TAILOR

Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization
TAILOR
Grant Agreement Number 952215

Strategic Research & Innovation Roadmap V2

Document type (nature)	Report
Deliverable No	D2.5
Work package number(s)	WP2
Date	2024-09-12
Responsible Beneficiary	TNO
Author(s)	Cor Veenman (TNO), Ruud Mattheij (TNO), Isabelle Tilleman (TNO), Freek Bomhof (TNO), Wico Mulder (TNO), Judith Dijk (TNO)
Publicity level	Public
Short description	This document is the second version of the *Strategic Research and Innovation Roadmap* (SRIR V2) of the TAILOR project and aims to add orderings and prioritisations to the first version, in order to create a guiding roadmap in the research landscape on trustworthy AI.

Document History			
Revision	**Date**	**Modification**	**Authors**
D2.5 Version 1	12-9-2024	Updated and expanded from SRIR v.1 (D2.1)	See above

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Annelot Bosman	ULEI	26-8-2024
Fredrik Heintz	LIU	18-8-2024

Table of Contents

Summary of the report

This document is the second version of the *Strategic Research and Innovation Roadmap* (SRIR V2) of the TAILOR project, which focuses on Trustworthy Artificial Intelligence through Learning, Optimization, and Reasoning.

The purpose of the roadmap is to identify research directions and technology gaps that need to be filled to achieve Trustworthy AI. By placing these research directions in the context of the needs, both from industry and from challenging application domains, we provide a framework for strategic innovation development in the context of both Trustworthy Artificial Intelligence (TAI) and Learning, Optimization and Reasoning (LOR). Through the definition of AI development layers and the mapping of research directions for challenging AI applications on these layers, we obtain a prioritisation needed for operationalisation.

Based on the application domains that were 'tested' in the roadmap, the following research areas are identified as the most important to pay attention to:

1. **Neurosymbolic AI** combining data-driven and knowledge-driven methodologies for the embedding of domain knowledge to improve task performance, user-centricity as well as enabling the embedding of values in models: value alignment.
2. **Transparency** and **transparency & traceability** as listed in the HLEG trustworthy AI requirements need special research attention for user-centricity and oversight.
3. For the involvement of stakeholders, user-centricity and value alignment a **multidisciplinary approach** is needed.
4. Research topics related to learning and deploying on federated data sources, such as **data sharing** for task performance and **federated learning** for privacy, i.e. value alignment.
5. **Uncertainty & risk** is important as value alignment topic, besides for the considered showcases, also for foundation models and generative AI as put forward in the Moonshot.

These areas are rooted in both TAI and LOR, but they also cover Ethical, Legal and Societal aspects (ELSA) and infrastructural topics.

Introduction to the Deliverable

The first version of the SRIR **D2.1** is found on the TAILOR webpage[1] on the Results page (at least until 2029). It gives a full overview of the research landscape for Trustworthy AI. The first version builds upon the insights of the academic research partners of the TAILOR project.

The aim of this second version of the SRIR is to:

- Extend the first version of the roadmap by incorporating additional, relevant research topics and priorities from industrial domains
- Operationalise the collected research topics into a guiding road map by combining and mapping research topics with priorities.

Process and people

[1] https://tailor-network.eu/

The second version of the SRIR is developed by TNO (main author: Cor Veenman), aided by Roberta Calegari (UNIBO), Barteld Braaksma (CBS), and Fredrik Heintz (LiU). The individual roles in this collaboration are presented in the table below.

Partner Acronym	**Name**	**Role**
TNO	Cor Veenman, Ruud Mattheij, Isabelle Tilleman, Freek Bomhof, Wico Mulder, Judith Dijk	Design of the methodology, development of SRIR v2, writing and updating document
UNIBO	Roberta Calegari	Methodology application and evaluation, collaboration process with ICT-48 CSA, PPP-AI, and AI4EU (described in D11.6)
CBS	Barteld Braaksma	Methodology application and evaluation
LiU	Fredrik Heintz	Methodology application, evaluation and guidance

1. Background and methodology

While AI research and technology has been around for a long time, the AI research domain has entered an extremely vivid, dynamic, and challenging phase. Over the last decades, we witnessed an extreme acceleration of activities and achievements related to AI. In a wide range of domains, AI systems enable significant gains in effectiveness and efficiency of task performance. Moreover, in an increasing number of (use) cases, AI systems are able to deliver task performances that are beyond human capacities. However, beyond the mere focus on task performance itself, there are rising concerns on the development and deployment of AI systems with respect to their trustworthiness.

Addressing these concerns, the AI HLEG[2] in 2019 put forward seven requirements that AI systems should satisfy in order to become Trustworthy AI systems. With the AI Act (Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence), these concerns and consequent requirements are embedded in a formal regulation. The TAILOR project, and this SRIR, specifically, aims to support research and development of Trustworthy AI systems by (1) defining key research topics by aligning academic and industrial view points, and (2) prioritising the resulting collection of research topics.

The first version of the SRIR focused on the identification of key research topics from an academic viewpoint (**D2.1**[3]**)**. The second version of the SRIR, hereafter referred to as SRIR V2 and presented as deliverable **D2.5**, addresses the review comments and extends the first SRIR to include industrial needs and the prioritisation of the research topics identified. A number of workshops were arranged internally as well as with other AI initiatives. The methodology used in the workshops as well as the obtained results are reported in deliverable

[2] https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1
[3] https://tailor-network.eu/research-overview/strategic-research-and-innovation-roadmap/

D11.6 Progress Report on SRIR Development in Collaboration with the ICT-48 CSA, PPP-AI, and AI4EU, v2[4]. The SRIR V2 allows for an effective operationalisation of the identified research topics via a guiding roadmap for a wide range of industrial application domains.

Audience

To increase the SRIR V2's focus, we specify the roadmap's intended audience. The SRIR V2 supports professionals bearing responsibilities to (1) make in-depth choices on their future research topics, and/or (2) to make the budgets and resources available for specific research areas. A clear example of the first group can be found in the European AI research and development community, while European Commission and research funding bodies at the national level are prime examples of the latter group.

Additionally, the SRIR can be used by any other entity or organisation with ambitions to innovate in the areas of trustworthy AI, learning, optimization, and reasoning. To facilitate this, the SRIR is designed to provide guidance through the overall landscape of Artificial Intelligence and related techniques. For such entities and organisations, the SRIR V2 not only shows techniques that warrant further development, but also shows several clear examples how those techniques can be applied to specific domains.

Lastly, the roadmap can be used by domain experts to sharpen their ideas as well as focus areas. The overview presented via the roadmap provides guidance on potential research areas to focus on, as well as on the current status of those research areas. These insights aid the decision making process on the feasibility of ideas.

Review comments

The first version of the *Strategic Research and Innovation Roadmap* (SRIR V1) was published on April 13th, 2022. It contained an overview of research topics that are relevant in the process of achieving trustworthy AI systems. The EC review process provided valuable feedback and good suggestions for improvements and two suggestions are particularly incorporated in this version.

The first point concerns the prioritization and ordering of research topics needed to provide clear directions in the roadmap. While listing relevant research topics is arguably important, it does not suffice when aiming for accurate planning and programming of research and development efforts on Trustworthy AI. As such, where the SRIR V1 outlined the overall research landscape, the SRIR V2 intends to help researchers and policy makers to navigate the challenging research landscape on trustworthy AI.

The second point concerns the reviewers' endorsement on the selected research areas where data-driven and knowledge-driven principles come together, resulting in the emerging fields of hybrid AI and neurosymbolic computing. SRIR V2 clearly indicates that, to develop ethical and legally compliant systems, these principles must come together, besides other potential gains.

[4] All TAILOR deliverables are found here: https://tailor-network.eu/about/deliverables/

Outline

SRIR V2 builds on SRIR v1, but additionally incorporates the industrial needs in research topics and provides an ordering and prioritisation of research topics towards a roadmap in the research landscape. As such, SRIR V2 is built upon the following foundation:

1. **Industrial needs and focus**
   To understand the needs of the industry in various application domains, TAILOR participated in a workshop under the ICT-49 initiative to understand the requirements for TAI in project exploitation by industries and SMSEs. Collaboration with AI4Eurrope resulted in sharpened focus in research needs and directions for TAI. Further, TAILOR organised seven Theme Development Workshops (TDW). These TDWs have explored the role of AI in their respective domains and identified the required core AI research topics for the next 5-10 years. The recommendations detailed in the TDW reports ("input for the roadmap") were added to the existing short list of research topics available in SRIR V1, resulting in the extended list of relevant research; see Annex **1**.
2. **Categorization in topic sectors**
   The existing short list of topics in SRIR V1 was subdivided into the categories (1) *Trustworthy AI* (**TAI**) and (2) *Learning, Optimization, and Reasoning* (**LOR**). To enable a more actionable categorization, we introduced two additional categories: (3) *Ethical, Legal, and Societal Aspects* (**ELSA**), and (4) *Data and Infrastructure* (**Infra**). In the remainder of this document, we refer to these categories as topic sectors, which are further presented in Section **2** below.
3. **AI development layers**
   We cluster and categorize all research topics and propose the term AI development layers. In order to fulfil all systemic needs, research is needed from basic functionalities and contextual awareness, to the enabling of oversight over AI systems. The development layers in AI systems are the subject of Section **3**.
4. **Prioritization based on showcases**
   Emphasized by the findings during the TAILOR project, it can be concluded that the prioritization of research topics strongly depends on the application domain. By aligning the layered research topics and the demands identified in a selected number of impactful showcases, a preliminary prioritization of research topics is established, as shown in Section **4**.

