

GenAI and Knowledge Generation in ATM

Current Research and Anticipated Prospects

Gurtner G, Cook A, Tanner G

Centre for Air Traffic Management Research

School of Architecture and Cities, University of Westminster

London, United Kingdom

g.gurtner/cookaj/tannerg@westminster.ac.uk

Abstract—This paper explores the role of generative AI for knowledge generation in air traffic management. It includes a high-level SWOT analysis, a review of the general state of the art and within SESAR. The paper presents the key results of a case study: a trend analysis to build a future concepts roadmap for ATM. It discusses the role of agentic AI and retrieval-augmented generation, both as a follow-up to the roadmap development and to illustrate broader use-case potential in SESAR’s transversal industrial research.

Keywords—Generative AI; GenAI; agentic AI; RAG; SWOT analysis; trend analysis; concepts roadmap

I. INTRODUCTION

Generative AI (GenAI) is gaining increased traction in many domains. We first perform a high-level SWOT analysis of such use, highlighting key positive trends and the unique contributions that GenAI brings to research, plus some areas whereby caution and protection are needed. After reviewing major work in AI being undertaken within SESAR, we focus specifically on a case study: a trend analysis to build a future concepts roadmap by the SESAR KTN, Engage 2. Work already completed and key results are presented. We then discuss the role of agentic AI and retrieval-augmented generation (RAG), proposed as a follow-up to the concepts roadmap development and, in addition, as an illustration of a broader use-case in SESAR’s transversal industrial research.

II. STATE OF THE ART

A. Generative AI at large

Generative AI (GenAI) is a broad, prevalent type of AI which, by definition, generates new content (e.g. text or images) given a user request and *a priori* work. A large language model (LLM) is a specific type of machine learning model, focused on predicting natural (human) language. LLMs are pre-trained on massive datasets, which are used by GenAI systems to generate outputs. Other types of models, e.g. for images, videos, etc., can be trained and are part of the GenAI class. (AI is thus much broader than generative AI, in the latter’s now-normalised sense of creating original content.) ChatGPT, Gemini and Claude, as ‘chat’ AI examples, are all based on LLMs. These LLMs are enclosed in special interfaces, allowing them to answer user/agent prompts, but also able to access tools such as internet search engines and retrieval-augmented generation (RAG) tools to augment and

enhance answers with up-to-date data, instead of relying on LLM’s static, pre-training data.

AI’s of different forms have been gaining ground over the last 20 years. The popularity of GenAI grew particularly rapidly from November 2022, when OpenAI released its new version of a chat AI called ‘ChatGPT 3.5’. This popularity pushed other companies to release their own versions of chat AI’s, in particular Google and Amazon. Notwithstanding several concerns raised since then, which we summarise below, this technology has been flagged as the most promising path to general artificial intelligence¹ and has found many applications. Moreover, in just three years, several new features have been added to such tools, including image recognition and generation, text and speech recognition, information retrieval, chain of thought, etc. Whilst not an exhaustive review, we next present a brief SWOT analysis of GenAI.

1) *Strengths - power*: A key issue with many challenges that humanity faces is their sheer complexity. Examples include multifactorial diseases such as cancer, and complex socio-technical systems such as transportation and mobility of the citizen. In the former case, rather than being targeted through intervention of a specific enzymic deficiency or hormonal excess, for example, cancer is the result of a complex dependency between nature and nurture, such as genetic predisposition, lifestyle, environment, infection, and chronic inflammation. A key strength, which GenAI is expected by many to bring to bear in oncology, as well as many other fields of medicine (e.g. antibiotic resistance [1]), is the ability to assimilate vast amounts of information and identify new potential links between components of systems, and properties thereof, to propose not only new solutions, but even new *types* of solution. Analogous advances in policy and technical interventions may be expected in transportation, and the role that ATM plays therein. The latter is a complex socio-technical system, set within the wider context of door-to-door citizen mobility, influenced by multifactorial determinants such as infrastructure (cf. ‘nature’), markets, regulation, policy and passenger utilities (cf. ‘nurture’).

With LLMs already very powerful at communicating in natural languages (and other types of languages, such as

¹An AI that can perform any cognitive task at the same level as a human. Some argue this is hypothetical; others that current AI systems already demonstrate some such properties.



programming), and particularly adept at synthesising text, extracting and connecting particular strands of information for users, they are useful for many academic applications, e.g. when researchers need to summarise information based on previously published academic papers, such as when performing a literature review². (See [2] for a comparison of ChatGPT and Microsoft Bing for medical informatics, for example.)

2) *Weaknesses - reliability*: These tools are not without issues, however. These mostly focus around concerns associated with reliability, some of which are being exploited by malevolent parties (see Section II-A4). They include, but are not limited to: bias in the datasets involved in training; hallucination; over-focusing; omission; and, consistency. Generative AI service companies invest significant effort to mitigate bias, and social scoring is now banned under EU law [3]. GenAI use is indeed now so widespread that literature reviews often focus on specialised fields in which issues may occur, various medical (diagnostic) areas being a notable example. Although hallucinations are generally becoming less of a problem, this is still a prevalent topic in terms of avoidance models (e.g. [4]) and detection tools (e.g. [5]).

3) *Opportunities - scope*: Frequently reported in the media, new applications for GenAI in particular, and AI in general, would require a review paper in its own right. These reach into many domains of daily life: aviation and education, two fields of importance for the readership of this paper, being no exception. There are many reviews³ of the types of jobs most likely (e.g., entry-level, white-collar) and least likely (e.g., blue-collar) to be impacted by GenAI, and those already being thus affected, including growth in AI-augmented roles. The ethical use of AI and safe integration of the human are extensively, and increasingly, researched and documented in many socio-technical contexts such as ATM. These include discussions of human-AI hybrid systems to “augment human ingenuity rather than replace it” although implementation “remains limited due to the high stakes involved and the challenges presented by AI’s unpredictable behavior and limitations in complex reasoning tasks” [6]. Even during the production of this section of the paper, further examples of the use of AI in aviation were published. Birmingham Airport announced⁴ the launch of a personal assistant, guiding passengers through their journey, accessible on smartphone devices or via digital help points around the terminal. Schiphol presented⁵ its vision for fully autonomous airside operations by 2050, in collaboration with KLM. Academia faces a persistent endeavour to keep abreast of emerging AI technologies, maintaining policy transparent and up to date, and striking the right balance between supporting appropriate use of new tools and skills in the learning environment, whilst mitigating abuse (such as

students passing-off work produced by GenAI as their own). Prominent amongst such enabling tools is prompt engineering [7], as used extensively in the case study of this paper (Section III).

