2024 SUMMER WORKSHOP

June 4-6, 2024 | Washington, DC

SYSTEMS ENGINEERING RESEARCH CENTER

ESI
Powered by industry, academia and TNO

DLR

TECOSA

Hosted by GW

Archimedes Workshop 2024 Workshop Report

The third annual Archimedes Initiative summer workshop, hosted by the Systems Engineering Research Center (SERC), gathered leaders from four prominent systems engineering organizations to discuss trustworthy AI, systems thinking, and other dynamic topics.

The Archimedes Initiative was co-founded in 2022 with the SERC; the German Aerospace Center (DLR) Institute of Systems Engineering for Future Mobility; the Netherlands Organization for Applied Scientific Research (TNO-ESI) Embedded Systems Innovation Center; and the Center for Trustworthy Edge Computing Systems and Applications (TECoSA) at the Royal Institute of Technology in Sweden. The international collaborators complement each other with a presence and expertise in different domains: DLR in automotive and maritime mobility, TNO-ESI in high-tech equipment, TECoSA in automobiles and aircrafts, and the SERC in defense.

The four-day gathering at The George Washington University, a SERC collaborator university, welcomed engineers from the partner organizations alongside guests from SERC sponsors in the Department of Defense and colleagues from the defense industry.

The first day featured presentations from the Archimedes partners, each focusing on the topic of trustworthy AI. Axel Hahn of DLR presented on “Trustworthy Vehicles: Let’s Learn from Live.” Freek Bomhof of TNO spoke on “Trustworthy AI as a Requirement in Applied Research Prototypes.” Martin Törngren of TECoSA discussed “Trustworthy AI Based Cyber-Physical Systems – the Good, the Bad and the Real Challenge.” SERC Chief Scientist Zoe Szajnfarber of The George Washington University (GWU) concluded with “Opportunities for Systems Engineering in the Trustworthy AI.” Short summary papers for each of these talks are included in this report.

The first day featured keynote remarks from Martin Stanley, who is the Emerging Technology Branch Chief at the Cybersecurity Infrastructure Security Agency (CISA) and represented the National Institute of Standards and Technology (NIST), and Ron Keesing, who is the Senior Vice President for Technology Integration at Leidos, as well as technical talks from representatives of the NIST-NSF Trustworthy AI in Law and Society (TRAILS) Institute that GWU co-leads with the University of Maryland.

Part 1 of this report presents four short papers summarizing each of the partner topics.

On the second and third days, SERC CTO Tom McDermott and researcher Dennis Folds led a workshop designed to extract research ideas from all participants as well as provide training on applied systems thinking tools. The workshop was based on a professional development course titled “Applied Systems Thinking and Decision Making: Chief Engineers Seminar Series,” which is a regular part of the systems engineering curriculum for the Federal Aviation Administration.

The theme of this workshop was “systems of trust.” The workshop started with a set of definitions of trustworthy AI captured from the presentations on the first day. The outcome of the workshop developed the initial focus areas for a strategic plan for trusted AI and autonomy research in the near- and long-term. The group defined the critical research activities necessary to accelerate the deployment of fully automated and trusted urban transportation systems.

Part 2 of this report lays out the context as a set of definitions of trust developed from the workshop presentations. Part 3 outlines a set of near-term, mid-term, and long-term strategies and research needs developed in the workshop.

DLR will host the workshop next summer. The Archimedes Initiative will meet in Oldenburg, Germany, for a workshop titled ‘Systems Engineering for the System Lifecycle – Rapid and Efficient System Evolution with Humans in the Loop.’ There is the need to evolve future systems in the field. DevOps cycle, continuous engineering, and interdisciplinary contributions require new approaches for systems engineering.”

Table of Contents

Archimedes Workshop 2024 Workshop Report 2
Part 1. Archimedes Partner Papers on Trustworthy AI 5
DLR: 5
Introduction and motivation 5
Relevance of trust and trustworthiness for systems engineering 5
Key ingredients for trust 7
Trust and Explainability 8
TNO-ESI: 9
Trustworthy AI, according to the EU 9
Workshop Presentation: Approaches to Trustworthy AI in Applied Research 9
Trustworthy AI at ESI: Definition and Future Research Directions 10
Conclusion 13
TECOSA: 14
Defining trust 14
Introducing trust in SE 14
SE for trust in future systems 15
SERC: 17
SE, AI, and Trust 17
AI for Systems Engineering (AI4SE) 17
Systems Engineering for AI (SE4AI) 17
Trust: Socio-technical systems as the unit of analysis for TEV&V 18
Part 2. Defining Trust, from the Workshop Presentations 19
Definitions of Trust 20
Concept of Trust in organizational leadership 20
Concept of Trust in computing systems 21
Definitional attributes of trust in AI/ML: 22
Part 3: A Roadmap for Trustworthy AI, from the Workshop 23
Horizon 1 (H1) 23
Horizon 3 (H3) 24
Horizon 2 (H2) 26
Summary 27

Part 1. Archimedes Partner Papers on Trustworthy AI

DLR:

This is summarized from Hoppe I, Hagemann W, Stierand I, Hahn A, Bolles A. Challenges for trustworthy autonomous vehicles: Let us learn from life. Systems Engineering. 2024;27:789–800. https://doi.org/10.1002/sys.21744. Italicized passages are directly from the paper.

Introduction and motivation

Why is it widely accepted that people get their driver's license after a very limited set of driving lessons—for example, nine hours in Germany? Is it the human intelligence and the related behavior we trust in, so that the corresponding risks are mitigated? The situation is quite different for trust in autonomous vehicles: Empirical studies demonstrate rather low individual trust in autonomous driving and "indicate [. . .] that a substantial proportion of people remain very sceptical about the idea of travelling in a fully automated vehicle". In Germany, for example, around half of the participants in a quantitative market survey (n = 1000) indicated that they would not use an autonomous vehicle (e.g., for public transport) at all, or were unsure. In the United States, the American Automobile Association found that 85% of surveyed Americans (n = 1107) were "fearful or unsure of self-driving technology". This finding is independent of the car model tested, and the level of non-acceptance has remained steady over the past few years.

Relevance of trust and trustworthiness for systems engineering

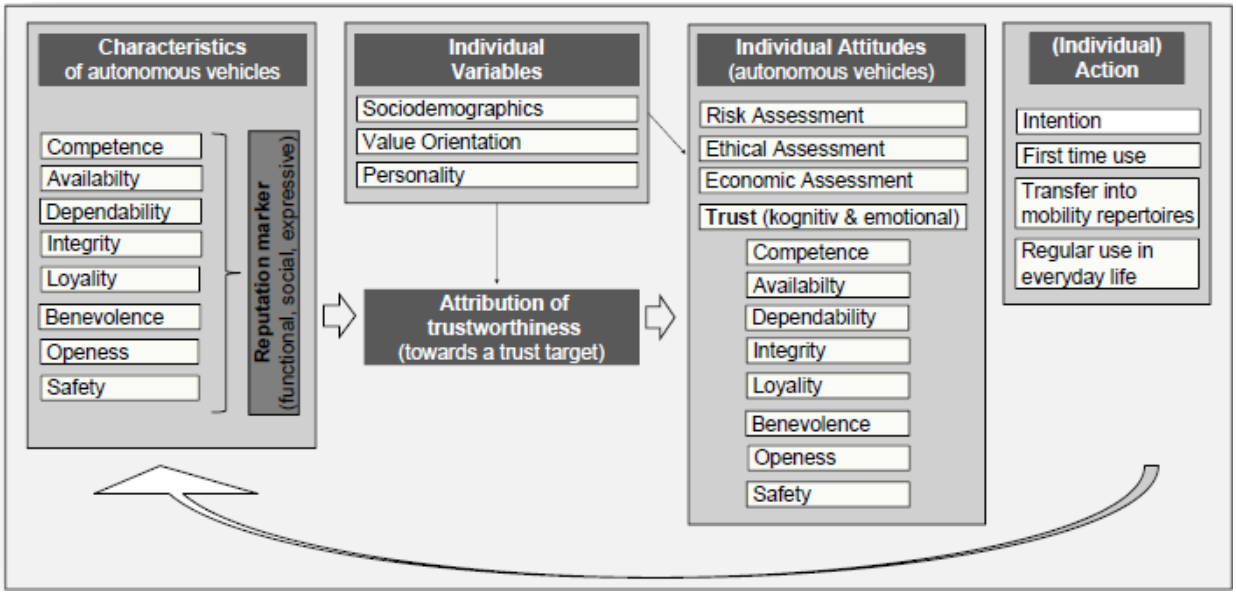
The empirical studies cited above call for a deeper and more grounded understanding of the current and future role of trust and trustworthiness in systems engineering, especially since autonomous vehicles are just one representative of a completely new class of autonomous systems that will cause substantial changes in our daily lives. Since only a small proportion of the surveyed persons will have had their own experiences with autonomous vehicles, it is all the more important to note the basic mindset on which first experiences with the technology will resonate.