Subsequently, the document covers our recommendations, in Section **5**, the expected future developments, in Section **6**, and Conclusions, in Section **7**.

2. Categorisation in topic sectors

The TAILOR project revolves around the high-level categories *Trustworthy AI* (**TAI**) and *Learning, Optimization and Reasoning* (**LOR**). To enable and support a more actionable categorization, two additional categories are introduced: *Data and Infrastructure* (**Infra**) and *Ethical, Legal, and Societal Aspects* (**ELSA**). These four categories are referred to as *topic sectors*. Within the topic sectors, we define the high level *topic clusters* to have a meaningful and intuitive overview of the (clustered) research topics per sector, as illustrated in the sector

map presented in Figure **1**. In what follows, the topic sectors, the corresponding clusters, and the research topics will be presented in more detail.

TAI – Trustworthy AI

Trustworthy AI, or TAI, is one of the two original core categories within the TAILOR project. The topics shared under this category are the core concepts of Trustworthy AI. The ethics guidelines for trustworthy AI, as developed by the HLEG, identify multiple high-level requirements that AI systems should meet. We use these high-level requirements as the topic clusters associated with the TAI topic sector:

1. *Human agency and oversight*
2. *Technical robustness & safety*
3. *Privacy & data governance*
4. *Transparency*
5. *Diversity, non-discrimination, fairness*
6. *Societal & environmental well-being*
7. *Accountability*

For industry, these qualities / features will have to be verified or validated to prove AI system compliance with the AI Act. The processes for this are not part of this deliverable.

LOR – Learning, Optimizing and Reasoning

Learning, Optimizing and Reasoning, or LOR, is the second of the two original core categories within the TAILOR project. The topics shared under this category are the core concepts representing the mathematical and algorithmic foundations onto which AI and its applications are built. Additionally, knowledge and acting (i.e., an AI's execution of tasks) are closely related to - and part of - the TAILOR research.

Here, it must be noted that paradigm-shifting developments in the domains of Foundational Models and Generative AI were not yet integrated in the TAILOR research plan and scope, given that these technologies were not yet in the forefront of scientific research at the time of writing of the TAILOR project plan. Acknowledging the relevance of these developments, the TAILOR team decided to establish specific topic clusters within LOR to host the research topics related to Foundational Models and Generative AI appropriately.

From this, we defined the topic clusters for LOR as follows:

8. *Learning*
9. *Optimization*
10. *Reasoning*
11. *Knowledge*
12. *Generative AI*
13. *Acting*

ELSA - Ethical, Legal, and Societal Aspects

The Ethical, Legal, and Societal Aspects of AI (ELSA) are important elements of the requirements for trustworthy AI. We identified the following topi**c clusters** for this topic sector:

14. Law & Regulation
15. Policy
16. Ethics
17. Politics, society & business
18. Governance

Examples of relevant Law & Regulation topics are the General Data Protection Regulation (GDPR)[5], the AI Act, the Digital Markets Act (DMA)[6], and the Data Governance Act (DGA)[7].

Examples of relevant Governance topics are the organization of quality, responsibilities, and accountability.

Infra - Infrastructure

The infrastructure (or Infra) on which an AI implementation runs, should facilitate its functionality in all respects. As such, it can be reasoned that these functionalities, too, relate to the trustworthiness aspects of AI. It is therefore important to include infrastructure-related topics in the roadmap, even if these topics are not directly related to AI or trustworthy AI. In other words, topics in this sector may not necessarily directly relate to TAI, LOR, or even AI in general. They are nonetheless essential, to provide an integrated approach towards trustworthy AI.

To adhere to this reasoning, we have defined the topic clusters for Infra as follows:

19. Data [storage, sharing, acquisition]
20. Hardware [high performance computing, edge AI]
21. Architecture [processing and AI architectures]
22. Systems engineering

[5] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[6] Regulation (EU) 2022/1925
[7] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022R0868

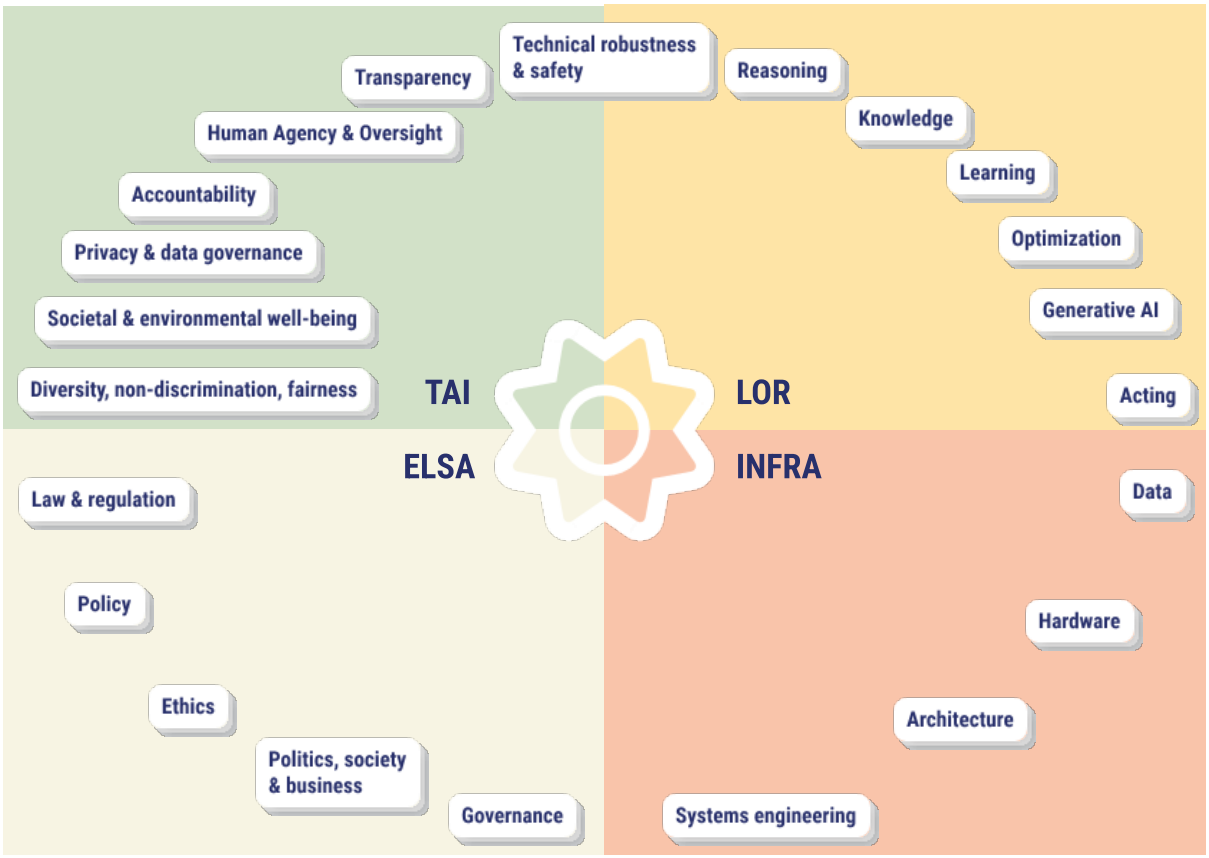


*Figure **1**: Overview of the four topic sectors; each of the topic sectors is subdivided into multiple topic clusters; each topic cluster consists of one or more research topics (not shown in this Figure).*

3. AI Development Layers

Historically, the development of AI systems found its origin in academic settings. Different applications for AI technologies were explored, in which the technology's *task performance* (expressed as, for example, accuracy) at a level that was generally considered to be acceptable was key. Over time, significant progress and successes were achieved, where learning, optimization and reasoning were leading in some form, either standalone or in some combination. Gradually, interaction with the user and *user centricity* received more and more attention[8], especially because machine learning methods yielded opaque or black box models. With the broader introduction of AI systems, also the misalignment with our human and societal values became apparent. This requires *value alignment* in order to fit the AI system with the norms of users, data subjects, organisations and society[9]. In a broader sense, AI systems are impacting our society and human well-being in terms of energy use and *sustainability* in general[10]. Currently, we are witnessing the regulation of AI systems in order

[8] L. Oberste and A. Heinzl, "User-Centric Explainability in Healthcare: A Knowledge-Level Perspective of Informed Machine Learning," in IEEE Transactions on Artificial Intelligence, vol. 4, no. 4, pp. 840-857, Aug. 2023.
[9] C. Huang, Z. Zhang, B. Mao and X. Yao, "An Overview of Artificial Intelligence Ethics," in IEEE Transactions on Artificial Intelligence, vol. 4, no. 4, pp. 799-819, Aug. 2023.
[10] D. Shin and E. Y. Shin, "Human-Centered AI: A Framework for Green and Sustainable AI," in Computer, vol. 56, no. 6, pp. 16-25, June 2023.

to formalize these aspects in regulations and standards towards the enabling of *oversight*[11]. All these stages can be seen as *development layers* in AI systems.