4) *Threats - abuse*: Also in an equally rapidly evolving and extensive landscape, are the threats enabled by (generative) AI. Further to the theme of academic utility and more obvious passing-off, is the subtle threat of how AI assistants might accelerate skill decay among experts, and hinder skill acquisition for learners [8]. Rather greater societal damage has recently been demonstrated by disruption across a number of European airports, including Heathrow, Brussels, Dublin and Berlin, caused by cyber-attacks⁶. Incidents such as these, as with many instances of GenAI use, demonstrate the wider principle of the technology being a force for good or bad. As for the latter, it can, for example, reduce the effort needed to launch sophisticated and targeted cyber-attacks, partly because GenAI can mimic human language and behaviour so well, e.g. interacting with (other embedded GenAI) systems, and because LLMs are efficient at writing code (which may be directed at illegal objectives). Conversely, GenAI may be used by companies to improve their ability to detect and manage such threats. Whilst GenAI service providers are active in installing ‘guardrails’ to prevent misuse of their systems (the reader may investigate these by entering prompts of questionable purpose), malefactors are often skilled at circumventing these protection mechanisms through so-called ‘jailbreaks’ to produce attacks. Also of concern is the prevalence of ‘LLM grooming’, whereby malefactors intentionally create bias and/or deception in training data by generating large-scale contamination thereof, thus “transforming generative AI systems into powerful vectors of disinformation. Unlike classical data poisoning—which corrupts developer-curated datasets—LLM Grooming operates through the systematic flooding of publicly accessible data sources, such as web archives, with fabricated narratives” [9]. Again, we have had space here to give indicative examples only, rather than any form of comprehensive review.

B. Generative AI in the SESAR programme

The European Union Aviation Safety Agency (EASA) produces dynamic guidance on its vision for the deployment of AI in aviation, aiming to serve as a basis for interaction with its stakeholders, and as presented in its latest AI Roadmap 2.0 [10]. This guidance is considered, *inter alia*, in the context of the certification of AI and advanced automation in aviation by the (SESAR) HUCAN project [11]. In parallel, and complementary to this guidance, sits EU legislation. The EU AI Act was published in the Official Journal of the European Union in July 2024, and entered into force in August 2024 [3]. It is the first comprehensive regulation on AI by a major regulator. It assigns applications of AI to three risk categories: unacceptable risk (e.g. banning government-run social scoring); high-risk applications (e.g. scanning CVs

²Good dedicated, commercial tools exist, such as SciSpace <https://scispace.com/> and Semantic Scholar <https://www.semanticscholar.org/>.

³See, e.g.: <https://www.independent.co.uk/news/world/americas/artificial-intelligence-jobs-risk-b2815932.html>

⁴<https://airliners.com/birmingham-airport-announces-new-ai-powered-improvements-to-customer-experience/>

⁵<https://www.futuretravelexperience.com/>

⁶<https://www.bbc.co.uk/news/articles/c62ldxyj431o>



to rank job applicants - subject to legal requirements) and, otherwise; largely left unregulated [12]. The ban of AI systems posing unacceptable risks started to apply in February 2025. GenAI, is not classified as high-risk but has to comply with transparency requirements and EU copyright law [3].

Further exemplifying current SESAR exploratory and industrial research projects deploying AI, whilst eight projects are within the dedicated ‘artificial intelligence for aviation’ flagship, AI is central to many projects within the other flagships [13]. Projects focusing on enhancing ATM through AI-powered digital assistants (DAs) and improved human-machine interaction include CODA, DIALOG, TADA, DARWIN and JARVIS. CODA is addressing controller mental workload by developing a DA to predict traffic and adjust airspace configuration to reduce workload. The ‘Teamwork Assistant’ being developed by DIALOG uses speech recognition and behaviour analysis to infer ATCO intent and manage workload in real time. TADA is using historical controller data and machine learning to develop a DA to support ATCOs in terminal airspace decision-making, such as suggesting AMAN arrival sequences. DARWIN is designing cockpit DAs to support single pilot operations while maintaining safety and reducing crew workload. One of JARVIS’ three AI-based solutions also supports crew and single pilot operations, with two further DAs being developed to support controllers and airport operators. RESPONSE has a focus on pilot cognitive state monitoring and is investigating technology enablers such as the integration of air-to-ground AI-powered DAs.

A cohort of SESAR projects is applying AI to optimise airspace usage, predict congestion, and balance demand and capacity. ASTRA is using AI to predict and resolve tactical air traffic hotspots for more efficient and greener trajectories [14]. HYPERSOLVER is developing a reinforcement learning-based AI system to manage ATC and ATFM holistically [15]. ISLAND uses AI and virtualisation to dynamically configure airspace and balance demand and capacity, whilst FASTNet seeks to integrate airport and network operations using AI to enable airport-to-airport collaborative planning within the network operations plan. KAIROS is applying AI to improve the quality of meteorological information and integrating this into decision-making tools for better flow management. Two projects are using AI-based solutions to help improve airport operations. ORCI uses AI to optimise vectoring instructions for better spacing and efficiency on final approach. TRUSTY aims to enhance remote digital towers with trustworthy AI for specific tasks such as runway monitoring, and decision support [16]. HUCAN is addressing certification and liability challenges of AI in ATM, providing design guidelines and legal analysis [11]. SynthAIR is developing AI-based synthetic data generation tools to support tasks such as chatbot training and sentiment analysis. AI4HyDrop is developing an AI-based framework for managing drone operations in restricted and urban airspace [17]. E-CONTRAIL combines AI and climate science to predict and mitigate contrail formation and aviation-induced cloud.

