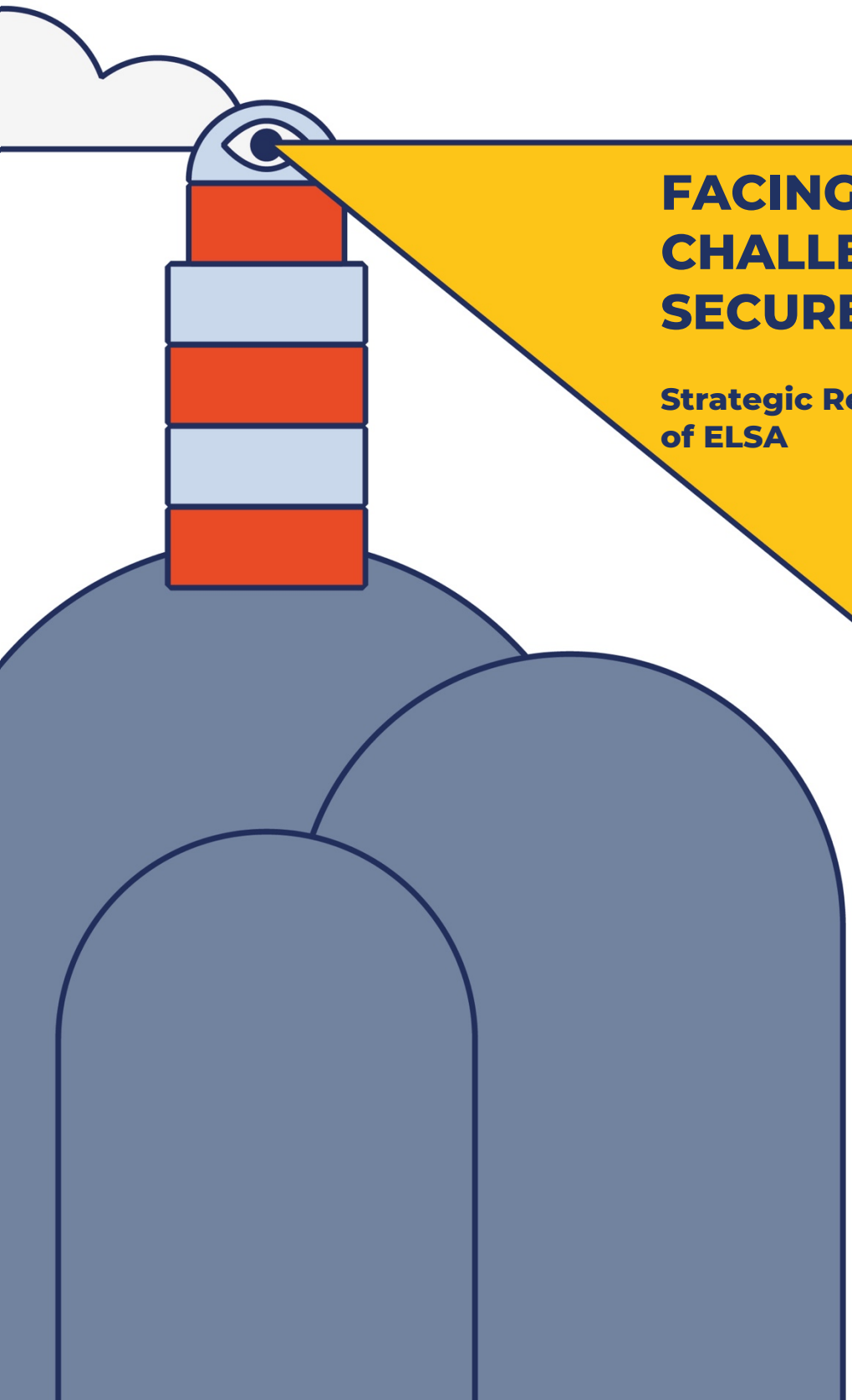
elsa
European Lighthouse on Secure and Safe AI

31.08.2023 – Version 1.0



FACING THE GRAND CHALLENGES OF SECURE AND SAFE AI

Strategic Research Agenda of ELSA

Funded by the European Union

1. Summary

Increasingly pervasive deployment of AI systems, often building upon machine learning, have highlighted the urgency of enforcing the principles of Trustworthy AI to make these systems work for the good of the people and society. Achieving this goal requires societal and policy actions, but also research in technologies and social principles that enable reaching these goals.