In new application domains, we see similar development layers, starting with explorations in research ─ and academic settings, up to the moment where the context in which a system operates relevant and, ultimately, oversight on the system is formalized and acted upon. Following this reasoning, we define development layers of AI systems as follows:

1. **Task performance**: this first layer focusses on exploration and initiation to establish a possible acceptable level of task performance. Later, the focus shifts to performance optimization towards the design of task specific technology.
2. **User centricity**: involving the user in a broad sense to make the system able to communicate (continuously) for preferences and other input on the one hand and user aligned queries and results on the other hand. Topics involved range from post-hoc explanations to tailored explainable AI and theory of mind.
3. **Value alignment**: embedding ethical, legal and societal values in the AI system, including policies and other contextual values and enabling the stakeholders/users to make suitable trade-offs. As such, topics range from fairness measures to holistic trustworthy AI.
4. **Sustainability**: for the building and during deployment of AI systems based on data intensive and parameter rich models immense amounts of energy are required. Topics range from energy usage to edge AI and frugal AI.
5. **Oversight**: once systems are functionally complete, it should be possible to audit and check for compliance with generic and domain specific laws, regulations and standards. Topics range from benchmarking to formal verification.

Here it has to be noted that in some domains the focus is on certain development layers a-priori, given that the respective topics need specific attention - for example: *value alignment* and *oversight* in sensitive domains, *sustainability* in energy-intensive applications, or *user centricity* for embodied AI agents.

Several research topics are hard to be linked to single development layers, since they are important in multiple layers. By assigning research topics to the most relevant development layer, a preliminary layering of research topics appears, see Figure **3**. It should be noted that there is no strict sequential development through the layers in AI development, given that AI-related research is an iterative or agile process in nature. That is, often evidence is needed for base level functionality at various layers, before deepening research in other layers. While research can, may, and will continue at the outer layers, it is imperative that the inner layers should not be ignored or deferred during the development process of trustworthy AI. That is, the inner layers are at least as important to the outer layers, especially for high risk AI systems. In Figure **3**, AI-related research topics are mapped onto the development layers. Here, it should be noted that the figure merely shows a subset of representative topics to achieve maximum readability.

[11] J. R. Carvalko, "Generative AI, Ingenuity, and Law," in IEEE Transactions on Technology and Society.

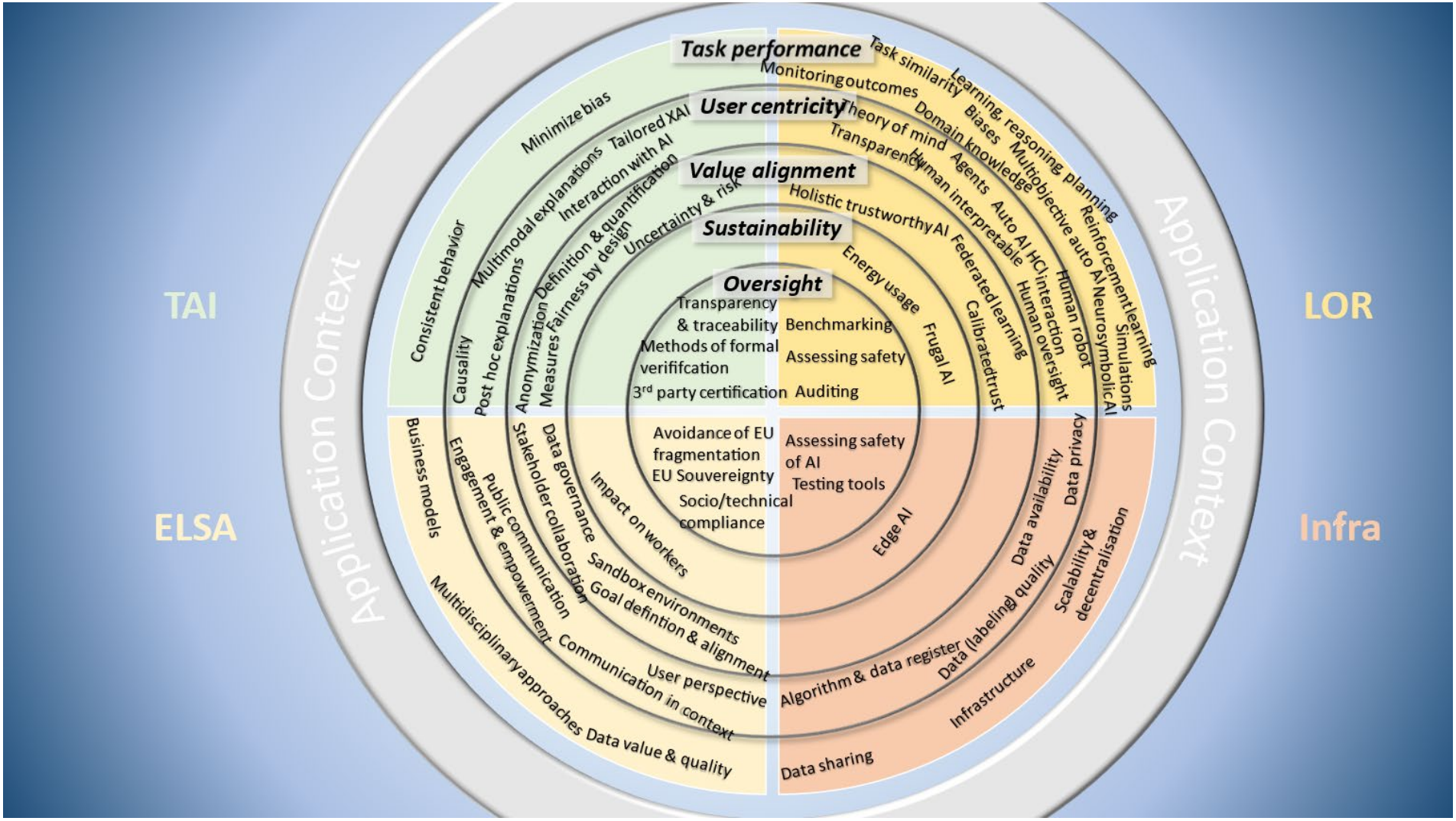


Figure 2: Illustration of the layer – sector map with representative topics projected on the AI development layers in the four topic sectors (TAI, LOR, ELSA, and Infra). Please note that the figure shows a subset of representative topics to achieve maximum readability.

4. Prioritisation based on showcases

The underlying idea behind the AI development layers is that there is a structure in AI system building and dependency between the proposed layers. The specific functional requirements and needs strongly depend on the application context, illustrated in Figure **3** as the light grey band at the outer ring of the figure. In other words, the articulation of specific research topics in individual layers is dependent on the maturity in the layer, and the specific application context.

Following this reasoning, we defined four showcases: impactful use cases with long-term challenging goals that serve as four different application contexts. The list of four showcases with goals, context and intended results for an AI system for that showcase can be found in Annex **2**.

In multiple expert sessions, the TAILOR community used the layer – sector map with the AI development layers in two workshops[12,13] to perform a (preliminary) identification of research topics deemed essential for the individual showcases. The aggregated results can be found in Annex **3**. The research topics that were selected more than once are listed in Table **1**. Clearly, given the limited number of research topics and showcases covered on the one hand, and the room for interpretation due to the limited descriptions provided for each research topic on the other hand, this exercise had its pragmatic limitations. Nonetheless, preliminary trends can be identified on the relevance of individual research topics.

Sector	**Most often mentioned research topics**
TAI	Transparency & traceability; Uncertainty & risk; Fairness by design; Tailored XAI; Minimize bias
LOR	Transparency; Domain knowledge; Federated learning; Learning, reasoning and planning
ELSA	Multidisciplinary approach; Social/technical compliance; User perspective; Goal definition & alignment; Stakeholder collaboration
Infra	Data sharing; Data availability & privacy; Edge AI

*Table **1**: List of research topics selected more than once among the four showcases supported by the TAILOR community. The research topics are ordered from left to right in number of mentions.*

Reflecting upon the (preliminary) outcomes of the exercise and discussing their implications with the TAILOR community strengthened and extended the list of main research topics identified in the context of trustworthy AI; see Annex **1**. Here, it must be noted that neurosymbolic AI as a form of hybrid AI was found essential for multiple showcases and development layers; appropriate attention to this topic in future studies is deemed warranted.

[12] Workshop "TAILOR SRIR – 2nd version"
[13] 4th TAILOR Conference – Trustworthy AI from lab to market

The role of the Fundamental Research

The design choice to focus SRIR V2 on showcase challenges arguably favours those research topics that work towards higher TRLs. Fundamental research, which does not (yet) have a clear application scope, is less prominent.

While it goes without saying that it is important that the SRIR V2 has a strong focus on application – and industry relevance, the role of fundamental, curiosity-driven research is necessary to make long-term impact. In the workshop in Vaals, a group of fundamental research questions was identified that were acknowledged as important to accelerate innovation in terms of trustworthy AI and Learning, Optimizing, and Reasoning, while linking demands from industry and society to curiosity-driven fundamental research areas.

As an exercise the links between these research questions and the research topics of the SRIR was established, as is presented in Table 2 below.