Furthermore, AI is integral to two Engage 2 catalyst-funded

projects. GenAI models for ATM is using GenAI to annotate NOTAMs with a tag system as a proof of concept. The ChatMET project is implementing an LLM chatbot as a simple natural language solution to access historical meteorological information. Having briefly reviewed some of the projects deploying AI in SESAR, we next present our specific case study of building the ATM concepts roadmap in Engage 2.

III. CASE STUDY - BUILDING A CONCEPTS ROADMAP

A. Background and scope

The goal of the Engage 2 ATM concepts roadmap is to provide an innovative set of concepts beyond the SESAR 3 timeframe with emerging and distinct relevance to ATM. As described below, we have used GenAI to automatically discover trends from articles and white papers. The development of the roadmap takes into consideration information on R&D topics, concepts, future priorities and challenges, existing research in the main SESAR programme, and links with other initiatives beyond ATM. This work builds on the foundation work of the original Engage KTN (2018-2022) and recent work carried out within Engage 2 investigating trends in skills required from the future ATM workforce and their expected impacts [18]. Mindful of the very fast development in GenAI, this technology was used for the first iteration of the trend analysis to explore ATM research literature, as we describe below. The time horizon for the trend analysis was set to cover near- and mid-term perspectives. Since we are interested in trends in aviation, we are invested in building a roadmap for (mostly European) aviation technological deployment, which means trying also to capture approximate dates and interdependencies among trends. The analysis covers trends with a horizon of 30 years. With a longer scanning horizon, more variables and unknowns come into play and factors beyond technological advancements, such as economic cycles, political changes and societal shifts will have increased impact. Kindly note that further details of the methodology and results described in this section on the case study, are expected to be published imminently on the Engage 2 website, as Deliverable 5.1⁷.

B. Methodology

We used GenAI models to extract knowledge and to find patterns within large sets of texts, to find trends in technological developments. Instead of relying on single prompts and high-level queries, several GenAI models were fed sequentially, or in parallel, with the results of previous models. This approach allows us to have a much higher quality for the final results, avoiding common traps like hallucinations, over-focusing and omission, or assigning all output items with the same credibility. Prompt engineering techniques were used at each step, affording the production of answers which were as relevant as possible, and having proper formats for the next sequence. Two providers of ‘small’ and ‘big’ general models were selected: OpenAI and Mistral. OpenAI is the

⁷<https://engagektn.eu/resources/public-reports/>



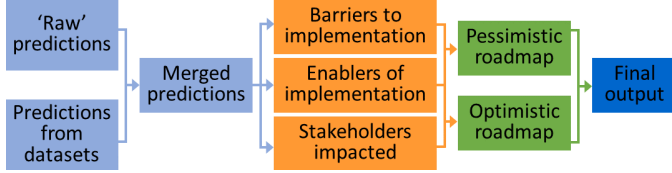


Figure 1. Workflow

provider of the well-known ChatGPT. We used the ChatGPT API directly, in particular, the models GPT-4.1 and GPT-4.1 mini. We used two Mistral models: mistral-small and mistral-medium, two API placeholders corresponding to the best models of these sizes developed by Mistral at any point in time (this API was accessed in May 2025 for the purpose of the work described in this article). We refer to ‘small’ models for GPT-4.1 mini and mistral-small and to ‘big’ models for GPT-4.1 and mistral-medium, which are respectively comparable in their capabilities. Small and medium models are suited to different tasks. Small models are best for performing more straightforward tasks such as generating summaries or simple extraction, whereas big models are good for making more complex decisions, taking into account more context and investigating the meaning of the prompts in greater depth.

Figure 1 shows the roadmap workflow. Using the LangChain framework for developing applications powered by LLMs⁸, at each step an LLM took the output of the previous step as input to produce the next one. LLM predictions were made independently. The ‘raw’ predictions were made without the ATM (research) data in order to capture more ‘obvious’ trends, which might not be captured through the datasets, whereas predictions from datasets (see Section III-C) used the ATM (research) data *and* LLM background knowledge. The merged predictions were then made using a big model, the model being oblivious at this stage to the underlying data, and only having access to the two sets of predictions. The model was prompted to decide on the 20 most important trends.

The next steps ran in parallel, each focusing on an important dimension of the roadmap: barriers, enablers and impacts on stakeholders (including environment and society impacts). For this stage (shaded orange in the figure), the models had access again to the two datasets and were tasked with making informed decisions supported with citations from the sources. The following step (shaded green) comprised priming two independent models to be either ‘optimistic’ or ‘pessimistic’ (see Section III-E1), and to ask them to build roadmaps with corresponding dates of implementation, considering the three previous dimensions. This output was then consolidated through concatenation of the preceding steps. This sequential workflow allowed us to have nuanced views of a complex system and to avoid (typical) hallucinations by reducing the dependency on a single model. Note that we did not need a RAG. Indeed, the information was sufficiently compressed to be included in the context window of the models.

⁸See: <https://python.langchain.com>

C. Data sources

Data obtained for the modelling comprised conference papers, SESAR project deliverables and reference materials. These three groups were treated as ‘trusted’ sources. Earlier iterations had revealed that some of the information provided by the SciSpace repository⁹ was not as relevant as those from the ATM datasets. Data already used by the original Engage KTN or in previous iterations (see [18]) were not re-analysed. In total, 525 documents were used.

1) *Conference papers*: Full papers were sourced from three ATM/air transportation research-focused conferences. A total of 316 conference papers were obtained from the most recent editions of the SESAR Innovation Days (‘SIDs’), the USA/Europe ATM Research and Development Seminar (‘ATM Seminar’) and the International Conference on Research in Air Transportation (‘ICRAT’), as per Table I.