Ongoing discussions inside the technological sphere name possible reasons for the surveyed mistrust towards or non-acceptance of autonomous vehicles. These discourses often assume that deficiencies in the existing technical functionality (such as sensor errors) and their lack of robustness in real-world settings are the main causations for mistrust. The multidisciplinary and mostly social science-based research on the individual acceptance of autonomous driving agrees on the importance of robust technical functionality. It is seen as a precondition for trust, often evaluated in empirical studies as "perceived technical performance" or "perceived usefulness". However, empirical findings also indicate that negative framings (pre-attitudes) people have about the technology (e.g., "Autonomous cars are dangerous") and their overall mistrust towards these systems overshadow the perception of usefulness and tolerance of a certain rate of (minor) errors or unexpected behavior.

On the other hand, it is assumed that "overtrust" is just as apt to lead to mistakes during use as a lack of trust is. The case of "overtrust" becomes prominent in media publics particularly for fatal errors involving automated cars and assisted driving—say, when drivers fully rely on the car while using their mobile. A scandalizing media debate triggered by events like these can then negatively impact societal trust in the technology. To combat the loss of societal trust in light of single events like these, systems engineering has started to integrate results from machine ethics, law and political science from an early stage, with the goal of ensuring a high level of harmonization of the technology with existing norms, values and ethical standards (especially with respect to safety and security).

On the very first brink of a new era in transportation, that is, the transition from automated to fully autonomous vehicles, individual trust can serve as a bridge for overcoming psychological barriers. The same goes for societal trust, for example, for political decision makers who set the course for legal frameworks allowing autonomous driving and for clarifying legal consequences in the case of wrong decisions or accidents for which autonomous vehicles are partly or fully responsible ("liability"). However, individual and societal trust are key not only for the very first stage of technology acceptance, as more process-oriented and long-term models demonstrate. Users will be confronted with continuing uncertainties even after they have fully decided to use autonomous vehicles (such as buses, trains and cars), due to the inherent nature of AI technology. Autonomous vehicles will be able to make their own decisions. For this purpose, they should cooperate with humans in a trustful but informed way, including a certain rate of failure tolerance, especially in safety-critical situations. The more autonomous vehicles cross the threshold from being "technological tools" with limited functionality to being "autonomous teammates" with their own strengths and weaknesses, that is, from automated to fully autonomous vehicles, the more that dialogue between humans and machines should help each to get to know and trust the other.

The paper goes on to address the state of the research on trust from an engineering viewpoint, from the humanities, and from social sciences. The concepts of trustworthy AI must draw research from all three disciplines. In the engineering disciplines, it is commonly agreed that trustworthiness is a collection of other well-established technical properties such as reliability, availability, maintainability, safety, security, privacy, usability, explainability, ethics and robustness. Many of these properties have been investigated for decades in engineering. From the humanities arises the concept that trust is a mutual relationship between the trustor and the trustee; and also the idea that trust is a stronger relationship than reliance (or dependability) but also more dynamic. From the social sciences trustworthiness is given when a trustor assesses that an autonomous vehicle has the three characteristics of abilities, benevolence and integrity. The following figure from the paper creates an integration of empirically informed assumptions into theoretical models of trust and trustworthiness:



Key ingredients for trust

Bringing together both humanities and social science perspectives, individual trust and trustworthiness can basically be described as the process-oriented, mutual exchange of information between a minimum of two agents (e.g., a passenger/certification auditor and an autonomous vehicle) towards a certain scope ("trust target"). Thus, trust depends deeply on interaction, in which the ability, benevolence and integrity of the trustee (e.g., the autonomous vehicle) is positively assessed by the trustor (e.g., the passenger or the auditor of a certification authority). Technical trustworthiness is defined within the engineering tradition as the sum of intersubjective, measurable, technical parameters of a system, which describe certain properties of it, often according to respective technical standards and norms. There is broad consensus in the social science literature on the acceptance of autonomous driving that technical parameters (functionalities) are no guarantee for the emergence of trust. In the humanities as well, trust depends deeply on the relationship of trustor and trustee, their motives and expectations.

Systems engineering has long been centered on defining the "abilities" of systems (as capabilities, functions, and requirements). One might infer that the primary aspect of trust related to ability is reliability – a guarantee of correct functionality. The engineering disciplines have well-established toolboxes to ensure reliability of the products they are building. The increased complexity of fully autonomous systems opens up the question of when and how components are used – one may trust the braking system in a vehicle, but do we trust the vehicle's decisions about when and how the brakes are applied?

When dealing with the benevolence of a technical system, more objectifiable characteristics are needed. In order to make "benevolence" a fruitful term for systems engineering, ethical and legal research topics become crucial in their ability to provide specifications on which behavior counts as "good" or "bad" and thus furthers informed and well-calibrated trust.

Integrity means that the system acts according to norms, standards and principles that are important for the trustor and defined beforehand. Integrity is a concept with a strong connection to time: the perception of integrity assumes a pre-post evaluation between norms, standards and principles defined beforehand and the factual behavior shown in specific contexts. It is outside the scope of the engineering disciplines to define these norms and standards in the context of autonomous vehicles, but it is likely inside the scope of systems engineering to derive these standards into system capabilities and requirements.

Trust and Explainability

Even though engineers have succeeded in technically integrating parts of the three key ingredients—abilities, benevolence and integrity—into an autonomous system, this does not necessarily imply that the user or auditor will attribute trust to the autonomous system. One of the major reasons for this could be that users and even auditors are not always put in a situation in which they are able to perceive and observe all the sophisticated engineering hidden behind the surface and therefore unable to assess the degree of benevolence, integrity and functional excellence with which a system is acting. Moreover, current research on safety verification and explainability of AI systems shows that system engineers themselves face the “black box problem” of AI-based system components. There is a need for an additional conceptual and integrative layer—namely, explainability—to communicate technical measures undertaken to make the system trustworthy with respect to its ability, benevolence and integrity. Engineering science has already identified explanations and the explainability of systems as an important tool for proving trustworthiness to humans. However, such an explanation can fail or remain insufficient due to a lack of shared knowledge and the individual expectations of humans.

TNO-ESI:

Trustworthy AI, according to the EU

Trustworthy AI has three top-level requirements, according to the EU [1], that should be met throughout the system's life cycle:

1. it should be **lawful**, complying with all applicable laws and regulations
2. it should be **ethical**, ensuring adherence to ethical principles and values and
3. it should be **robust**, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

The EU has created a framework with seven requirements that AI systems should meet [1], focusing on the ethical and robustness components:

1. **Human agency and oversight**: including fundamental rights, human agency, and human oversight
2. **Technical robustness and safety**: including resilience to attack and security, fallback plan and general safety, accuracy, reliability, and reproducibility
3. **Privacy and data governance**: including respect for privacy, quality and integrity of data, and access to data
4. **Transparency**: including traceability, explainability and communication
5. **Diversity, non-discrimination, and fairness**: including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
6. **Societal and environmental wellbeing**: including sustainability and environmental friendliness, social impact, society and democracy
7. **Accountability**: including auditability, minimization and reporting of negative impact, trade-offs and recovery

[1] https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

Workshop Presentation: Approaches to Trustworthy AI in Applied Research

During the Archimedes Initiative Workshop on trustworthy AI on 4-6 June 2024 in Washington DC, Freek Bomhof, director of the Appl.AI program at TNO, gave a presentation about trustworthy AI within TNO.