Fundamental research questions	SRIR research topics
Trustworthy Reasoning and Integrating Explicit Knowledge	
How can we use Generative AI - models for trustworthy reasoning with knowledge that pre-exists and is available or has been learned?	#24 Neurosymbolic systems #41 Hybrid AI #48 Trustworthy LLMs
How can such knowledge be used during the learning procedure?	#41 Hybrid AI
How should it be represented?	#48 Trustworthy LLMs
How to assess the reliability of the inference?	#61 Measuring trustworthiness - Trust multidimensional definition
Trustworthy Real-World Interaction, Perception and Planning	
How can we observe, interact with the real-world, plan and take the right actions in a trustworthy manner?	#39 embodiment – this topic addresses some interactions, mainly to improve and enrich human-robot interaction
How can we learn about these low-level actions?	New topic, not yet in SRIR research topics
How can we integrate social reasoning (theory-of-mind reasoning)?	#34 Theory of mind
Trustworthy Autonomous Assistants	
How can we develop interactive tandem systems for specific tasks and domains?	#43 AutoAI HCI
Guarantees for Trustworthy AI	

How to define and verify robustness for generative neural networks?	#125 Guarantees for AI – however specifically for generative neural networks the topic needs to be extended
What can we learn across various runs of a formal verification system (to make future runs more efficient)?	#65 consistent behaviour in certification
How do we obtain formal robustness guarantees for large machine learning models (e.g., the transformers underlying LLMs)?	#66 methodology for formal verification – this topic is needed but must be extended for LLMs
How do we train large machine learning models so that they satisfy stringent robustness guarantees?	This is a new topic that feeds into #8 Verifiability of systems
How can we effectively and efficiently verify / certify robustness properties of large-scale hardware, software and cyber-physical systems?	#11 Formal methods from safety engineering
How can we automate the design of such systems in a way that ensures they have these properties?	This feeds into #125 Guarantees for AI, but it adds the 'automation' aspect to it.
Frugal AI, Frugal ML, Green AI, Sustainability	
How do we achieve the kind of performance seen in current AI systems (e.g., LLMs, but also deep reasoning systems, such as the ones used for hardware and software verification) with substantially reduced resources?	#51 Sustainability (energy usage) of LLMs
Complex Humans - Systems - Society Interaction Modeling, Learning, Reasoning and Controlling	
How do we deal with problems that do not have a well-defined solution, but rather a constant interaction where many different considerations and constraints have to be balanced in a constantly evolving and changing system?	Links to #81 systemic approach & life cycle management, but extends to the design process of AI, and adaptability (self-awareness)
How can we understand different needs and preferences of actors (which can be individuals, groups, organizations, etc.) and how can we find reasonable trade-offs between these?	Links to #62 trustworthiness metrics, but has a much broader scope.

Table 2: Fundamental research linked to roadmap research topics

We can conclude that a number of prominent research questions on fundamental aspects of AI can be linked to a large part to the SRIR research topics. In other words: the SRIR does not only structure application-oriented AI research but is also useful for more fundamental AI

research. This also provides ways to link fundamental research to more application-oriented research.

In some cases, the fundamental research questions could give rise to additional research topics to be included in the SRIR list.

The moonshot

A good example of the merge between fundamental research and application-oriented research can be found in the Moonshot proposal by CLAIRE and euRobotics, titled "Trustworthy, Multicultural Generative AI Systems for Safe Physical Interaction with the Real World"[14]. The relevance of this moonshot for the SRIR V2 is that, due to its relation with the identified research topics, we can use it to link and justify the topics in a single target application: the moonshot itself. Accordingly, we use the moonshot as a backdrop for the roadmap because it is expected to accelerate all other related topics.

Using the outputs in different perspectives

There are multiple ways in which the SRIR V2 roadmap methodology can be used. In this Section, we will outline several of them. Here, we would like to encourage the reader to combine these methods if none is directly applicable to the reader's case. It goes without saying that the roadmap is intended to be versatile and multidimensional.

The first perspective that we consider is that of a **policy maker**. Under the notion that policy makers aim to improve the lives of EU citizens through the identification of EU-wide needs and target policy actions are requires, they require insights into the rules, laws, guidelines - and their implications - that are required needed to foster and empower both the researchers and the parties that implement and use the innovative technology at hand, while, at the same time, keeping the innovative technology and its use safe and responsible. As such, the clear overview of the relevant research topics in the SRIR V2 is relevant for policy makers.

Another perspective that we consider is that of a **project sponsor**, i.e., the person who receives project proposals and decides which of the proposals are granted funding. We consider two kinds of project sponsors: the *domain-specific project sponsor* and *the societal benefit project sponsor*. The first focuses on accelerating progress in a specific domain or even towards a specific challenge, while the latter focuses on accelerating progress that benefits EU residents in general as well as the EU industry. While characterized by distinctive elements, both kinds of project sponsors will, in their own domain, benefit from the insights provided by the identification of relevant research topics in the SRIR V2. The domain-specific project sponsor can follow the outcomes of the most related showcase. Otherwise, if the application context is too different, it is recommended to apply the followed road mapping methodology for that application context. Specific research topics will show up that are linked to the sectors and development layers depending on needs and challenges. The social benefit project sponsor, on the other hand, should consider the aggregated outcomes of the road mapping methodology based on the showcases; see Section **5**.

[14] Moonshot proposal 07.11.23 (for release) (claire-ai.org)

5. Recommendations

The roadmapping exercise for the SRIR V2 led to the selection of a well-defined (sub)set of prioritized research topics. The selected and prioritized set of research topics followed from their perceived importance in the four impactful showcases and the moonshot. The research topics support the development in one or more AI development layers. Below, we summarize these research topics (in bold face) and the development layer(s) (in Italics) they apply to:

1. **Neurosymbolic AI** (LOR) combines data-driven and knowledge-driven methodologies for the embedding of **domain knowledge** (LOR) in various AI maturity development layers: to improve *task performance*, *user-centricity* as well as enabling the embedding of values in models: *value alignment*.
2. **Transparency** (LOR) and **transparency & traceability** (TAI) as listed in the HLEG trustworthy AI requirements need special research attention for *user-centricity* and *oversight*.
3. For the involvement of stakeholders, *user-centricity* and *value alignment,* a **multidisciplinary approach** (ELSA) is needed.
4. Research topics related to learning and deploying on federated data sources, such as **data sharing** (Infra) for *task performance* and **federated learning** (LOR) for privacy, i.e. *value alignment*.
5. **Uncertainty & risk** (TAI) is important as *value alignment* topic, besides for the considered showcases, also for foundation models and generative AI as put forward in the Moonshot.

6. Future developments

The SRIR V2 may, of course, be used by individual partners. The expectation is that further implementation and elaboration of the SRIR will take place in the European context. For this, two links have been identified: towards (1) the AI NoE research mapping (VISION project)[15] and (2) the AI on Demand platform (AI4Europe project)[16].

AI NoE research mapping

A noteworthy development is the EU AI ecosystem map[17], which aims to create an overview of the parties working on AI research and the topics and domains they work on. Currently, the categories listed in this overview are based on a combination of

- The twelve high level categories of AI research as identified by CLAIRE
- The list of keywords used for AAAI (version 2023)[18]

The categorization has been discussed extensively with the NoE community, acknowledging that any categorization involves making choices that are not optimal for everyone. The AI

[15] https://www.vision4ai.eu/ai-ecosystem-mapping/
[16] https://www.ai4europe.eu/
[17] eu-ai-ecosystem.tnods.nl
[18] https://aaai-23.aaai.org/keywords/

Ecosystem mapping has a different categorization because its goals are different. However, it makes sense to link the SRIR v2 research topics to the EU AI Ecosystem map: this may create support for a widely adopted and used categorization of AI research. A harmonized categorization may provide more comprehensive ways to look at the AI research area. For instance, it will be easier to do an analysis of the research on Trustworthy AI based on the SRIR, and compare that to the location and embedding of that research in the European AI community. The EU AI ecosystem map is planned to be hosted on the AioD platform. Its contents will be maintained by a Joint Topic Group in the AI, Data & Robotics Association (Adra): EMIR - Ecosystem Mapping and Information Repository[19]. To explore the link between the SRIR v2 and the Ecosystem map, connections between the Adra Topic Group and the TAILOR consortium have already been established.

AI-on-demand platform

The AI-on-Demand (AIoD) platform, created through the AI4Europe project[20], is a digital platform and experimentation environment that aims to be a one-stop shop for anyone looking for AI knowledge, technology, tools, services, and experts. To goal of this platform is to benefit the entire AI community, including research from academia, students, SMEs and tech providers.

There are two links to the AIoD platform. The first one is that the AIoD platform provides a building block for the development of Trustworthy AI in Europe. As such, it will support a number of topics that are outlined in the TAILOR SRIR. The second link is that the TAILOR SRIR itself could be featured on the AIoD platform, as a resource for AI researchers and policy makers.

AIoD as building block for Trustworthy AI

At the time of writing, the AI on Demand platform offers the following catalogues:

- Organisations
- Projects
- Open calls
- AI Assets
- Research bundles
- Educational resources
- News
- Events
- Case studies

Feedback from the TAILOR community indicates that there is room for extension of these catalogues, for instance with code repositories, datasets, models, papers and educational materials. This may evolve into a combined European alternative for GitHub, OpenML, and arXiv.

[19] Ecosystem Mapping & Information Repository | Adra Association (adr-association.eu)
[20] Home Page | AI-on-Demand (ai4europe.eu)

The biggest condition for success is that the platform should be useful and visible. Endorsement from parties such as IJCAI and ECAI could help with this, as well as commitment from the EU and AI4Europe.

The community believes that when it comes to sharing of datasets and models, it is highly important that parties can have/request data access without the ability to copy directly. This is why it is also suggested to connect the AioD platform to the European data spaces, which may also help bring the data and AI communities closer together. In that way, the AIoD as a building block for Trustworthy AI in Europe becomes even stronger and more relevant.

TAILOR SRIR in the AIoD platform

Since the AIoD evolves into a one-stop shop for AI researchers and developers in Europe, it would make sense to highlight the TAILOR SRIR in the platform as well.