The European Union has tasked the ELSA consortium to build a network of excellence on research in secure and safe artificial intelligence (AI). ELSA is a virtual centre of excellence that builds upon the ELLIS network and spearheads efforts in foundational safe and secure AI methodology research addressing three major challenges: The development of robustness guarantees and certificates, privacy-preserving and robust collaborative learning, and the development of human control mechanisms for the ethical and secure use of AI with a focus on use cases health, autonomous driving, robotics, cybersecurity, media and document intelligence.

ELSA is taking a foundational and interdisciplinary approach to these challenges that are characterised and outlined by this **Strategic Research Agenda**. The ELSA's approach is characterised by several cornerstones:

Threat Modelling and Risk Analysis: Methods and solutions are based on rigorous definitions of threats and risks. Only once threats and risks are characterised, well defined statements of properties like robustness or privacy can be given. This is foundational and best practice in domains like cybersecurity and needs adoption in machine learning (ML) and AI – in particular once adversaries need to be considered.

Striving for foundational research, guarantees, insights: In order to innovate in compliance with European values, methodological research plays a key part in building trustworthy AI/ML applications and systems. Such advances should have their footing on rigorous and foundational research, so that trust in the resulting technologies is sustained and not eroded by false promises.

Interdisciplinary aspect: The success of arriving at Secure and Safe AI technology hinges on the capability of integrating knowledge and insights far beyond the core AI/ML domains. On a more technical dimension, e.g. formal and symbolic methods from verification over cryptography to cybersecurity play key roles. On a less technical dimension, e.g. ethics, legal and human factor research are indispensable.

System view – MLTrustOps: We need to arrive at a holistic view of the design, processing, life-cycle, and impact of AI/ML systems in order to arrive at security and safety properties. Hence, we are proposing MLTrustOps to include all relevant aspects into an inclusive view of AI/ML systems and applications.

Governance and Legal Aspects of Socio-technical Systems: With the realisation that AI/ML systems do not only become part of our IT landscape but also form socio-technical systems that are increasingly deeply ingrained in our society, we

need to realise the profound effect. Governance and legal aspects need not only ensure compliance but well being of the whole society and aspiring for common good.

Understanding inherent limitations and tradeoffs in Trustworthy AI: While the focus of research and innovation needs to be developing foundations and solutions to the most pressing challenges, it is equally important to shed light on inherent tradeoffs and potential impossibilities. These can inform technology as well as the public discourse and avoid false promises.

Openness, Transparency and Accountability: An Open Source approach is a key ingredient towards a transparent and accountable approach to AI development that fosters safety and security – in particular in the context of foundation and large language models.

Beyond these guiding principles we define 3 main **Grand Challenges** as part of this Strategic Research Agenda that also targets research towards key **Use Cases** measured by **Benchmarks**[1]:

Grand Challenge – Technical Robustness and Safety: Current AI systems suffer from several fundamental issues undermining their trustworthiness, and thus preventing their adoption in cybersecurity-related and safety-critical applications. The first grand challenge formulated within ELSA aims to overcome these issues by developing new methods for creating safe, robust, and resilient AI systems, while considering specific threat models and practical attacks for the applications at hand.