The presentation by Freek Bomhof focused on the methodologies employed to ensure Trustworthy AI within applied research contexts. The objective was to examine the application of fundamental principles of Trustworthy AI in ongoing TNO projects and derive insights from the analysis.

To gather data, responses were solicited from flagship projects within the Appl.AI initiative regarding their adherence to the seven principles of Trustworthy AI. These responses and findings from Knowledge Investment Plan reports and strategic plans were incorporated into the presentation. The compiled data is summarized in Slide 34 of the presentation.

Key Findings are the following:

Domain Focus and Principle Adherence: Projects with a narrow domain focus exhibited a less comprehensive adherence to the principles of Trustworthy AI. Conversely, more generic applications demonstrated a more balanced integration of all seven principles. This means that the 7 principles are not 'automatically' addressed at all times, which is understandable but maybe not desirable.

Explicit Trustworthiness Focus: Research projects explicitly addressing trustworthiness, such as GRAIL, MMAIS, and AIOL, inherently included all principles of Trustworthy AI comprehensively.

Audience Reactions and Workshop Observations are:

- Regulatory Sandboxes: The 'regulatory sandboxes' concept attracted significant interest. An international expert group on Regulatory Sandboxes was established at KU Leuven during the Data Week / LAILEC conference, enabling the connection of interested experts from George Washington University to this initiative.
- European and American Approaches: The American audience did not find any elements of the European approach to Trustworthy AI unnecessary or superfluous. Similarly, the American approach did not contain elements that surprised the presenter, indicating a consensus on Trustworthy AI assessment in the academic sphere. The EU AI Act is recognized as having more regulatory power than Biden's Executive Order.
- Environmental Sustainability: The topic of environmental sustainability of AI systems was raised by several attendees, reflecting its growing importance in AI research discussions.
- NSF Institute for Trustworthy AI in Law & Society (TRAILS): This initiative aims to align AI development closely with societal needs, serving as an interesting counterpart to the Dutch Nationale Wetenschaps agenda programme.

Trustworthy AI at ESI: Definition and Future Research Directions

Definition of Trustworthy AI at ESI

The focus of the work in ESI around trustworthy AI aligns closely with the EU framework, emphasizing ethical and robust components over lawful considerations. ESI's mission is to enhance the Dutch high-tech ecosystem by fostering efficient innovation and high-quality system development. Trustworthy AI is integral to both the development processes and the resulting systems.

Critical aspects of Trustworthy AI at ESI include a human-centric design approach, where AI systems facilitate trustworthy design decisions and alternatives, prioritizing transparency, robustness, safety, sustainability, and accountability. Human decision-making remains central to meeting stakeholders' needs and preventing potential biases. All AI-focused ESI projects should explicitly consider the seven principles of Trustworthy AI, as outlined in the EU framework, using tools such as the Trustworthy AI Assessment List.

We see trustworthy AI playing an essential role in the **development process**, where we use AI for system engineering - **AI4SE**, and the **developed systems**, where we use system engineering for AI - **SE4AI**.

- In the development process, AI should suggest trustworthy design decisions and design alternatives, for example, in light of transparency, robustness, safety, sustainability, and accountability. Humans should be in the decision-making loop and be responsible for end-to-end decisions. AI should consider the stakeholders' characteristics and needs and prevent potential biases.
- The developed solutions should be trustworthy, such that they are robust and safe, guarantee the privacy and integrity of the data, and are fair and sustainable.

Trustworthiness in the Context of AI4SE at ESI

In the context of AI4SE, we envision system architects and engineers will heavily use so-called digital assistants. These assistants will provide help with various tasks in the system development process, including requirements engineering and system design, making trade-offs, implementation, validation, and verification. As ESI, we will also develop digital assistants as part of methodologies to improve engineering effectiveness. Trustworthiness is key to getting our methodologies accepted by system architects and engineers. The results of the methodology should be trustworthy, such that no problems are introduced into the system that is being designed.

To ensure that our methodologies are trustworthy, we will use the latest guidelines, best practices, and frameworks on trustworthiness. Within TNO, various groups are working in consortia, building roadmaps on trustworthy AI and developing solutions to reason and assess trustworthiness in solutions. We will leverage these results and consider them in our research projects.

Trustworthiness in the Context of SE4AI at ESI

In the context of SE4AI, we see research questions on designing trustworthy AI-enabled systems and keeping systems trustworthy over their lifetime. Some key questions include:

- **System aspects**: How do you ensure that an AI-enabled system is safe, reliable, and robust? How do you ensure that the behavior of an AI-enabled system stays safe once the system starts learning on the fly and changes its behavior in response? How can we ensure transparent decision-making so that system users understand the system's behavior? How do we reason and make trade-offs on trustworthiness aspects like dependability, safety, reliability, and robustness? Inspiration can be found in the Appl.AI MMAIS project, that is addressing how to make trade-offs for AI safety.
- **Testing and monitoring**: How do we test AI-enabled systems before they leave the factory? How do we monitor system behavior during operation?
- **Diagnostics**: How can we diagnose whether an AI component or another component causes a system failure? How can we diagnose and trace the decision-making of an AI component when that component causes system failure?
- **Upgrades**: How can you upgrade AI-enabled systems that have learned in the field during operation? Do you need to start learning from scratch again? Can you transfer the learnings, and how can you trust them in the upgraded situation?
- **Performance**: How can the performance of an AI-enabled system be predicted? How do AI components not unexpectedly degrade the system's predictability?

Future research directions

For ESI, the main concern is not specifically to design and develop trustworthy AI but to help our partners design and maintain **high-tech systems that remain trustworthy even once they embed or interact with AI**. Therefore, we must adapt system engineering techniques to high-tech systems enabled by AI and develop system engineering methods and techniques that are trustworthy, even if they are AI-based. For that, we need to enforce the principles of AI trustworthiness when developing methods and tools for designing and maintaining (AI-enabled) high-tech systems. Conversely, methods and techniques for system design, optimization, and maintenance, such as computable models, simulations, and digital twins, must be adapted to enforce the principles of AI trustworthiness on high-tech systems enabled by AI.

For AI-enabled systems, we aim to ensure trustworthiness, particularly robustness and reliability of the system throughout its lifecycle, guaranteeing its dependability and evolvability in a changing context. We talk here about 'lifecycle management' in the context of trustworthy AI (see the approach taken in Appl.AI's SEAMLESS project).

Specific methods should be developed to guarantee the dependability of **evolving AI-enabled systems**, especially when exposed to new data after a period of time in operation. Comprehensive stress tests should be performed before the system leaves the factory to improve its robustness. Continuous monitoring and adaptation capability should guarantee the system's reliability when deployed in a changing context.

Abrupt changes can disrupt system performance, especially for systems with tight control loops. Deploying **updates and upgrades** to systems embedding AI requires a well-thought policy addressing many challenging trade-offs. For instance, preserving learning done in the field could cause unpredictable behavior, but removing them and starting over could represent a significant loss for the system performance in the deployment environment.

In the event of a system failure, the ability to reproduce the faulty behavior is essential for diagnosing and fixing the issue. This requires **transparency in system operations and the ability to analyze and explain the system's behavior**, which is critical to maintaining user trust and achieving a successful system's maintenance. Explanations should be adapted towards end users, practitioners/domain experts, and researchers.