Since the AI Ecosystem mapping (as discussed in the previous section) is already considered to be included in the AIoD platform, the TAILOR SRIR could 'piggyback' on that. A prerequisite for this would be to establish an explicit link between the research topics in the TAILOR SRIR, and the categorization of AI research that is proposed in the AI Ecosystem mapping.

7. Conclusions

The first version of the Strategic Research and Innovation Roadmap (SRIR V1) outlined the landscape of research topics for AI innovation considering the essential principles of *Trustworthy AI* (TAI) and methodologies based on *Learning, Optimization and Reasoning* (LOR). The second version of the SRIR, i.e., SRIR V2, aims to add orderings and prioritizations to the first version, in order to create a guiding roadmap in the research landscape of trustworthy AI.

To facilitate the mapping procedure, we added two additional categories of research topics to TAI and LOR that are deemed pivotal: *Ethical, Legal and Societal Aspects* (ELSA) and *Data and Infrastructure* (Infra). We additionally proposed five development layers derived from the historical stages in developments in AI research: *task performance*, *user centricity*, *value alignment*, *sustainability*, and *oversight*.

Subsequently, we introduced four impactful showcases that are characterized by their long-term challenging goals, to serve as an application context. Based on these showcases, the TAILOR community mapped out the road for the SRIR V2. As a result of the structured creation of SRIR V2, it became apparent that several research topics were recognized as highly relevant, and prioritized accordingly:

Neurosymbolic AI, *domain knowledge*, *transparency*, *multidisciplinary approach*, *data sharing*, *federated learning*, *uncertainty & risk*.

For specific application domains, additional research topics may gain importance, while the relevance of others may be reduced. As such, it is recommended to apply the same roadmapping methodology followed by the TAILOR community to identify the relevant research topics accordingly for the application domain at hand.

8. Annex 1: Research topics

This Annex presents the extended list with research topics used to define SRIR V2. For each research topic, the following information is provided:

ID	A unique number that uniquely identifies the research topics
Short name	A short name for the research topic
Origin	Where the research topic comes from: - SRIR v1 - TAI+LOR workshop (TAILOR WS) - Theme Development Workshop (TDWS)
Description	A bit more elaborate description of the research. The underlying 'full' descriptions are available and are derived directly from the sources.

ID	Short name	Origin	Description
2	XAI - Transparency	SRIR v1	Transparency by design
3	XAI - post hoc	SRIR v1	post hoc explanations
4	XAI - human interpretable	SRIR v1	human interpretable formalisms
5	XAI - Multimodal	SRIR v1	Generating multimodal explanations
6	XAI - Causality	SRIR v1	Causality
7	XAI - Engagement	SRIR v1	Empowering and engaging people
8	Safety & robustness - Verifiability	SRIR v1	Verifiability of systems
9	Safety & robustness - Failure behaviour	SRIR v1	Capability profiling, calibration of failure behaviour
10	Safety & robustness - Benchmarks	SRIR v1	Robustness benchmarks
11	Safety & robustness - Formal methods	SRIR v1	Formal methods from safety engineering
12	Fairness - by design	SRIR v1	Fairness by design
13	Fairness - Legal	SRIR v1	Alignment with legal requirements
14	Fairness - Socio-technical compliance	SRIR v1	Compliance in socio-technical systems
15	Accountability - Formalize fairness	SRIR v1	Formalize fairness objectives and mitigate biases
16	Accountability - Measures	SRIR v1	Scientific and methodological measures, quality standards and procedures to better model the development process of learning methods
17	Privacy - Risks	SRIR v1	defining formally and detecting automatically privacy risks raised by AI systems handling different kinds of personal data
18	Privacy - Anonymization	SRIR v1	designing data anonymisation and attribute hiding algorithms that are robust to sophisticated attacks
19	Privacy - Privacy by design	SRIR v1	designing AI algorithms that respect by design privacy constraints
20	Privacy - balancing privacy & utility	SRIR v1	balancing privacy with utility, fairness, interpretability
21	Privacy - Data subject sovereignty	SRIR v1	control by data subjects, consent management
22	Sustainability - Infrastructure	SRIR v1	Availability of infrastructure to work on 'AI for good'
23	Sustainability - Frugal AI	SRIR v1	Reduce carbon footprint of AI: frugal AI
24	Paradigms - Neuro-symbolic systems	SRIR v1	Design of neural symbolic systems (incorporate rules in neural networks; constraint programming; knowledge graphs from texts)
25	Paradigms - Innovation adoption	SRIR v1	Innovation adoption: AI knowledge in SMEs and startups; lack of trust; lack of quality data; scalability; academic prototypes to operational systems
26	Deciding - Human Oversight	SRIR v1	Autonomous behaviour guided by human oversight
27	Deciding - Safety	SRIR v1	Assessing safety and formal verification, model checking and automated synthesis to guarantee safety specifications
28	Deciding - Acting	SRIR v1	Reasoning and planning for acting; learning strategies/plans from data; learning heuristics for planning

29	Deciding - Learning reasoning planning	SRIR v1	Learning models from data, and then do reasoning and planning
30	Deciding - Learning from experiences	SRIR v1	Learning from past experiences and simulations, for refining strategies/plans or models
31	Deciding - Monitoring outcomes	SRIR v1	Monitoring the actual outcome of actions; recognizing possibly unexpected outcomes; reasoning, planning and learning how to deal with unexpected outcomes
32	Deciding - Capability self-assessment	SRIR v1	Self-assessment of capabilities; adjustable autonomy
33	Deciding - Benchmark behaviour	SRIR v1	Benchmarks to evaluate behaviour and performance
34	Social contexts - agents	SRIR v1	how to empower individual AI agents to communicate with each other, collaborate, negotiate and reach agreements/consensus and how they coordinate to fairly share common resources, and how they differentiate to accomplish collaborative tasks together
35	Social contexts - Theory of mind	SRIR v1	theory of mind models
36	Social contexts - simulations	SRIR v1	agent-based social simulations
37	Social contexts - multi-agent	SRIR v1	Multi-agent learning; including privacy
38	Social contexts - domain knowledge	SRIR v1	Model domain knowledge for agents in social contexts
39	Social contexts - embodiment	SRIR v1	Using embodiment features and social clues to enrich interaction
40	Social contexts - calibrated trust	SRIR v1	Explainability of social multi agent systems; calibrated trust
41	Automated AI - Hybrid AI	SRIR v1	AutoAI for Hybrid AI Systems
42	Automated AI - Task similarity exploitation	SRIR v1	Task similarity to efficiently explore which algorithms may work well on a new dataset (or task)
43	Automated AI - AutoAI HCI	SRIR v1	combining AutoAI techniques with HCI, hybrid between human and automated processes
44	Automated AI - multi objective AutoAI	SRIR v1	Robust, efficient and multi-objective AutoAI is needed to enable real-world AI applications
45	Automated AI - Meta-learning	SRIR v1	Leveraging multi-task- and meta-learning; learning which algorithms and configurations work well for an application domain
46	Automated AI - Benchmarking AutoAI	SRIR v1	Benchmarking AutoAI
47	Foundation models - Biases	SRIR v1	Biases in large models
48	Foundation models - Trustworthiness of LLMs	SRIR v1	Trustworthiness of LLMs; integration of LLMs with knowledge representation and reasoning; reproducibility, accuracy
49	Foundation models - EU Sovereignty	SRIR v1	Sovereignty: EU LLMs, including infrastructure
50	Foundation models - LLM Explainability	SRIR v1	Explainability of LLMs
51	Foundation models - LLM energy usage	SRIR v1	Sustainability (energy usage) of LLMs
52	Foundation models - Avoid EU fragmentation	SRIR v1	Avoid fragmentation, bundling forces in Europe
53	Make LOR more trustworthy - Transparency and Traceability	TAILOR WS	Data Transparency and Traceability: Need to be explicit about data used and their sources, as well as provide governance
54	Make LOR more trustworthy - Audit Checklist Development	TAILOR WS	Audit Checklist Development: Develop a checklist with relevant attributes and process for auditing
55	Make LOR more trustworthy - Better understanding of LOR	TAILOR WS	Enhancing LOR: Better understanding of large LOR-based systems; Check that the optimisation objective corresponds with what the user really cares about; modularizationof LOR components; definition of objectives for optimization
56	Make LOR more trustworthy - Risks	TAILOR WS	Risks: LOR-based systems can become very complex when there are many different rules, even specialists struggle to do so
57	Achieve trustworthiness by using LOR - user understanding	TAILOR WS	Enhancing Scientifc Results: use LOR to gain insight in beliefs and goals of the user so that the system can optimize better and explain better
58	Achieve trustworthiness by using LOR - Formal proofs	TAILOR WS	Dissemination: invest in communication & awareness; links with formal proofs community

