggestions be used as evidence?
2. AI suggests a wrong action, and the ATCO follows the suggestion, leading to an occurrence:
 - Is the ATCO liable for breaching the duty of professional negligence?
 - Does the ATCO have an appropriate mental model about how AI will function?

Human-Machine Interaction

To reconcile this framework and address these questions while upholding the Just Culture principle, it is important to look at human behaviour and training in the context of human-machine relations. We need to clarify who will make decisions,

when and why they will be made, and based on which standards and training.

This is especially important in situations where there is a hybrid mode with significant interactions between humans and machines. The aim should be to reduce overconfidence in the machine and other unintended consequences.


As automation is introduced and tasks and responsibilities are increasingly delegated to technology, liability for damages is expected to shift from human operators to the organisations responsible for designing, developing, deploying, integrating, and maintaining the technology. However, the functioning of AI challenges traditional tests of intent and causation, which are used in virtually every field of law. These kinds of tests, which assess what is foreseeable and the basis for decisions, could be ineffective when applied to black-box AI.

The solution to this problem should not be strict liability or a regulatory framework with specific transparency standards for AI. Instead, a flexible system could lead to a more suitable approach as it adapts the current regime of causation and intent tests. In this sense, it impacts the requirements for liability for those situations when AI operates autonomously or lacks transparency. On the other hand, it maintains traditional intent and causation tests when humans supervise AI or when AI is transparent.

Just Culture and AI

So far, our approach to machines has been guided by a simple principle: we know the inputs, we understand how they work, and we know the expected outputs. This has led to a focus on human considerations regarding mistakes, negligence, and faults.

With the introduction of AI, we may have to deal with machines that can make mistakes. It would be unfair, wrong, and even unethical to place all the responsibility solely on humans and their oversight duty.

This paradigm shift is important not only in retrospect, ex post, when allocating liability or conducting safety evaluations, but also in advance, ex ante, when prevention and precautionary measures need to be applied. This approach contributes to reinforcing the 'Just Culture' principle, which should not be amended but should consider the involvement of AI as a player in the playbook. 

“With the introduction of AI, we may have to deal with machines that can make mistakes. It would be unfair, wrong, and even unethical to place all the responsibility solely on humans and their oversight duty.”

References

EASA (2023, May 10). EASA artificial intelligence roadmap 2.0: A human-centric approach to AI in aviation. <https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-roadmap-20>

EUROCONTROL (2020, March 5). The FLY AI report: Demystifying and accelerating AI in aviation/ATM. <https://www.eurocontrol.int/sites/default/files/2020-03/eurocontrol-fly-ai-report-032020.pdf>

European Commission (2021, April 21). Proposal for a regulation laying down harmonised rules on artificial intelligence. <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence>

ICAO (2019, January 8). Working paper: Artificial intelligence and digitalisation in aviation, 2019. https://www.icao.int/Meetings/a40/Documents/WP/wp_268_en.pdf

SESAR Joint Undertaking (2020, December 17). European ATM master plan 2020. <https://www.sesarju.eu/masterplan2020>



Federico Franchina is Professor of Maritime, Air and Transport Law at Università degli Studi di Messina. He was formerly a Legal Expert at EUROCONTROL and Expert at the Superior Council of Public Works Expert at the Superior Council of Works at the Ministry of Infrastructures and Transport in Italy.