Ensuring that there are **mechanisms for human oversight and intervention** is a safety and ethical concern. The system should not operate in an isolated, self-centered, human oversight-superseding manner, especially when interacting with humans.

In the event that the AI must make decisions that could harm humans, ethical frameworks must guide its actions. **The system should be designed to minimize harm and ensure fairness**, avoid disparities in service based on sex, ethnicity, or other factors, and guarantee data privacy, quality, and integrity.

By following these guidelines, AI-enabled systems' robustness, reliability, and ethical integrity can be maintained, ensuring they continue to perform well even in changing contexts.

Conversely, starting with **system architecting and design, passing through system testing, monitoring, maintenance, and upkeep**, we should guarantee the trustworthiness of AI used as a tool to translate market and technology choices into practical system concepts, enforce design guidelines for **dependability and quality**, optimize the **system design for performance, diagnose system failures and maintaining them dependable**.

Proposing **human-centered solutions** for design and maintenance techniques on the one hand and AI-enabled systems on the other is paramount to fostering human acceptance and supporting human agency and oversight, allowing humans to remain central in decision-making processes. This includes developing solutions tailored to the concerned human profile (newbie, expert, system or software architect, field service engineer, final user, etc.) to enhance their system understanding for design, maintenance, and operation. Such enhancements might need to address socio-technical challenges in high-tech industries.

Finally, considering the broader socio-economic implications of AI deployment could allow AI systems to be designed and maintained while keeping sustainability and societal well-being in mind. This is especially true when the concerned systems impact public well-being or health, and their operation must follow standards of quality imposed by regulatory agencies.

Conclusion

In conclusion, ESI's priority is to ensure that AI-enabled systems remain trustworthy, robust, and reliable throughout their lifecycle.

We foresee the need for research related to trustworthiness in both SE4AI and AI4SE. Concerning the first, SE methods and techniques must be adapted to make AI-enabled high-tech systems compliant with trustworthiness principles. Concerning the second, the methods themselves should comply with trustworthiness principles when using AI. The ultimate aim is to ensure system dependability and maintain transparency for issue diagnosis. Ethical frameworks and human oversight could be crucial to prevent catastrophic malfunctions. Human-centered solutions could enhance user understanding and support their agency. Socio-economic implications could foster sustainability and societal well-being, especially in the context of mandatory regulatory compliance. By following these guidelines, AI-enabled systems and the techniques to design and maintain them can remain reliable and ethically sound, fostering human acceptance.

TECOSA:

Defining trust

The state-of-the art provides several definitions of trust and trustworthiness. We have chosen to rely on existing definitions. In doing so we still want to emphasize the following:

- Trust should be seen in the context of relations, specifically between humans and cyber-physical systems (CPS), and/or between different CPS. Psychological research has shown that trust between humans and robots depends on several attributes including predictability, dependability, and faith, relating to e.g. controllability, understandability, and performance, [1]. Trust can also be studied in the context relations between machines/CPS, as relevant for collaborating systems. A related concept is that of "trusted computing" [2].
- Trust could be seen as slightly different between humans/machine vs. between machines, since psychology and human cognition is involved in the former, and since in machine to machine interaction, non-human solutions (algorithms and protocols) can be devised.
- Trust alone is not enough; one also needs to consider potential overtrust, responsibility and liability. This becomes apparent with some of the developments in automotive solutions for automated driving at L2 and L3 levels of automation (SAEJ3016). Improper training, definition and information of capabilities and responsibilities, and (over-) promotion may easily lead to complacence and unawareness of how is in charge, see e.g. [3]

Introducing trust in SE

We found two potential interpretations to this question; (a) How to make sure trust is part of SE (processes, methods, competences), and (b) How to ensure that the results of SE can be trusted!

Interpretation (a) is close to question number 3 (next section). Considering interpretation (a) first, we believe that trust should in principle be covered by current SE and systems thinking, assuming a starting point that considers a system of interest with interactions in its environment, the concerns of the involved stakeholders, and all life-cycle phases. It is clear however that trust represents a multifaceted endeavor and also a trajectory that brings in many concepts and concerns that are somewhat new for SE, such as ethical considerations, and behaviors failures of deep learning systems. For this purpose, we believe there is a need for explicit support to ensure that trust aspects are properly incorporated into SE. As stated in [4], new risks related to complex CPS may otherwise be underestimated (e.g. due to a lack of awareness such as with limited insights into cyber-security risks) and that dependencies and trade-offs between trust-related attributes may not receive sufficient attention. This thread is continued in the next section.

For the other interpretation, (b), ensuring trust in SE work relates to ensuring trust in the organizations carrying out the work, including through demonstrating safety culture and transparency (e.g. in terms of safety cases and relevant trust related performance indicators), and relevant compliance.

SE for trust in future systems.

SE has a central role in dealing with trust through systems thinking (providing a systemic view) and its process frameworks (providing a systematic approach). As already mentioned, we believe that architectural frameworks dedicated to trust and trustworthiness are important to guide engineering efforts by supporting a careful dependency and trade-off analysis involving a broader set of experts.

It is further essential to understand the domain in which you are operating, such that appropriate tools can be adopted, [5]. When solutions and systems go into the complex domain, there is a need to develop new methodologies, including requiring controlled testing (such as AI sandboxes).

In the complex domain, it also becomes necessary to design for run-time supervision and data gathering to be able to handle the unforeseen and emergent effects. Such supervision is generally needed both for real-time supervision and potential actions as well as for longer term analysis and actions, [6-7], with vehicles at higher levels of automation as a prime example.

Systems in which trust and dependability are critical are spreading in our society with the proliferation of more advanced CPS. This generally requires an emphasis on complexity management and considerations across multiple disciplines. Specifically, when including more machine learning (and thus data-driven) components, the corresponding CPS will be provided with new capabilities that however also rely on software and hardware, and feature new failure modes, [8].

When it comes to the consideration of AI in future cyber-physical systems (or just IT-systems), a key aspect is to enhance the level of awareness of risks, pros and cons and side-effects, while developing further guidelines and frameworks, and interacting with standardization and regulatory bodies.

The EU AI act brings both attention to key aspects such as risks of AI based systems, but also many questions and concerns, including for how SMEs should handle a large number of complex regulations.

References

[1]: K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, P. A. Hancock, A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems, Human Factors 58 (2016) 377–400.

[2]: Z. Huanguo, L. Jie, J. Gang, Z. Zhiqiang, Y. Fajiang, Y. Fei, Development of trusted computing research, Wuhan University Journal of Natural Sciences 11 (2006) 1407–1413.

[3] Tesla Drivers Say New Self-Driving Update Is Repeatedly Running Red Lights (futurism.com).

[4] Rusyadi Ramli and Martin Törngren. Towards an Architectural Framework and Method for Realizing Trustworthy Complex Cyber-Physical Systems. Joint Proceedings of RCIS 2022 Workshops and Research Projects Track: co-located with the 16th International Conference on Research Challenges in Information Science (RCIS 2022) / [ed] Joao Araujo and Jose Luis de la Vara, CEUR-WS , 2022, Vol. 3144.

[5] David Snowden and May Boone. (2007). A Leader's Framework for Decision Making. Harvard Business Review. 85 (11): 68–76. PMID 1815978

[6] Martin Törngren and Ulf Sellgren. Complexity Challenges in Development of Cyber-Physical Systems. In Principles of Modeling; Lohstroh, M.; Derler, P.; Sirjani, M., Eds.; Springer, 2018; Vol. 10760, Lecture Notes in Computer Science, July 2018. doi:10.1007/978-3-319-95246-8_27.