TABLE I. CONFERENCE PAPERS

Conference	Full papers
SIDs 2024	54
ATM Seminar 2021, 2023	54, 41
ICRAT 2020, 2022, 2024	55, 43, 69
Total	316

2) *SESAR project deliverables*: 203 deliverables published by SESAR 2020 and SESAR 3 projects were sourced and selected from CORDIS¹⁰, including concept definitions and requirements. Data management plans (for example) were excluded. SESAR 2020 material included final reporting since these projects had all closed. Only initial material was available for the on-going SESAR 3 projects. By March 2025, exploratory research Call 1 (‘ER1’) projects had published 130 reports, of which 22 were considered potentially relevant. Thus, deliverables for 14 out of the 18 ER1 projects, and 13 of the 32 industrial research Call 1 (‘IR1’) projects were sourced. The materials covered all nine SESAR 3 flagships, plus transversal activities, as per Table II.

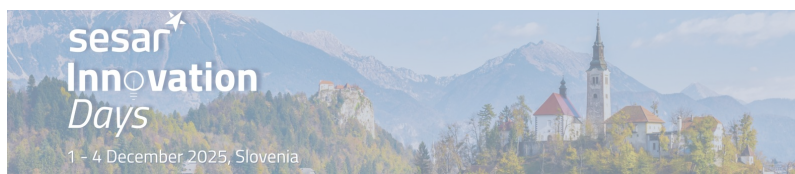
TABLE II. SESAR DELIVERABLES

Programme	Call	Project coverage	Deliverables
SESAR 2020	‘ER4’	41/41	90
	‘IR Wave 3’	5/5	46
SESAR 3	‘ER1’	14/18	22
	‘IR1’	13/32	45
Total		73/96	203

3) *Joint Undertaking reference material*: Six key reference documents and supporting material were sourced from SESAR 3 JU (ATM Master Plan 2025 [19]), Clean Aviation JU (Strategic Research and Innovation Agenda, 2024 [20]) and Shift2Rail/Europe’s Rail JU (Rail Strategic Research and Innovation Agenda, 2020 [21], draft Master Plan, 2022 [22]). Having obtained the material outlined above, the first step was

⁹<https://scispace.com/>

¹⁰<https://cordis.europa.eu/>



to convert the pdfs into more machine-friendly formats, such as ‘Markdown’. This was done using the optical character recognition model from Mistral. This model was able to extract written characters from pdfs, even characters within images, and save all associated figures. Tags were automatically added for tables, sections, etc., so that they could be treated by subsequent models. The second step was to produce summaries of the sources. In the case of conference papers and project deliverables, most of the relevant information was extracted via ‘focused’ summaries, i.e. summaries produced by LLMs which already ‘knew’ what subsequent LLMs would be interested in. In most cases, at least for conference papers, these summaries aligned well with abstracts written by the authors themselves.

D. Results

The model thus produced detailed results inferred from the ATM datasets and the internal knowledge of the LLMs. The results provided explanations about the stakeholders impacted, barriers and enablers, as well as the environmental and social impacts for the 20 most important trends. Examples of the trends included ‘electrification and advanced battery technologies’, ‘global seamless biometric security and identity’ and ‘quantum computing-optimised ATC’. However, an initial attempt of mapping the strategic deployment objectives (SDOs) as described by the ATM Master Plan [19] to the trends, produced less robust results. The mapping issue could be due to the SDO descriptions not being sufficiently detailed for the LLM to fully understand their meaning. Two stakeholders in particular, ‘regulatory authorities’ and ‘technology and software providers’, were found to be impacted across most trends (i.e. these trends were ranked as having a high impact). Conversely, almost all trends were found to have a low impact on ‘energy and fuel providers’, ‘logistics, freight and delivery’ and ‘pilots, crew and aircrew’. The model was able to explore in greater depth the implementation, or evolution, of the trends themselves, by selecting enablers and barriers to the trends. The key *enablers* across most trends were as per the column headers in Table III). Of note, whilst green hydrogen was included as part of the ‘synthetic fuels and direct air capture integration’ enabler, hydrogen technologies *per se* did not feature in the main trends identified, an issue which the team has flagged for investigation in the next iteration of the model. The three *barriers* ranked as high for most trends were: ‘regulatory and certification’, ‘technological and performance limitations’ and ‘infrastructure and integration’. The ‘data, privacy and cybersecurity’ barrier received the most low rankings across the trends (five instances), however this barrier was still predominantly ranked as high (nine instances). Table III shows some of technological improvements, and other developments, which formed the three key enablers for five of the trends.

Trends related to personalised and high-speed air travel, such as ‘on-demand personalised air travel’ and ‘global supersonic and hypersonic travel’, exhibited consistently high negative environmental impacts across multiple categories. For instance, supersonic travel was linked to high nega-

TABLE III. THE THREE KEY ENABLERS AND SELECTED TRENDS

Trends	Key enablers		
	1. Technology & infrastructure	2. Gov. & regulatory support	3. Investment & collaboration
Electrification & advanced battery technologies	Advancements in battery tech.; charging infrastructure	Policy; emissions regulation	R&D investment; aero-battery collaboration
Synth. fuels & direct air capture integration	Electrolysis; renewable energy; carbon capture tech.	Government policies; carbon pricing	Infrastructure & supply chain investment
Unmanned freight & cargo drones	Autonomous navigation; battery tech.; communication	Regulatory framework; airspace integration	Cost-effective manufacturing & scalability
On-demand, personalised air travel	eVTOL technology; urban infrastructure	Regulatory framework; ATM integration	Booking platforms
Quantum computing-optimised ATC	Quantum hardware; communication networks	Regulatory & standardisation frameworks	Integration with classical ATC

tive impacts on greenhouse gas emissions, stratospheric pollution, and noise, alongside medium negative impacts on waste and resource consumption. In considering the social impacts, trends involving automation and surveillance, such as ‘fully autonomous commercial aircraft’, ‘unmanned freight and cargo drones’ and ‘global seamless biometric security and identity’, presented a mixture of positive and negative impacts. While these innovations may improve operational efficiency and security, the model flagged concerns around privacy, exclusion, and employment disruption. For example, biometric systems may exacerbate issues of gender bias, while autonomous aircraft could lead to significant job displacement, it was reported.