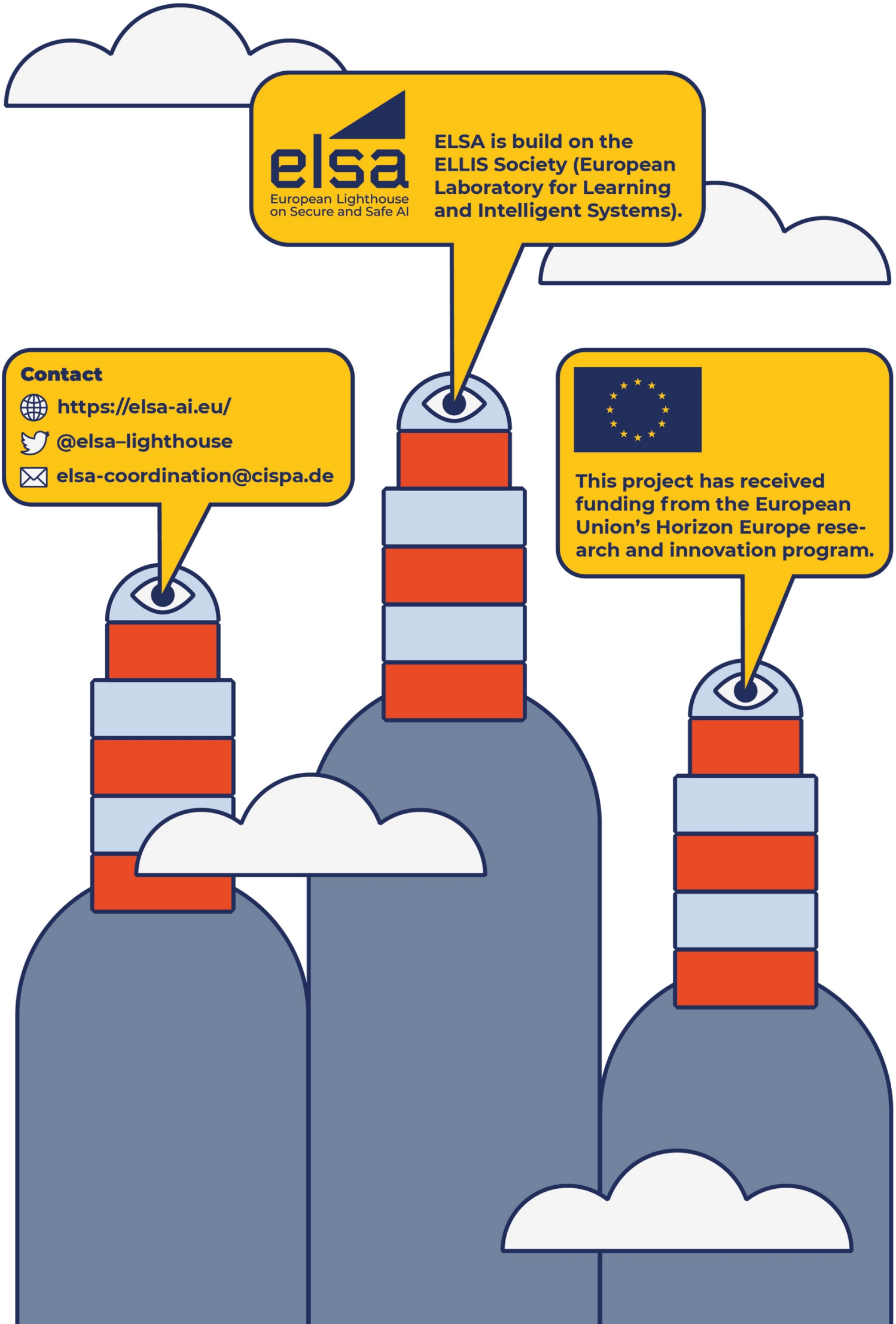
Grand Challenge – Robust Private Collaborative Learning: Modern machine learning depends on ever larger data sets collected from many sources. Our aim is to improve privacy by enabling flexible distributed learning with formal guarantees for preservation of data subject privacy and robustness to adversarial manipulation of learning.

Grand Challenge – Human Agency and Oversight: Machine learning models need to work for the society and its individuals. From the technical aspect, we improve transparency of ML models, particularly those utilising deep learning. From ethical, legal and regulatory aspects, we address the problems of AI assurance and meaningful human oversight embedded within a regulatory governance regime.

Outlook: While the research community has already achieved significant progress along this research agenda, there are equally significant gaps to close in order to provide key methodology and deploy them in practice. The recent advances and deployments of Large Language Models amplify the shortcomings and needs for Secure and Safe AI. AI remains a technology with substantial risks and it is on us to innovate in compliance with our societal values and decide where and how to use it in order to leverage its potential for societal good. We need a decisive and sustained investment in order to take leadership, lay the foundations for the future, and shape this technology in a European understanding.

---

[1] https://benchmarks.elsa-ai.eu/


elsa
European Lighthouse on Secure and Safe AI
ELSA is build on the ELLIS Society (European Laboratory for Learning and Intelligent Systems).
Contact
https://elsa-ai.eu/
@elsa–lighthouse
elsa-coordination@cispa.de
This project has received funding from the European Union's Horizon Europe research and innovation program.