[7] Anders Cassel et al. DevOps Safety – do not get left at the station! Position paper at Safecomp 2024 (available here: https://www.safecomp2024.unifi.it/vp-22-accepted-position-papers.html)

[8] Martin Törngren. Cyber-physical systems have far-reaching implications. Hipeac Vision 2021. https://doi.org/10.5281/zenodo.4710500

SERC:

SE, AI, and Trust

SERC has spent a lot of time thinking about the unique role of Systems Engineers in the AI space. We want to make sure we're differentiated and no (accidentally) competing with algorithm developers on their core. We see two key areas for contribution:

- AI4SE, which focuses on the application of AI in support of systems engineering processes to enable enhanced decision-making, optimization and efficient effort allocation. Fundamentally this is about applying (generally not innovating on) AI tools to do Systems Engineering better.
- SE4AI, which focuses on leveraging (and extending) systems engineering principles to develop AI Enabled Systems (AIES) that are safe, robust etc. Here, the AI is typically complex and state of the art, but it is not the part developed by SEs. We're focused on extending SE methods to support this new class of system.

Of course, these two streams are interconnected. There is a need for SE4AI on the AI4SE tools, but also integrating AI into the SE workflow may change what SE work looks like.

AI for Systems Engineering (AI4SE)

Within AI4SE multiple SERC researchers have been building and playing with prototypes to assess their potential in multiple aspects of SE work. These range from document summarization and fact extraction to templating processes and brainstorming. This has led to three categories of research work: 1) AI to support information processing tasks; 2) AI to support the generation of SE content; 3) alternative architectures for human AI collaboration in process tasks.

We see characterizing the first two as critical for enabling the third. Right now, a lot of the focus has been on replacing or augmenting existing tasks or subtasks. However, AI has different strengths and weaknesses than human performers and it follows that the right configuration of work should change to match. Therefore, there is a need to consider both the architecture of the work (set of tasks and dependencies) and the mapping of opportunities and risks of AI replacement, augmentation and/or transformation. This has already been shown through cases in multiple contexts (examples provided).

This notion of architecture is important for SE4AI too. It's important to realize that system behavior is not only about how the algorithm works (although that's important), or even one human working with one cognitive assistant. Most AIES are embedded in complex socio-technical systems, and behavior must be measured at that level. Existing SE practices get us most of the way there, but this is a scale and level of open-ness of the system boundaries that there is a need for method development.

Systems Engineering for AI (SE4AI)

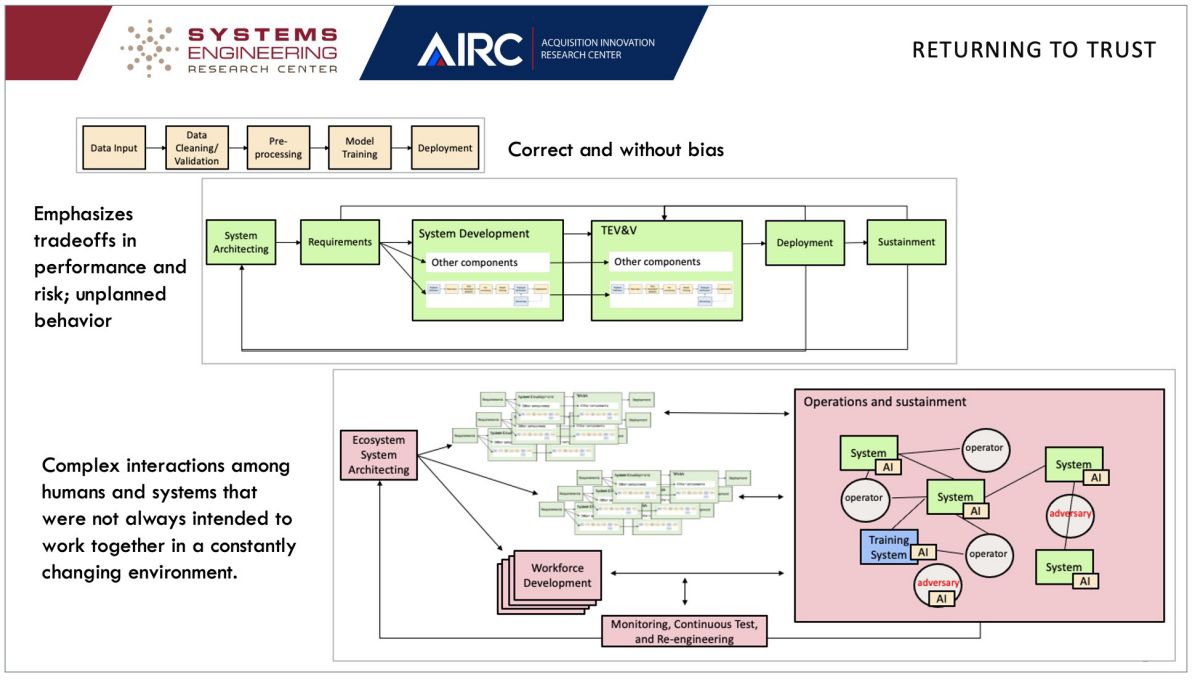
To that end, within SE4AI SERC research has broadly focused on 1) Modeling AI as part of a complex system of autonomous systems; 2) measuring AI "-ilities" relevant to SE; and 3) new methods for Test & Evaluation (T&E) as a continuum. Importantly, for all of these, we recognize the need to study this system at multiple levels, including at the socio-technical level.

Trust: Socio-technical systems as the unit of analysis for TEV&V

Returning to trust, we emphasized a distinction between trust – which is by the user and a property of the relationship, trustworthiness, which is a property of the artifact, and trustworthy AI, which combines both aspects, asking if humans *should* trust the artifact.

The aspect of the interaction is particularly important because different types of people (with different backgrounds, different expertise etc.) build trust in different ways. Three canonical examples include: a) a developer perspective which typically focuses on accuracy at the algorithm level; b) a domain expert, who tends to assess if "the AI" agrees with their experience; and c) an end user who tends to trust third parties. These kinds of trust all make sense when AI is replacing a known function (like identifying if a mass on an x-ray is a tumor as the doctor would), but how they apply is more complicated when the AI is embedded in the system in a complex way. Consider the role of AI in Uber's enterprise. How would a domain expert judge if the matching algorithm is working based on their experience? Understanding how trust formation interacts with system use is critical to understanding complex AIES, because overtrust and undertrust both affect system behavior as a whole.

More broadly, while there's still work to do classifying which tasks match (and don't match) the capabilities of AI, we see most of the opportunity for SEs in thinking about how AI affects the architecture of work. We also see a need to shift the T&E and Verification & Validation (V&V) discussion to the socio-technical system and consider investing in test beds as a shared resource. The below slide emphasizes the levels of analysis for assuring AI. All levels of analysis are important – focused on ensuring the algorithm is producing correct and unbiased results, that an AI-enabled system is behaving as required, and that the necessary post deployment monitoring is in place to ensure that the human(s)-technology(ies) systems continues to behave as desired even as context changes. While the first two levels are relatively better understood and leverage existing assurance tools, the latter requires an evolution of systems principles and is a critical focus moving forward.