59	Achieve trustworthiness by using LOR - Other reasons for distrust	TAILOR WS	Risk: trustworthiness does not only come from LOR per se (eg, also user interface decisions infuence trustworthiness); fake news and malevolent intentions can destroy trust
60	Measuring trustworthiness - Multidisciplinary research	TAILOR WS	Multidisciplinary research - social/legal: Psychological research is needed to better understand the concept 'trustworthiness' and align it with our technical measures; evaluation in social experiments; co-design with end-users
61	Measuring trustworthiness - Trust multidimensional definition	TAILOR WS	Trust definition: the concept 'trustworthiness' encompasses different other concepts (fairness, accuracy, ...) and we should adopt a common understanding of this
62	Measuring trustworthiness - Metrics misfits	TAILOR WS	Challenge/risk: how to balance trustworthiness dimensions; changing (societal) perceptions of trust; metrics risk to become the goal (‘malicious overfitting’)
63	Trustworthiness certification - testing tools	TAILOR WS	Testing tools, reputation mechanism, risk assessment, standards: trustworthy AI dimensions experimentation playgrounds to measure if and to which extent a tool is compliant with a given TAI dimension, to provide a kind of "certifcation"; multiple user groups needed; link to regulatory sandboxes; formal verification
64	Trustworthiness certification - third party certification	TAILOR WS	Third-party certifcation: inspiration from food labelling; build upon existing sectorial approaches
65	Trustworthiness certification - consistent behaviour	TAILOR WS	Definition of trust: Trust as consistency, explainability, reliability, AI act: Trust is related to expectations. We generate trust by means of consistent behaviour; link to requirements in AI act; simple explanations; also address mistrust and distrust
66	Trustworthiness certification - methodology for formal verification	TAILOR WS	Technological challenge: Experimental evaluation is possible (e.g., simulation), but doesn't guarantee formal verifcation, which is very difcult. Rare or unexpected cases may be overseen; link to social sciences
67	Trustworthiness certification - Standardization	TAILOR WS	Standardization bodies (de jure) and startups/companies already proposing labels (de facto).
68	Trustworthiness certification - User perspective	TAILOR WS	Challenge in trust definition: labels can be bypassed; certification alone does not guarantee trust; skill level of user is also relevant (context, usage scenario)
69	Public private research - Multidisciplinary	TAILOR WS	Multidisciplinary approach: Big tech can basically hire all disciplines they need (they do) and science is often divided in silos. Interdisciplinarity is badly needed, and inclusion of soft skills; Communication and social science are crucial
70	Public private research - trustworthiness beyond compliance	TAILOR WS	challenge: when we can include trustworthiness in a business model, then commercial companies will be quick to pick up (‘responsible AI beyond compliance’)
71	Actions & priorities - Definition and quantification	TAILOR WS	Definition and quantification: quantify axes of trustworthiness; define accountability;
72	Actions & priorities - Guidelines and best practices	TAILOR WS	Guidelines: describe best practices
73	Actions & priorities - awareness & dissemination	TAILOR WS	Awareness, dissemination: engage in dialogue with the public; make companies accept the need for certifcation and make governments follow the rule of law according to the AI Act; there is already a lot of misunderstanding ('the robots are taking over') so communication may backfire; how to deal with conspiracy theories?
74	Actions & priorities - Continuous revision & flexibility	TAILOR WS	Continuous revision & flexibility: revise current insights in the context of new developments (eg, generative AI – could LOR make LLMs more trustworthy?)
75	Public sector - education	TDWS	Education on AI for employees; understanding, expectations, acceptance
76	Public sector - AI ecosystems performance	TDWS	Measure performance of AI ecosystems
77	Public sector - Algorithm register	TDWS	Algorithm register
78	Public sector - Procurement and market creation	TDWS	Procurement and market creation
79	TDW - data sharing	TDWS	Data: overcome information silos, governance and technology for data sharing to promote availability, quality and accessibility of data
80	TDW - AI requirements & certification	TDWS	Requirements for AI: make requirements concrete while maintaining overall view; certification
81	TDW - systemic approach & LCM	TDWS	Systemic approach and life-cycle management: procurement, AI system performance monitoring, end-of-life policies and procedures
82	Mobility - holistic trustworthy AI	TDWS	Grasp Trustworthy AI holistically; many dimensions/aspects of trustworthy AI are connected and influence each other

83	Mobility - data sharing	TDWS	Data is key: available databases, better standards, need for a common big data pool, redefine rules for certification, data transparency towards users
84	Mobility - public communication	TDWS	Communication with the audience
85	TDW - public communication	TDWS	Trustworthy AI and Explainable AI: media coverage, explain terms better (eg, differences between trustworthy and explainable), knowledge management in the AI community
86	TDW - data value & quality	TDWS	Data: value of data for customers; bias
87	TDW - AI training & education	TDWS	Academia and education: more AI training needed; attract more interdisciplinary AI researchers; exchange between industry and academia to help researchers gain practical experience
88	Health - data governance	TDWS	Ownership of Health Data needs (global/EU) privacy by design governance guidance for all involved multi-stakeholders based on the new data economy principles
89	Health - data availability	TDWS	Availability of public data sets difficult due to patient data security
90	Health - infodemics	TDWS	Developing trustworthy AI tools is key to fight infodemics
91	Health - legal & privacy	TDWS	Legal question of data security and anonymity -> how to guarantee these when using federated learning? -> legal framework required
92	Health - XAI & user types	TDWS	Explainability is often not very relevant for the end user (i.e. patient, clinician, careprofessionals), as long as end user benefits from it
93	Health - XAI need	TDWS	Industrial stakeholders are not necessarily interested in creating better (more explainable) models and optimising the models
94	Health - data sharing	TDWS	Need advocacy on public healthcare systems and organisations to create funding opportunities and to support clinical data collection and sharing.
95	TDW - definitions	TDWS	Trustworthy & Explainable AI are closely linked – more precise definitions needed for both
96	TDW - education	TDWS	AI competency courses in middle & high school to make it easier for students to take AI related courses at university – increasing the number of people with expert AI knowledge significantly
97	TDW - standards	TDWS	Incentives for the participation in standards development needed (Especially for academics and SMEs
98	Energy - energy efficiency through AI	TDWS	Energy efficiency using AI: Focus on achieving improved energy efficiency through AI-enabled software solutions. This includes energy consumption anomaly detection and time-series forecasting.
99	Energy - ML for optimization of usage	TDWS	Machine learning applications: Explore the application of machine learning solutions for energy optimization in various sectors such as transportation (land, air, and sea), heavy industries, heating and ventilation, data centers, and energy storage.
100	Energy - business models	TDWS	Business models and incentives: Develop suitable business models that align with conflicting objectives such as carbon neutrality and cost minimization to promote the adoption of AI solutions for energy efficiency.
101	Energy - cross-domain collaboration	TDWS	Collaboration and cross-domain innovation: Foster collaboration among stakeholders, including universities, automation suppliers, and industry end-users, to combine machine learning and optimization approaches for energy efficiency.
102	Energy - XAI	TDWS	Explainable AI: Address the importance of explainability in AI solutions for the energy sector to gain trust and acceptance from decision-makers and regulatory authorities. Develop tools and frameworks for explainable AI models.
103	Energy - integration of energy systems	TDWS	Integration of energy systems: Focus on the integration of energy management and distributed smart building systems to achieve a resilient grid infrastructure powered by renewable energy sources. Emphasize the partnership between humans and smart energy systems.
104	Energy - data quality & availability	TDWS	Data quality and availability: Overcome challenges related to data availability, quality, and accessibility for effective AI implementation in energy systems. Explore data-driven approaches for decision-making and optimization.
105	Energy - scalability & decentralisation	TDWS	Scalability and decentralization: Consider scalability as a crucial factor for AI solutions in energy systems, especially when dealing with large and interconnected systems. Explore the potential of edge computing and decentralized coordination mechanisms.
106	Energy - interdisciplinary research	TDWS	Interdisciplinary research and knowledge transfer: Encourage interdisciplinary collaboration and knowledge transfer among experts in various fields to address the complex challenges at the intersection of energy and AI.
107	Energy - AI for EV grid integration	TDWS	Integration of electric vehicles (EVs) into the grid: Leverage AI models to manage the integration of EVs into the grid effectively. Focus on predicting user behavior, optimizing charging infrastructure, and using EV batteries as energy reservoirs for grid stability and flexibility.
108	TDW - XAI	TDWS	Explainable AI and Trustworthiness: There is a need for AI models and systems to be explainable, transparent, and trustworthy is of importance across all domains. It is crucial to address the black box effect of AI models and develop tools and methods for explainable artificial intelligence. This involves considering interpretability at different levels, including feature interpretability, model interpretability, and decision interpretability.
109	TDW - Human AI collaboration	TDWS	Human-AI Collaboration and Interaction: The concept of human-AI ecosystems and collaborative sustainable buildings highlights the importance of effective interaction and collaboration between humans and machines. AI should be designed to work