E. Building the concepts roadmap

1) *Roadmap progress thus far:* Initially, two contrasting roadmaps were prepared, presenting optimistic and pessimistic outlooks to 2050. The optimistic roadmap envisioned rapid technological progress and regulatory alignment. For example, advancements in battery density leading to an early emergence of electric regional aircraft, AI-powered predictive analytics enhancing ATM, and culminating in fully autonomous commercial flights and net-zero carbon growth by 2050. Key enablers included advancements in AI, aerospace-battery industry collaboration, synthetic fuels, and quantum computing, supported by harmonised regulations, public trust initiatives and infrastructure development. Conversely, the pessimistic roadmap anticipated significant delays and fragmented progress due to technological, regulatory and societal barriers. It highlighted slow adoption of electric regional aircraft, limited scalability of synthetic fuels and slow integration of digital and autonomous systems. Public resistance, cybersecurity con-

cerns, and lack of global coordination were recurrent themes. A comparison of the two roadmaps revealed that while both often referenced the same technologies, they differed markedly in timing, scope and feasibility.

2) *Roadmap into the EngageWiki*: Various roadmaps published by Europe’s Joint Undertakings and agencies have been reviewed, along with a sample from national bodies. The Clean Aviation JU have published roadmaps with detailed planning for each of three new generation of aircraft, showing phased timelines, challenges, targets and TRLs [20]. Roadmaps in the ATM Master Plan also present detailed timelines for the development and deployment of the key technologies towards the 2045 vision [19]. In contrast, the UK’s National Quantum Computing Centre’s technology roadmap sets out a pathway for the development of general-purpose quantum computing. Instead of a timeline, critical milestones were identified, such as the demonstration of an early quantum advantage [23]. Engage 2 plans to develop a relatively simple visual representation of its roadmap to display the AI-based results. An enabler-based approach will be used, rather than a timeline, with up to five enablers per technology. The inclusion of an off-the-shelf, interactive chatbot is also under consideration as a way of offering additional functionality to the user.

IV. WHERE NEXT?

A. *Agentic AI - an introduction*

GenAI (such as ChatGPT) demonstrates its utility and adaptability by producing content in a range of styles and in different contexts, based on user prompts. Agentic AI, in contrast, refers to largely *autonomous* artificial intelligence systems, which may learn, make decisions, and take actions towards specific and complex goals, with minimal human supervision. It typically uses LLMs as its main resource (‘brain’) to assimilate and process data and coordinate tools. Unlike ‘basic’ LLMs, agentic AIs are proactive systems that exhibit agency by setting their own workflow dynamically, making plans and re-assessing them autonomously in order to reach high-level objectives. Agency is a spectrum, and chatbots, such as those provided by OpenAI, Google and Amazon through web apps of their APIs, are typically low on this spectrum, in the sense that they execute atomic tasks. However, the distinction is becoming increasingly blurred, with the introduction of ‘chain-of-thought’ and ‘deep thinking’ chatbots, which are able to assess and modify their answers before submitting them to the user, in addition to using external tools (such as internet searches). However, more agentic systems are typically able to adapt to dynamic environments, iterating on outcomes, which allows them to handle complex tasks pro-actively, even in real time, such as optimising schedules or re-routing passengers. We discuss further the core characteristics of agentic systems, in particular for application in knowledge generation, in Section V. For a review of agentic AI, focusing on its evolution and three ‘core’ aspects (patterns, types, and environments), concluding that “challenges persist in aligning AI with human values, ensuring adaptability, and addressing ethical constraints”, see [24].

Exploring the theme of better managing hallucinations through the use of agentic AI, recent work [4] deploys a competitive market where AI agents buy answers from upstream generative models, reselling them to users who differ in how much they fear hallucinations, cf. value accuracy. In this context, it is suggested that ‘hallucination-sensitive’ domains (such as law and medicine), “endogenously lead to more serious verification efforts in agentic AI markets”.

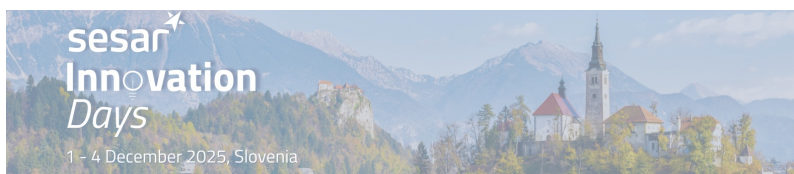
B. *Maturing the concepts roadmap*

Roadmaps aim to summarise what we know about a field and to extrapolate and forecast what the future may look like for the corresponding domain. They represent as much a forecast as a tentative prescription of the future, taking into account what seems to be feasible. They thus usually require significant effort and knowledge in compilation, domain intuition, and forecasting, which is typically performed by *groups* of humans, especially for complex domains. This is because of the variety of knowledge and cognitive processes needed. A good example is the SESAR ATM roadmap produced by the SESAR 3 JU [19], built on layers and layers of deep understanding of the system, as well as policy motivation. Since AI is an automation of the human cognitive process, and an LLM is an approximation of a human brain, it is natural to try to use a *group* of LLMs, instead of isolated ones, so the system can reach some kind of consensus. It is also logical to raise the level of agency of the AI, so that it autonomously solves some tasks. In the context of a roadmap, both are crucial. The static system depicted in Figure 1 cannot perform several important cognitive tasks:

- Each step has no self-reflection. For example, if quantum computing were to be selected as an important topic, there is no feedback to check if this topic is indeed relevant, or just a hallucination.
- Isolated processes tend to consider information from one source as being as important as another. In reality, complex weighting processes occur in the human brain, which may also arise in LLMs, but are not explicitly transferred from one LLM to another in a multi-step environment. Prioritisation tasks, crucial for more reliable outcomes, are thus lacking.
- Each step is run on a single LLM and thus its answer is final. In reality, because of the noise in such models, several answers could be provided by the same LLM, which may differ significantly.
- A single LLM gives answers based on its default training, unless prompted to do otherwise; a roadmap requires complex interactions and consensus-seeking processes - only possible with multiple LLM passes.
- None of the LLMs used in this study are fine-tuned, or can be considered expert in their own field, e.g. thanks to dedicated RAG systems.