Part 2. Defining Trust, from the Workshop Presentations

The workshop presenters all discussed related perspectives on how trust in AI and automation can be developed. Trust is a systems thinking problem. It involves the relationships between trusting agents (behavioral) and the relationships between dimensions of trust (system qualities). Axel Hahn from DLR started the discussions with a presentation centered on autonomous vehicles. "Autonomous vehicles are a new class of systems and will substantially change our daily lives." He outlined a number of behavioral and system characteristics present in fully autonomous vehicles:

The following behavioral and system characteristics of trust and trustworthy AI were noted from all of the presentations:

- **Trust** is a multi-party agent-based relationship: A trusts B with respect to C. Trust is a relationship both between agents (human or AI) and the larger groups (organizations, communities, societies) that those agents are shaped by. Trust is a property of these relationships.
- **Trustworthiness** is a property of the artifacts exchanges in the relationship. Trustworthy AI is a product the relational properties between user and system and the artifacts produced by the system. Focusing on trust forces to think about these relationships.
- Trust is dynamic: Trust develops over time, changes with experience, and shapes all agents experiences in turn.
- Successful fully autonomous systems, including those with AI components, don't begin fully autonomous. They start with supporting human analysis, then assisting human tasks, then augmenting human tasks, then achieve fully autonomous capabilities. This is a methodology to achieve trust over time, by demonstrating incremental value in human workflows.
- Trust spans all physical, cognitive, informational, and emotional aspects of a relationship.
- The foundations of Trust come from Integrity: acting in accordance with values, norms, standards, and principles. There are both physical and social standards.
- Trust is demonstrated by actions or behaviors versus intent (motives) and demonstrated capabilities (abilities) that produce results.
- Trust includes aspects of benevolence (good will) and vulnerability that drive social relationships. Trust evolves in a human context that defines good or bad.
- Trust has cost, systems must be designed to be trustworthy.
- Relationships between humans and AI result in new "supra-human" tasks that humans can now do with help of a machine. These relationships are emergent.
- Successful automated machine systems act as a valued partner to humans: they are accurate, predictable, adaptive, resilient, increase security and reduce risk (assured), and are perceived to behave ethically or with fairness. These are system qualities.
- Trust in automated machine systems should develop organically, starting at lower risk points and gradually demonstrating value or return on investment.

- Trust needs to be considered as a human factor in systems engineering. The automation paradox must be considered in system design: under what circumstances do agents step in to supervise or control an automated task. A contextual control mechanism or "dial" supports trust.
- Governance of these systems should be established gradually, collecting and building from better data reflecting human responses, and identifying the stopping point. Governance must be scalable, balancing speed versus risk.
- Accountability extends the relationship between human and machine to a relationship between a social actor and a forum, where the actor has to explain and justify their conduct, and the forum can pass judgment. This expands the domain of Trust from the use of the Trustworthy AI to the organizations that produce it and the societies that judge its trustworthiness. Control, governance, and accountability are important definitional aspects of Trustworthy AI.
- Challenges to trust include misinformation, inequities in control, quality versus productivity gain, inherent complexity – particularly with software, interaction complexity, predictability, safety, security.

Recent publications such as the EU's AI Act, the TAILOR Handbook of Trustworthy AI, and NIST's AI Risk Management Framework provide a number of taxonomies and dimensions for trustworthiness in AI. Trust is both a behavioral quality of human-human, human-machine, and machine-machine relationships, and a system quality of the computing systems that are required for machine learning and machine automation. In the workshop we produced a set of definitions of these qualities drawing on both behavioral aspects human leader-follower relationships and system qualities of robust computing systems. Thes are outlined in the next section.

Definitions of Trust

Concept of Trust in organizational leadership

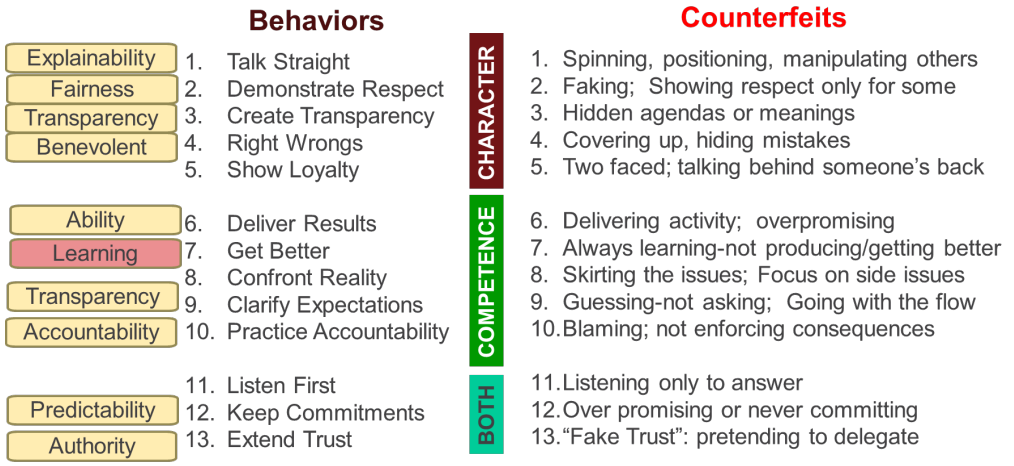
In organizational leadership, there is a defined relationship between **intent**, **behavior**, and **trust**.

"A person has **integrity** when there is no gap between intent and behavior…when he or she is whole, seamless, the same—inside and out. I call this "congruence." And it is congruence—not compliance—that will ultimately create credibility and trust." [Steven M.R. Covey, The Speed of Trust, 2006].

Steven M. R. Covey in The Speed of Trust defines four cores of trust in a leadership and organizational context as:

- Integrity – the fundamental motives of the agent and the behavior that follows.
- Intent – the fundamental motives of the agent and the behavior that follows.
- Capability – the capability of the agent to successfully produce and accomplish tasks.
- Results – the credibility of the agent established as a function of past, current, and future expected performance.

Covey continues to define 13 behaviors of high trust leaders. We have associated these with a set of qualities and behaviors captured from the workshop presentations.

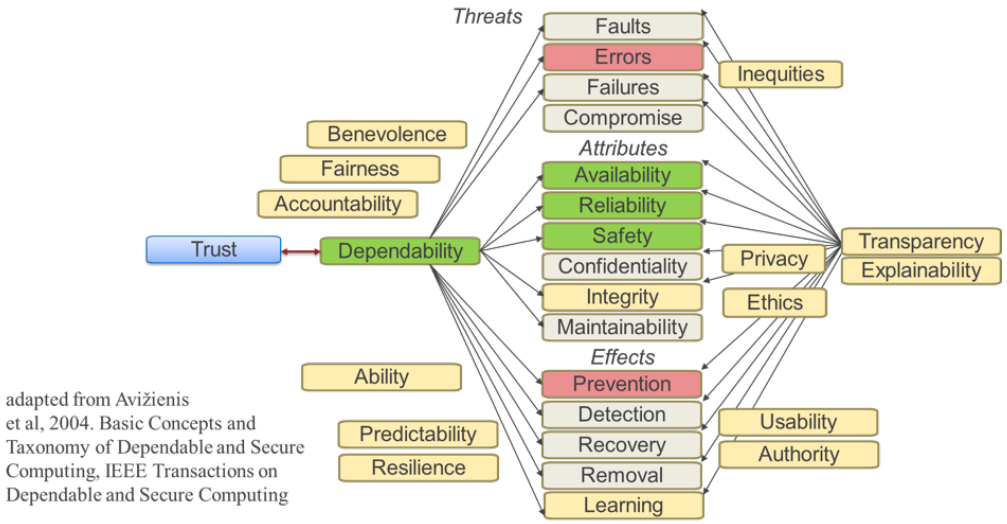


Covey, Stephen M. R., and Merrill, Rebecca, The Speed of Trust, Free Press, 2006, ISBN-13: 978-0-7432-9730-1

Concept of Trust in computing systems

In computing systems, there is a defined relationship between **dependability** and **trust**. This relationship is defined by the dependence of one system on another, and the **acceptance** that the other system is also dependable. [Avižienis et al, 2004. Basic Concepts and Taxonomy of Dependable and Secure Computing, IEEE Transactions on Dependable and Secure Computing].

This dependence can be either human/machine or machine/machine.



adapted from Avižienis et al, 2004. Basic Concepts and Taxonomy of Dependable and Secure Computing, IEEE Transactions on Dependable and Secure Computing

These two views effectively create a categorization for the definitional attributes of trust.

Definitional attributes of trust in AI/ML:

Ability: possession of the means or skill to do something that has use; talent, skill, or proficiency in a particular area.

Usability: the degree to which something is able or fit to be used.

Learning: the process of acquiring new understanding, knowledge, behaviors, skills, values, attitudes, and preferences through experience, study, or by being taught.