			together with humans as collaborators in teams, aiming for shared objectives and mutual support. The field of human-machine interaction, including preference elicitation, aggregation, and understanding human requirements, plays a significant role in achieving successful human-AI collaboration.
110	TDW - data quality & privacy	TDWS	Data Quality, Accessibility, and Privacy: There is a need for high-quality data, structured data acquisition, and semantic dataspaces for facilitating data communication. Additionally, privacy concerns regarding the collection, processing, and sharing of data as well the importance of data anonymization and local processing should be taken into account.
111	TDW - optimization and decision-making with AI	TDWS	Optimization and Decision-Making: AI's role in optimization and decision-making processes is important for AI in the energy sector particularly in the context of energy management, EV charging, and industrial generation scheduling. AI can provide smart solutions for optimizing energy consumption, designing efficient systems, and achieving global optimization
112	Disinfo - user-centric XAI	TDWS	Users should be at the core of the development of future XAI tools
113	Disinfo - communication in context	TDWS	Communication surrounding AI should not only focus on solely technical or factual information but should be contextualised in the current political, ethical, and moral context
114	Disinfo - public communication	TDWS	Further measures to reach the public regarding AI and its capabilities and limits need to be developed and tested
115	Disinfo - goal definition & alignment	TDWS	Abusive language detection and automated moderation may be misaligned in target constructs and broad goals
116	Disinfo - transparent moderation	TDWS	AI should be able to make moderating processes more transparent for users on social media platforms
117	Disinfo - data sharing	TDWS	The issue of data sharing at larger scale remains very challenging due to tension between technical, ethical, and regulatory aspects associated with it
118	Disinfo - data governance	TDWS	Clarification is needed regarding the data governance rules in order to reduce hesitation to share data on researchers' side
119	TDW - adaptive XAI	TDWS	It is incredibly important to make explanations of the behaviour of AI as flexible as possible to cater for the needs of users
120	TDW - minimize bias	TDWS	It should always be a priority to utilise models which minimise bias by design and properly integrate the uncertainty of predictions used
121	TDW - user-centric XAI	TDWS	Users should be at the core of the development of future XAI tools
122	TDW - cross-disciplinary research	TDWS	Tighter collaboration between computer scientists and legal researchers and lawyers is needed
123	TDW - data sharing infra	TDWS	Creation of a large European-wide infrastructure enabling access to large datasets would be beneficial
124	TDW - sandbox	TDWS	Set-up of sandbox environments to facilitate specific types of research is recommended
125	Manufacturing - guarantees for AI	TDWS	Fundamental research is needed between academy and industry into methods and their application to provide guarantees about AI systems. Furthermore there is a need for research into the different dimensions of trust, for example, robustness with respect to changing work conditions, interpretability as a means for true human-machine collaboration), and verification that the AI fits the intended purpose.
126	Manufacturing - federated learning	TDWS	By processing training sets in parallel Federated Learning could be more efficient and scalable for large-scale manufacturing ecosystems. Federated Learning also enables Privacy by Design by training partial data sets separately and combining only the resulting models without sharing the data.
127	Manufacturing - Human Robot collaboration	TDWS	Human-robot collaboration (HRC) can help integrate humans into the production process without replacing them by relieving them of burdensome tasks and highlighting their strengths such as flexibility, experience and understanding.
128	Manufacturing - practical constraints	TDWS	Products are in many cases constrained by restrictions, which also pose difficulties. In this case an evolving Digital Twin - evolved during the use of the product - can also enhance the knowledge about constraint, not just on the engineer side but also on the user side.
129	Manufacturing - verification of trustworthiness	TDWS	The trustworthiness of systems deployed in a spacecraft or space station must be extremely high and must be verified independently from the vendors who provide the various parts of the system.
130	Manufacturing - data labeling quality	TDWS	When it comes to data labelling, sustainable approaches aim to combine psychology with elements of labelling technologies to ensure high data quality (standardisation) which is needed to apply AI in manufacturing. However, data often appears messy, because they are not labelled accurately due to the lack of motivation of human annotators. These operators should be involved more to understand the reasons for labelling the data.
131	Manufacturing - data availability	TDWS	The availability of training data is the main problem of using deep learning methods. E.g. in a production environment, most data will show undamaged parts while actual defects are rare. A solution to this is to simulate measurements based on scenes that are generated by parametric models or the real world. By investigating the parameter space of such models, training data can be generated in a controlled way.
132	TDW - tailored XAI	TDWS	Explainable AI by design as a key to increase trust and most probable more reliable systems. Therefore, meaningful, and tailored explanation to different users and stakeholders is very important.

133	TDW - interaction with AI	TDWS	Users and experts should be able to interact with the AI. Robust and model agnostic explanation methods are required to leverage the full breadth of available AI methods and models, to create trust or interpret the model.
134	TDW - reinforcement learning	TDWS	Reinforcement learning (RL) can help to solve an optimization problem in the absence of a model, i.e. via learning from experience.
135	TDW - uncertainty & risk	TDWS	AI could help to deliver some form of risk certainty management, including knowledge about the uncertainty of certain constraints.
136	TDW - stakeholder collaboration	TDWS	Joint labs with the industry should be established for a better understanding and to create teams that can carry out the implementations required by the industry.
137	TDW - impact on workers	TDWS	Acceptance (explainability) and inclusion (controllability) of workers to take away the fear, including an demographic worker structure is essential for a successful implementation and further usage.
138	TDW - risks of AI	TDWS	Dynamic risk management should be considered when implementing AI.
139	TDW - edge AI	TDWS	The combination of Edge Computing and Edge AI together with high-performance communication technologies such as 5G will act as an enabler for industrial process improvements

Note that there is no topic 1, resulting from the way the underlying Excel was created and the (otherwise meaningless) id numbers were assigned.

9. Annex 2: Showcases

Showcase 1: Health Support

- Assistant for a long and healthy life
  - Supports patients to shape their behavior to prevent the development of diseases, to reverse already established diseases, or enable them to live with a disease in a comfortable way.
- Application context
  - The system exploits historical data from other patients and expert knowledge.
  - The data is stored at federated locations, and has different confidentiality restrictions within and between these federated locations. The data is available in different formats.
- Output
  - The system generates advice for patients that is aligned with the patient's characteristics, their goals, their physical and mental opportunities and limitations, and is based on the latest scientific knowledge.

Showcase 2: Energy Community

- Energy Prosumer Community Planner
  - Provides a plan for power consumption for individuals in a community supported by a dynamic tariffs plan.
- Application context
  - The system exploits current and historical data from community members
  - The data is stored at federated locations with confidentiality restrictions.
  - Community members can have different priorities and preferences
- Output
  - Energy consumption plan per member including differentiated tariff structure