Building a realistic roadmap may thus require more complex agentic setups, in particular using:

- Reactive agents: e.g. one instructed to use an LLM, inspect the output, and modify its query iteratively until



the request from the user has been answered satisfactorily. This type of agent can also use tools as needed, such as an internet search, or RAG query. Agents may also call other agents, leading to multi-agent processes.

- Dedicated RAGs: focused corpuses of texts allow agents to be more expert in specific domains. Dynamic RAGs can also be built to allow an agent to self-teach about a specific domain, e.g. using an online connection.
- Fine-tuned or highly-primed agents: in order to reach a realistic consensus, agents need to come from different ‘cognitive places’, which is what we used already with the pessimistic and optimistic behaviours described in Section III-E1. More behaviours need to be injected, such as more politically-oriented ones (which could be fine tuned on real politicians), more scientifically oriented, more cautious, more ambitious, etc, in order to better represent the variety of opinion, and reach a more robust consensus.

The overall structure of the new multi-agent setup may thus resemble the following. Dedicated agents, with ingrained fact checkers, could be tasked with extracting information on demand. An orchestrator could supervise the extraction of the main domains, asking the specialised agents to extract and check knowledge. Agents dedicated to prioritisation may review this knowledge and ask the specialised agents for more information, as needed. At this point, different agents with high-level specific behavioural profiles (politicians, scientists, etc.) start the consensus process. This will require the creation of highly specialised agents *created* by the orchestrator, with the tasks of acquiring deep knowledge about each important domain. This may include the dynamic creation of RAG systems, and the self-training of these experts. The high-level profiles then have everything they need to build the roadmap. Through an iterative, consensus-forming process, these agents can suggest important milestones, check with experts, ask for potential correlations with other domains, ask the experts again, etc., until the orchestrator decides that the roadmap is sufficiently robust. Since logs of this process will be kept, a final chatbot could be given access thereto and explain to a user the rationale behind any decision, highlighting potential conflicts between agents, and how they were solved.

Such considerations are particularly relevant for high-stake domains such as ATM, where strategic decision-making is needed. The reliability and relevance of AI helping humans is still an open question. As with many automation tools, their usefulness regarding long-term strategies will become evident over a longer time horizon. However, there is reasonable confidence that AIs will be useful for strategic decisions. Firstly, because human-level intelligence has already been demonstrated by AI in many domains, *by non-specialised models*¹¹. Secondly, as with cooperation, because consensus-forming processes, and other social interactions, were key to

historical human successes, whereby agent-driven AI swarms will likely be a good answer to such current limitations. The capacity to make crucial decisions in uncertain environments with delayed consequences, arises as much from the human brain as from our interactions with others.

C. Human expertise and interaction

While the documents for the roadmap have all been produced by humans in the first place, one might question the injecting of intermediate expert opinions in the construction of the roadmap *per se*. Indeed, human experts do have access to other types of knowledge than GenAI, through their experience, and thus may have other intuitions on certain subjects. In general, a successful GenAI-led, automated process may need to go through a phase of cooperation with humans to be more efficient (and accepted by humans). In the case of high-level knowledge, it might be that even in the long-run, humans need to be involved in the process, as we point out below.

In the meantime, one has to provide humans with a natural way of interacting with the process, be this at the construction stage, when the AI builds the roadmap, or at the consultation stage, when the roadmap is finished and the human wants to explore it. Fortunately, as noted before, GenAI systems have been designed for human interaction.

Regarding the construction phase first, there are several possibilities for including human feedback in the process. In the sequential process presented in this paper, one could imagine intermediate steps where humans check the output of the AI, and offer advice that may be integrated into the prompt of the next stage. For example, after running the ‘stakeholders impacted’, a human could flag that one of the impacts does not make sense. The output could then be directly amended by the human expert, or such advice be incorporated in the prompt.

Still in the same sequential process, one could imagine several iterations for each step. Taking the same example again, the human expert could provide comments on an output about the stakeholder’s impact, which would be incorporated into a prompt fed to the *same* AI, that would refine the output, and so on. This AI could have a memory of the previous iterations, or not, depending whether one wanted to achieve a faster equilibrium (with memory) or a more objective one (without memory). Note that this process could also be used to fine-tune an agent.

In a more agentic type of process, the human could also be considered as a node, on a par with other AI agents. Indeed, an expert human could be considered as a reviewer agent, a weighting agent, or a primary knowledge producer (for example, listing stakeholder impacts for a certain technology). Human experts could provide qualitative assessments of some output (classifying them as ‘trusted’ or ‘uncertain’ output, etc.), or quantitative ones through weights or scores. They could take part in the final decision-making process, for example by participating in a consensus-building process, dialoguing directly with AIs (or other experts) to make the

¹¹This is crucial because humans rely on a massive amount of data from their experience and knowledge before making strategic decisions, instead of relying on hyper-specialised cognitive processes, such as vision.

final assessment. They could also provide refined advice, which could help to improve the quality of the roadmap, suggesting (or asserting) changes of vocabulary, semantic subtleties, cosmetic changes, etc., that would make the roadmap more readable and acceptable for other humans.

Technically speaking, none of the above is particularly difficult to implement. Humans could interact via different means, e.g. by e-mail to provide reports, or standard prompt windows for fast interactions and dialogues. The supporting infrastructure needs to be asynchronous, a current standard for agentic systems, and additional care must be taken when dealing with sensitive data. Apart from this, the infrastructure is not different from a full AI system, since nodes should be agnostic to the source of their information in a graph-based infrastructure.

The interaction with the final results for the broader readership could also be enhanced through the use of GenAI, as opposed to the more static roadmaps produced so far¹². While a fairly simple roadmap is planned to be deployed on the EngageWiki website (see section III-E2), a chatbot, as mentioned above, with access to the internal rationale leading to the roadmap and/or with the proper references, could help stakeholders and the broader readership when trying to understand the meaning of the roadmap.