Authority: the power or right to give orders, make decisions, and enforce obedience; the official permission or right to act, often on behalf of another.

Integrity: the quality of being honest and having strong moral principles; the honesty and truthfulness or earnestness of one's actions; no corruption of information.

Dependability: the quality of being trustworthy and reliable; constancy; ability to provide services that can be trusted within a time-period.

Trustworthy: able to be relied on as honest or truthful, worthy of confidence.

Predictability: consistent repetition of a state, course of action, behavior, or the like, making it possible to know in advance what to expect.

Resilience: the capacity to withstand or to recover quickly from difficulties.

Benevolence: the quality of being well meaning; disposition to do good; kind and helpful.

Fairness: the quality of treating people equally or in a way that is right or reasonable; free from bias or injustice.

Inequity: lack of fairness or justice.

Accountability: acknowledgment of and assumption of responsibility for actions, products, decisions, and policies; the fact of being responsible for what you do and able to give a satisfactory reason for it, or the degree to which this happens; assurance that an individual or organization is evaluated on its performance or behavior related to something for which it is responsible.

Confidentiality: the state of keeping or being kept private; preserving authorized restrictions on access and disclosure.

Part 3: A Roadmap for Trustworthy AI, from the Workshop

The highlight of the systems thinking workshop was a facilitation focused on the near-term issues, long-term aspirations, and mid-term innovation pathways necessary to create "systems of trust." The interactive facilitation employed the Three Horizons framework.[1] After receiving an overview of the workshop focus and tool framework, the group ideated dimensions of the present (Horizon 1), ideal future (Horizon 3), and the transition pathways (Horizon 2) that might take us from the present state to the ideal future state.

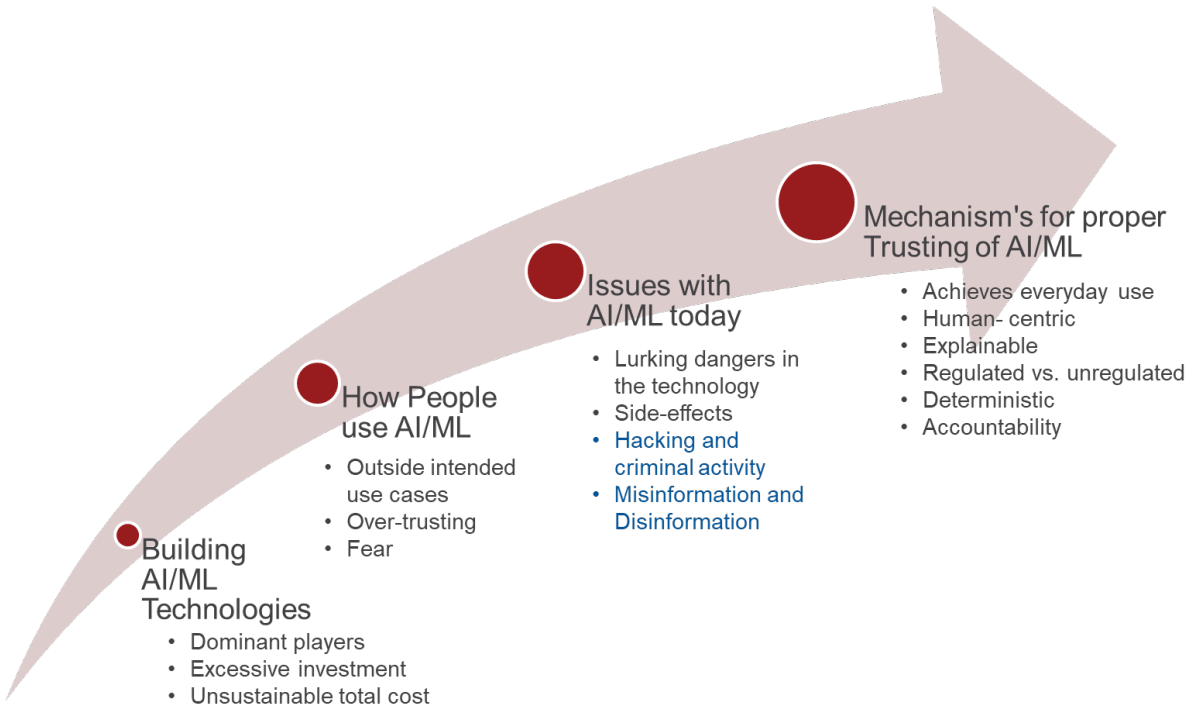
In order to make plans and strategies for the future, we have to think of what the future will look like. However, this is much harder than many can conceptualize, especially when considering the complexities between information technologies, software, AI algorithms, and future more automated systems. The workshop applied the Three Horizons Model to give us a deeper understanding of the significance of what we usually call short, medium, and long-term futures. Attention to the three horizons always exists in the present moment, and our evidence about the future is rooted in how people (including ourselves) are behaving now. The model is based on the observation that business, technologies, political policies, and entire communities exhibit life cycles of initiation, growth, peak performance, decline, and more. These cycles can be viewed as waves of change in which a dominant form is eventually overtaken and displaced by another.

A key difference between the Three Horizons framework and other strategic foresight methods is that it allows the selection of futures to be aspirational – instead of focused up front on strategy or intentionally broad. It tunes our awareness toward the signals of change that align with our aspirations, without the pressure of correctness or accuracy. The purpose of the tool is not to reach agreement on possible futures, rather to gather a large set of insights on current, near-term, and far-term states from the participants. Initially we use the framework to identify future dilemmas and emerging interactions to populate a system map for further study.

Horizon 1 (H1)

The first horizon describes the current way of doing things, and the way we can expect it to change if we all keep behaving in the ways we are used to. H1 systems are what we all depend on to get things done in the world. Innovation and change in our H1 systems is happening, but it is about sustaining and extending the way things are done now in a planned and orderly way; uncertainties and risks are to be eliminated or prepared for – the lights must be kept on. Nothing lasts forever, and over time, we inevitably find that our H1 ways of doing things are falling short – no longer meeting expectations, failing to move towards new opportunities, or out of step with emerging conditions.

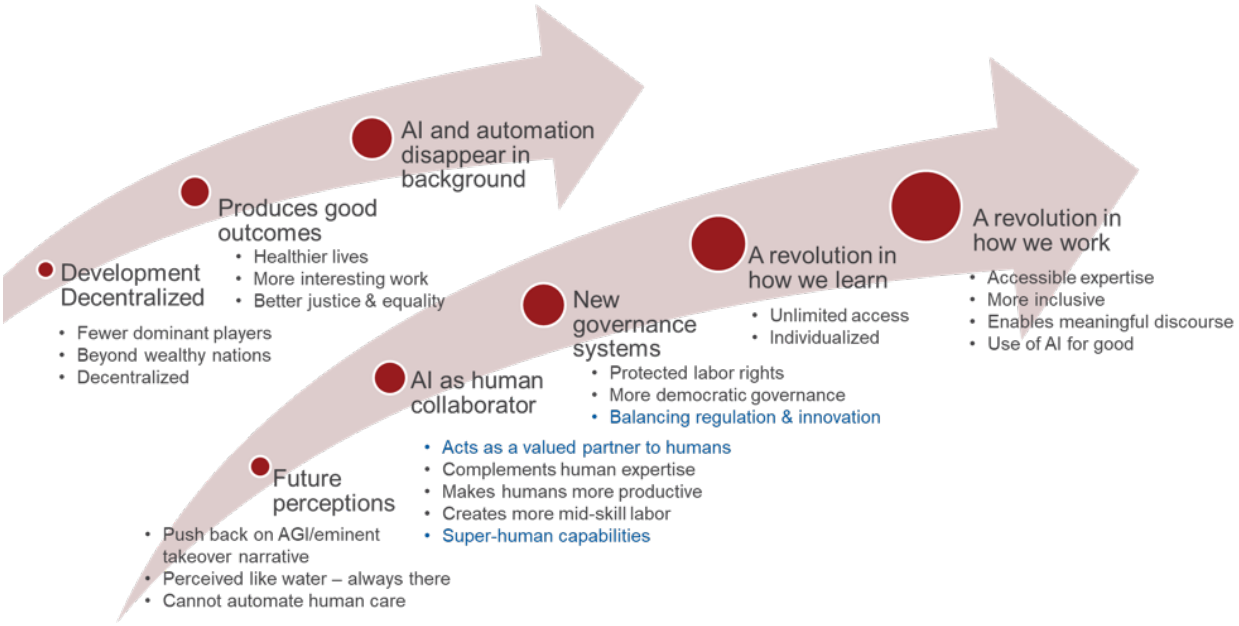
[1] *The Three Horizons framework is a facilitation model for analyzing transformative innovation developed by the International Futures Forum.* http://www.internationalfuturesforum.com/three-horizons