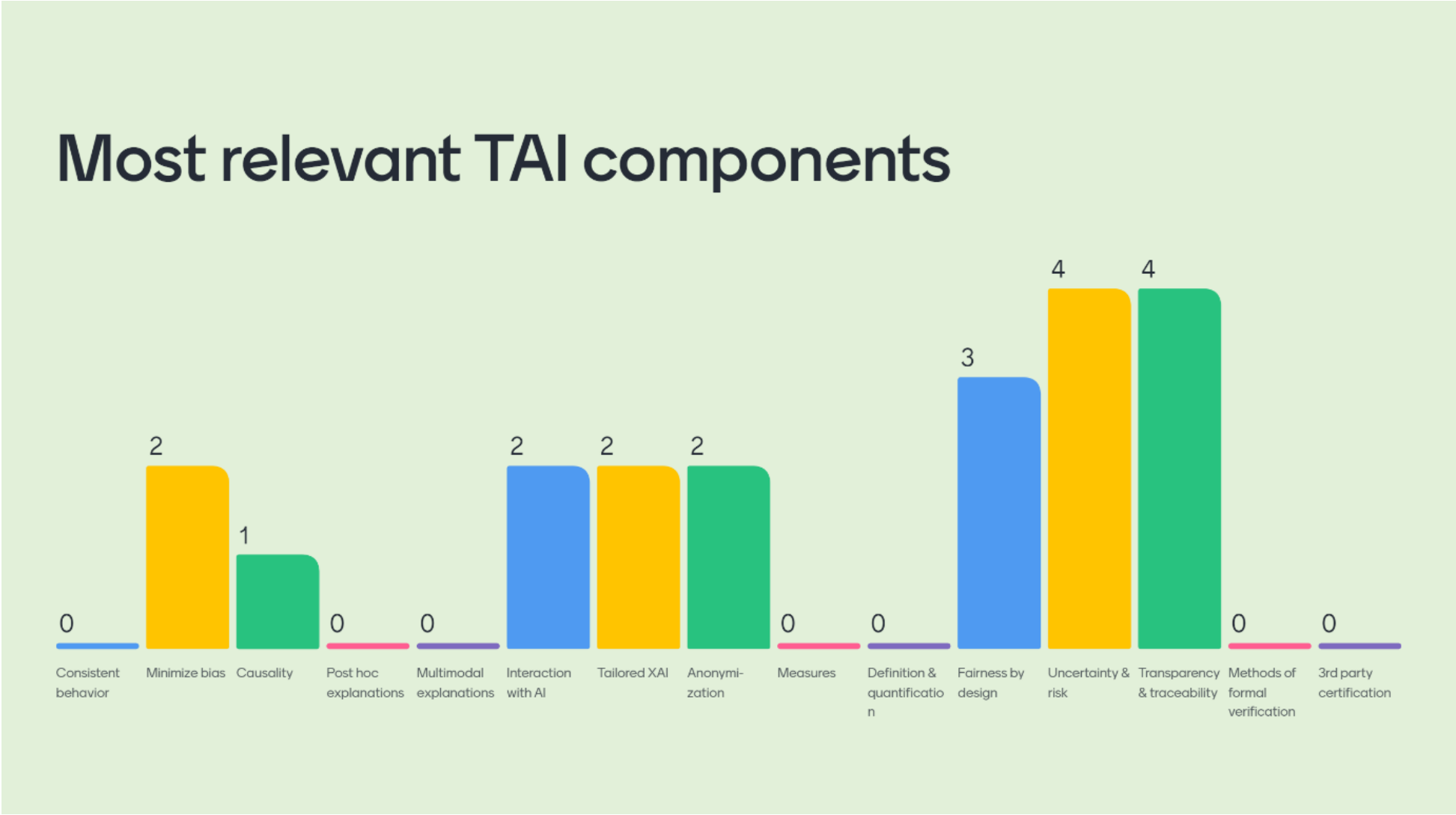
Showcase 3: Logistics

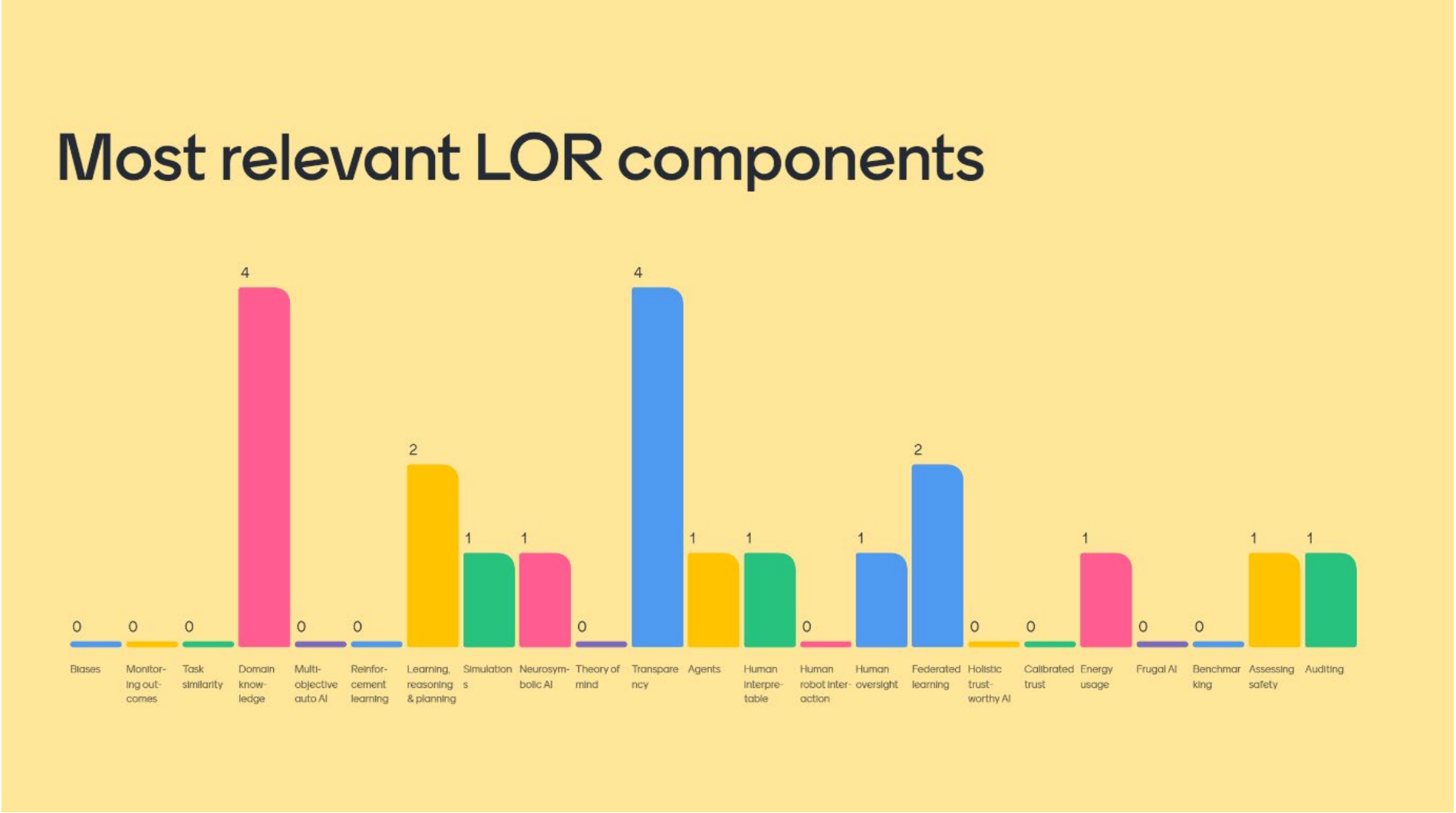
- Optimal Planning for Multiple Collaborating Parcel Services
  - Optimizes the planning for the vehicle fleets of multiple parcel services in a transparent and fair way, while incorporating external factors (such as traffic delays) that require adaptations.
- Application context
  - The data of the parcel services are confidential and at federated locations.
  - The parcel services can have different financial, enviromental and consumer friendliness criterions
- Output
  - Dynamic real-time planning for multiple parcel service fleets

Showcase 4: City Design

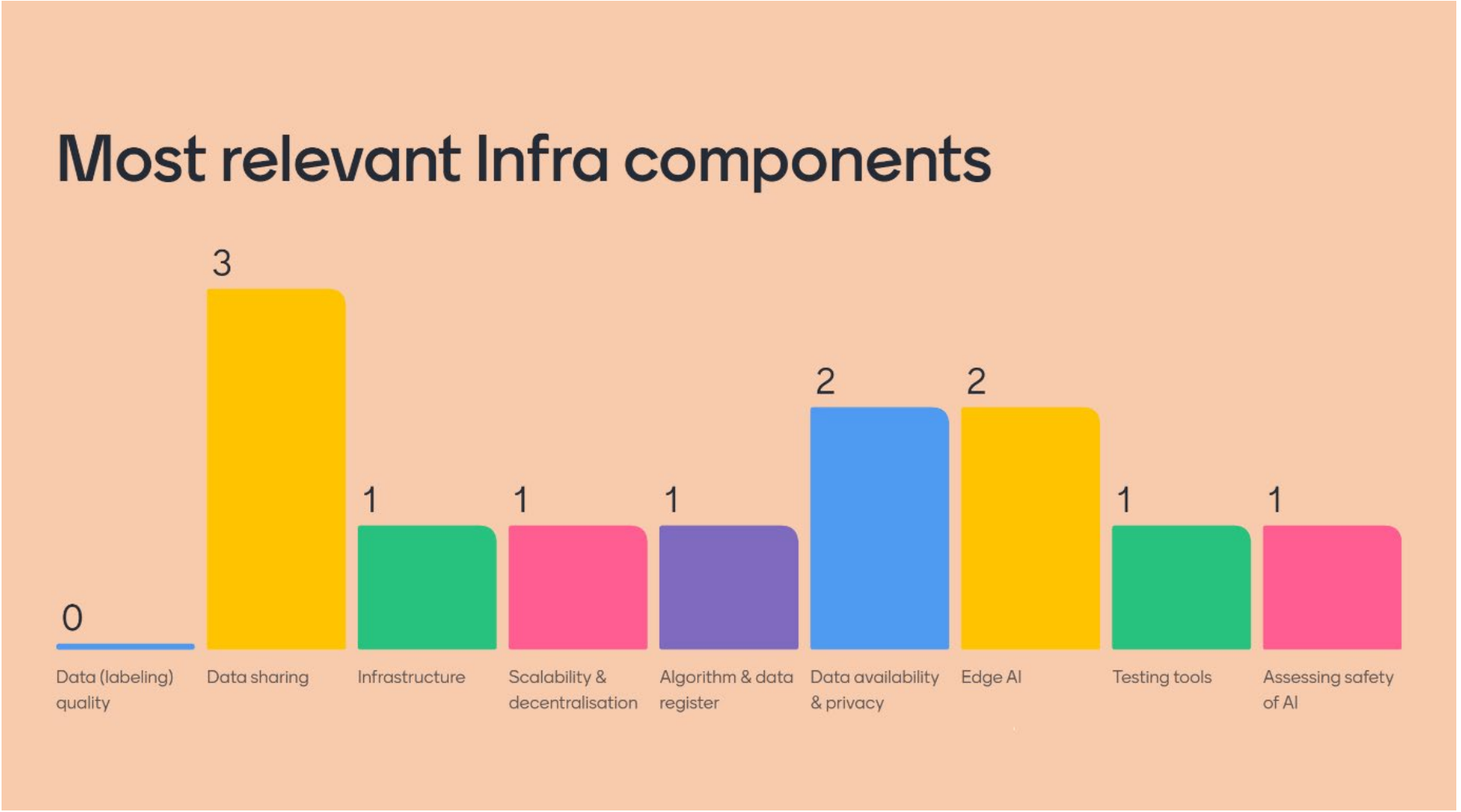
- Housing planning
  - Plans optimal housing locations including impact on other locations
- Application context
  - Data about existing environment (population, infrastructure, mobility options), while data may be incompatible, missing or confidential.
  - Expected demographic developments, needs, and economical conditions.
- Output
  - A map of the toal area with per location a plan with report in case the location would be suitable.

10. Annex 3: Research topic ranking per sector




Most relevant LOR components
0
0
0
4
0
0
2
1
1
0
4
1
1
0
1
2
0
0
1
0
0
1
1
Biases
Monitor-ing out-comes
Task similarity
Domain know-ledge
Multi-objective auto AI
Reinfor-cement learning
Learning, reasoning & planning
Simulation s
Neurosym-bolic AI
Theory of mind
Transpare ncy
Agents
Human interpre-table
Human robot inter-action
Human oversight
Federated learning
Holistic trust-worthy AI
Calibrated trust
Energy usage
Frugal AI
Benchmar king
Assessing safety
Auditing


Most relevant ELSA components
1
4
1
1
0
1
2
1
2
2
0
1
0
1
3
Business models
Multidisciplinary aproach
Data value & quality
Public communication
Communication in context
Engagement & empowerment
User perspective
Data governance
Goal defintion & alignment
Stakeholder collaboration
Sandbox environments
Impact on workers
Avoidance of EU fragmentation
EU Souvereignty
Socio/technical compliance


Most relevant Infra components
0
3
1
1
1
2
2
1
1
Data (labeling) quality
Data sharing
Infrastructure
Scalability & decentralisation
Algorithm & data register
Data availability & privacy
Edge AI
Testing tools
Assessing safety of AI