D. Transversal opportunities in IR2

In the SESAR transversal projects PEARL (Performance Estimation, Assessment, Reporting and simuLation)¹³ and AMPLE3 (SESAR 3 ATM Master Planning and Monitoring)¹⁴, respectively looking after performance assessment and supporting the updating of the ATM Master Plan, we see a significant potential for GenAI going forward, especially in their successor (IR2) projects. This is due to the particular strengths of GenAI in converting large corpuses of text into high-level knowledge (across multiple SESAR Solutions as multifactorial input cases), making links between unconnected elements (e.g., performance dimensions and architecture components), and extrapolating information (e.g., from specific Solutions to wider, network contexts).

There could thus be scope for GenAI-driven automatic reviews that deploy RAG with all relevant fine-grained information related to air transportation, including SESAR, similar to the approach for the ATM concepts roadmap, above. A static LLM could manage the high-level knowledge extraction (interacting with KPA champions to check the results, for example), with, at the next level, a dynamic agent, automatically reading reports, asking dedicated questions of other expert AI agents (e.g. operational questions about ATFM delay), forming consensus, producing detailed reports and sending them back to individual projects (after review by KPA champions).

This could be supported by a chatbot AI answering questions from individual projects (such as clarifications and

support, and interfacing with the expert AI agents, to open dialogues on user guidance and training. Such an implementation would need to comply with two critical technical specifications. (i) The rate of hallucinations needs to be low. This could be assessed manually by the project, using shadow mode procedures where human and AI work on the same task in parallel, then checking the AI output against the human's. (ii) Documents compiled for the RAG mostly need to be treated confidentially. Use could thus be made of open models (such as LaMA 3.1, DeepSeek-R1, Qwen3, Mistral 8x7B, or gpt-oss) on local machines. This will ensure that no confidential information is 'leaked'. The results from these processes could serve as validation exercises in IR2 for AI-powered tools, allowing these in particular, and SESAR in general, to capitalise on similar AI uses by national governments and international businesses.

V. CONCLUSIONS

Across our illustration of current, and potential future, AI research in SESAR, considering generative AI in particular, we have seen a wide range of work being executed in ways hitherto not imagined, before the advent of these technologies. We have been mindful, however, to flag potential cautionary notes where appropriate. In our case study of the concepts roadmap, the omission of hydrogen as a key fuel technology and the weak analysis of the SDOs are areas for us to investigate further as we advance the technology into the use of agentic AI and retrieval-augmented generation. Of note, hallucination was not an apparent problem, as indeed we had hoped to manage this in our model. In reviewer feedback on the technologies identified, one criticism received has been that the mapping is too conservative – something we aim to explore in the next phase, through prompt engineering under increased autonomous control.

An important aspect of the future use of AI is whether agentic development in general will gain traction or stay focused on specific tasks. Looking at the wider industry, it seems that developers in general are moving towards more abstract positions, leaving more and more low-level development to AI systems, which are increasingly capable of producing reliable standard code. (Although we note in passing that this has led to reports of coders being increasingly less able to de-bug code when subsequent errors arise, having not been part of the core creation process.) This trend, similar to wider trends in the past, shows automation increasing across industry, and may lead to agentic systems becoming more prevalent. These systems are, however, currently far from reliable and largely untested, even for simple applications such as presented in this paper. While their added value has been proven, for example in standardised tests, it is as yet unclear how reliable and original they are for general forecasting and knowledge generation. Moreover, it is still completely unknown whether massive multi-agent AI systems could catastrophically fail due to emergent phenomena.

This raises the important question of the validation of such procedures. By their nature, there is no 'ground truth' in

¹²Static roadmaps are also very useful because they can be easily shared through social media and other means.

¹³<https://www.sesarju.eu/projects/pearl>

¹⁴<https://www.sesarju.eu/projects/ample3>



such processes, because roadmaps are a blend of forecasting, creative processes, and political will. Thus, in this context, the concept of validation is deeply linked to the *perceived* value of the AI output, by (human) experts. Any kind of validation of the results presented here would thus require (intensive) rounds of discussion with experts. This is indeed something that we plan for the future. However, this begs the question: will humans only consider roadmaps (or other types of strategic document) in which they readily perceive value, or will they, at some point in the future, trust the AI to produce strategic thinking that they are not able to fully comprehend or intuit, just as we already have confidence in machines to give us answers (e.g. in engineering) that we are not able to fully understand?

Indeed, although potentially prescient, agency deeply touches the human psyche. (Generative) AI drives not only the automation of computational cognitive processes, but also reaches into *creative* processes, a domain which is often viewed as the last true domain with human supremacy – and roadmaps typically rely on both types of process. It might be that only ideas and creative thoughts, which are sufficiently close to what a human could think of, will be considered interesting, or credible. This may mean that humans will remain in charge of certain high-level tasks, especially in safety-critical domains such as aviation and ATM, because they consider AI supremacy as being too alien. This is equally relevant to political projects – roadmaps included.

ACKNOWLEDGMENT

This project is supported by the SESAR 3 Joint Undertaking and its founding members under Grant Agreement No. 101114648. The University of Westminster, a UK participant, received funding from UK Research and Innovation under the UK government's Horizon Europe funding guarantee No 10083017.