Although the current wave of AI and ML applications has been evolving for a number of years, it became viral with the introduction of OpenAI's ChatGPT model. In the past 2 years, numerous large language models and countless applications have been evolving in the race to take advantage of this new technology. Most experts note that we are just at the beginning of this wave and how it will evolve over the next several years is difficult to predict. Key issues to be resolved relate to automation (what human tasks will be automated) and trust. A further issue is sustainability in terms of total cost, particularly energy use. The dangers and side effects of the technology, as with many new technologies, need to be managed. Research is needed on the mechanisms for proper trusting of AI systems, which as they develop more human-like qualities, will be much more complex and embedded in the socio-technical regime. As noted previously, research must converge across the technical, humanities, and social sciences – research will be trans-disciplinary.

Horizon 3 (H3)

The third horizon is the future system. It spurs reflection on new ways of living and working that will fit better with emerging need and opportunity. H3 change is transformative, bringing a new pattern into existence that is beyond the reach of the H1 system. There will be many competing visions of the future and early pioneers are likely to look quite unrealistic – and some of them are.

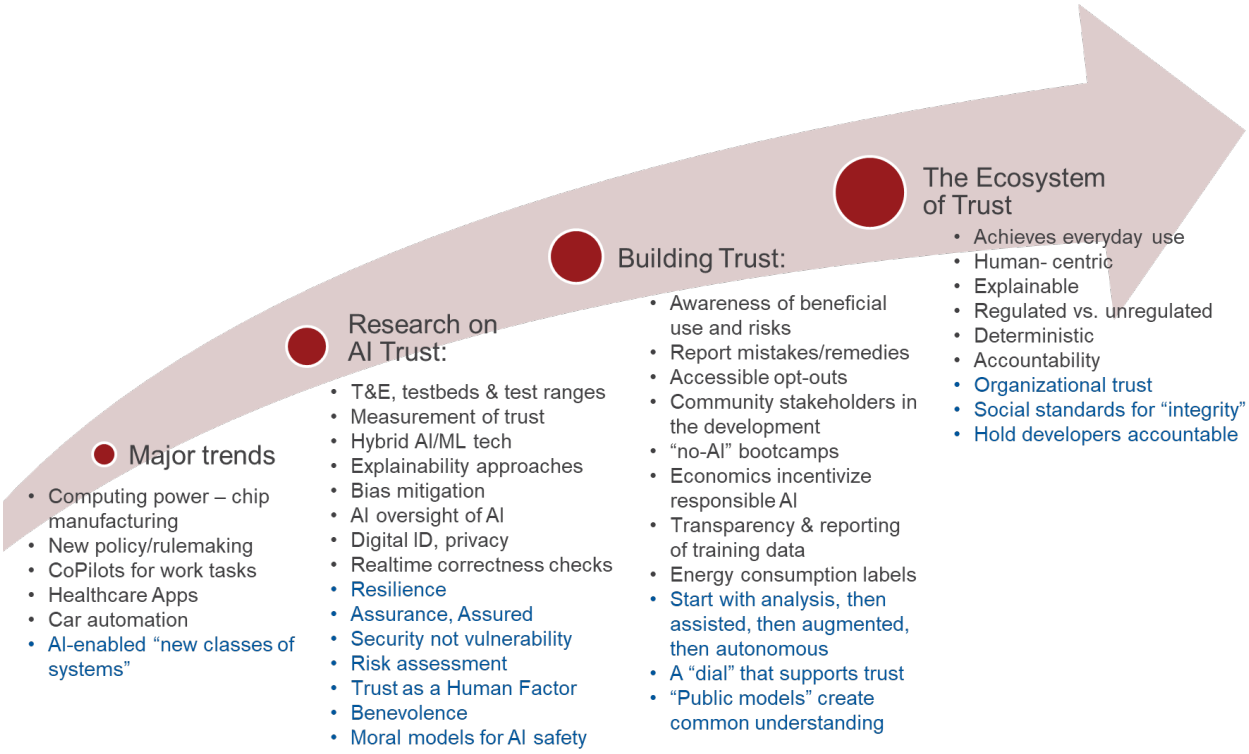


The number of ideas generated by the group was quite large. They have been divided into two waves: the first related to development of AI/ML applications and the second related to the human effects of the technology. In the developmental wave, we are already seeing a decentralization in availability of foundational or large language models. This should continue as LLMs get more specialized by domain. We want the technology to produce good outcomes, particularly in the areas of human health, job satisfaction, and equity. As with other machine automation technologies in the past 200 years, norms will develop, and the technology will become part of the background. With the scale of automation possible today, the opportunity to improve human lives and not just profit from them cannot be missed.

AL/ML applications will produce (and are already starting to produce) a revolution in how we learn and how we work. It is incumbent on the research centers and universities like the SERC, DLR, TNO, and TECOSA to become trusted agents in this transformation. We must be public developers and users of the technology, and we must demonstrate the new SE methods and tools required to ensure these technologies earn our trust over time. This includes not just the technology and applications, but also the governance mechanisms, human interaction, and public perspectives on trustworthiness of the technologies. You can see much of this agenda in the partner presentations and statements presented earlier.

Horizon 2 (H2)

The second horizon is the transition and transformation zone of emerging innovations that are responding to the shortcomings of the first horizon and anticipating the possibilities of the third horizon. H2 innovations seem very similar to your current products and services, and the overpowering temptation is to use the same metrics to assess their success. However, because these ideas are new, it takes time to get them configured effectively. This means that if you treat H2-oriented innovations just like H1-oriented innovations, you are likely to abandon them too quickly because it will seem like they are not performing well. You have to figure out a way to ring fence H2 innovation efforts.



The H2 ideated the start of a research agenda for engineering trust into AI/ML systems. The list of research topics on AI trust mentioned in this facilitation was quite long and there are many others not listed. As many of these topics have to do with application behaviors and system qualities, systems engineering research centers should play a central role in research on trustworthy AI. As we move forward, the Archimedes team will develop and publish a more detailed roadmap on trustworthy AI research.

Beyond research, the applied research and development centers in the Archimedes Initiative must be at the forefront of applying AI/ML technologies and demonstrating trustworthiness to the public. This spans all aspects of technology, humanities, and social science.

We called the long-term goals "the ecosystem of trust." Trust is a complex system characteristic and develops over time. It is clear (at least to us systems engineers) that trust will not be adequately addressed by disciplinary research and require a multi-disciplinary and systems thinking mindset in order to connect research to outcomes.

Summary

This Archimedes Workshop joined together our four systems engineering research centers to focus on a common agenda for trust in AI systems and development of trustworthy AI. The presentations covered the concepts of trust quite broadly and we believe quite completely. The systems thinking workshop started a joint process to develop an Archimedes research roadmap for trustworthy AI systems. The unique format (presentations and systems thinking workshop) both ideated the broad research needs and further coupled the community of research in this domain. Stay tuned for the publication of the joint Archimedes Initiative Roadmap on Systems Engineering Research for Trustworthy AI.