REFERENCES

- [1] Howard A, Reza N, Green PL, Yin M, Duffy E, Mwandumba HC, Gerada A, William Hope W, "Artificial intelligence and infectious diseases: tackling antimicrobial resistance, from personalised care to antibiotic discovery," *The Lancet Infectious Diseases*, vol. Artificial Intelligence and Infectious Diseases 3, 2025, DOI: [https://doi.org/10.1016/S1473-3099\(25\)00313-5](https://doi.org/10.1016/S1473-3099(25)00313-5).
- [2] Gwon YN, Kim JH, Chung HS, Jung EJ, Chun J, Lee S, Shim SR, "The use of generative AI for scientific literature searches for systematic reviews: ChatGPT and Microsoft Bing AI performance evaluation," *JMIR Medical Informatics*, vol. 12, 2024, DOI: <https://doi.org/10.2196/51187>.
- [3] European Parliament, "Artificial Intelligence Act," 2025. [Online]. Available: <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>
- [4] Iyidogan E, Ozkes AI, "Agentic AI and hallucinations," *Economics Letters*, vol. 255, 2025, DOI: <https://doi.org/10.1016/j.econlet.2025.112520>.
- [5] Heo S, Son S, Park H, "HaluCheck: Explainable and verifiable automation for detecting hallucinations in LLM responses," *Expert Systems with Applications*, vol. 272, 2025, DOI: <https://doi.org/10.1016/j.eswa.2025.126712>.
- [6] Ali H, Pham D-T, Alam S, Schultz M, Li MZ, Wang Y, Itoh E, Duong VN, "Human-AI hybrids in safety-critical systems: concept, definition and perspectives from air traffic management," *Advanced Engineering Informatics*, vol. 65(B), 2025, DOI: <https://doi.org/10.1016/j.aei.2025.103256>.
- [7] Knoth N, Tolzin A, Janson A, Leimeister JM, "AI literacy and its implications for prompt engineering strategies," *Computers and Education: Artificial Intelligence*, vol. 6, 2024, DOI: <https://doi.org/10.1016/j.caeai.2024.100225>.
- [8] Macnamara BN, Berber I, Çavuşoğlu MC, Krupinski EA, Nallapareddy N, Nelson NE, Smith PJ, Wilson-Delfosse AL, Ray S, "Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness?" *Cognitive research: principles and implications*, vol. 9, 2024, DOI: <https://doi.org/10.1186/s41235-024-00572-8>.
- [9] Danet D, "LLM Grooming: A new cognitive threat to generative AI," *SSRN*, 2025, DOI: <http://dx.doi.org/10.2139/ssrn.5461315>.
- [10] European Union Aviation Safety Agency, "Artificial Intelligence Roadmap 2.0 - human-centric approach to AI in aviation," 2023. [Online]. Available: <https://easa.europa.eu/ai>
- [11] Lanzi P, Spiller E, Jameel M, Christoffels L, Gigante G, Esposito R, Contissa G, Sanchi M, Galli F, Everdij M, Stroeve S, "Challenges and new directions for the certification of AI and advanced automation in civil aviation," *SESAR Innovation Days 2024*, 2024. [Online]. Available: <https://www.sesarju.eu/SIDS2024>
- [12] Future of Life Institute, "The EU Artificial Intelligence Act - up-to-date developments and analyses of the EU AI Act," 2025. [Online]. Available: <https://artificialintelligenceact.eu/>
- [13] SESAR 3 Joint Undertaking, "SESAR Innovation Pipeline: Air traffic management research and innovation (2024 highlights)," 2025. [Online]. Available: <https://www.sesarju.eu/node/4873>
- [14] Groia M, Vendruscolo T, Vaiopoulos P, Bonelli S, Gauci J, Bezzina M, Berling D, Jurvansuu M, Borovich N, Koopman C, Grech L, Zaidan R, Bortoli AD, Brambati F, "AI-Enabled Tactical FMP Hotspot Prediction and Resolution (ASTRA): A Solution for Traffic Complexity Management in En-Route Airspace," *Engineering Proceedings*, vol. 90(1), 2025, DOI: <https://doi.org/10.3390/engproc2025090091>.
- [15] Chen Y, Dalmau R, Alam S, "Spatio-temporal Graph-based Demand and Capacity Balancing Considering Flight Uncertainties," *SESAR Innovation Days 2024*, 2024. [Online]. Available: <https://www.sesarju.eu/SIDS2024>
- [16] Ahmed MU, Barua S, Begum S, Jmoona W, Cruze RSD, Veyrie A, Kebir S, Hurter C, *Role of Multi-modal Machine Learning, Explainable AI and Human-AI Teaming in Trusted Intelligent Systems for Remote Digital Towers*. Association for Computing Machinery, 2025, DOI: <https://doi.org/10.1145/3719384.3719389>.
- [17] Ruseno N, Chagas FS, Fas-Millán MA, Bechina AAA, "Analysis of API-Based Communication Performance for drone's operation in U-Space," *Transportation Research Procedia*, vol. 81, 2024, DOI: <https://doi.org/10.1016/j.trpro.2024.11.021>.
- [18] Trausmuth G, Cook A, Tanner G, Gurtner G, Takacs V, "D4.1 ATM sector changes," Engage 2 Consortium, Tech. Rep., 2024. [Online]. Available: https://engagektn.eu/wp-content/uploads/2025/02/Engage-2-D4.1-%E2%80%9393-ATM-sector-changes_rev01.00_14-OCT-2024.docx.pdf
- [19] SESAR 3 Joint Undertaking, "European ATM Master Plan (2025 edition)," 2025. [Online]. Available: https://www.sesarju.eu/MasterPlan2025_SupportingDocuments
- [20] Clean Aviation Joint Undertaking, "Strategic Research and Innovation Agenda (2024 edition, 2nd revision)," 2024. [Online]. Available: <https://www.clean-aviation.eu/about-us/reference-documents>
- [21] European Rail Research Advisory Council, "Rail Strategic Research and Innovation Agenda," 2020. [Online]. Available: https://rail-research.europa.eu/wp-content/uploads/2020/12/RAIL-Strategic-Research-and-Innovation-Agenda-2020-_FINAL_dec2020.pdf
- [22] European Commission, "Annex to the Commission Decision on the submission of the Europe's Rail Joint Undertaking Master Plan for adoption by Europe's Rail Governing Board," 2022. [Online]. Available: https://rail-research.europa.eu/wp-content/uploads/2022/03/GB-Decision_01_22_Annex_Master-Plan.pdf
- [23] National Quantum Computing Centre, "Technology roadmap," 2023. [Online]. Available: <https://www.nqcc.ac.uk/technology-and-research/technology-roadmap/>
- [24] Nisa U, Shirazi M, Saip MA, Pozi MSM, "Agentic AI: The age of reasoning - A review," *Journal of Automation and Intelligence*, 2025, DOI: <https://doi.org/10.1016/j.jai.2025.08.003>.

