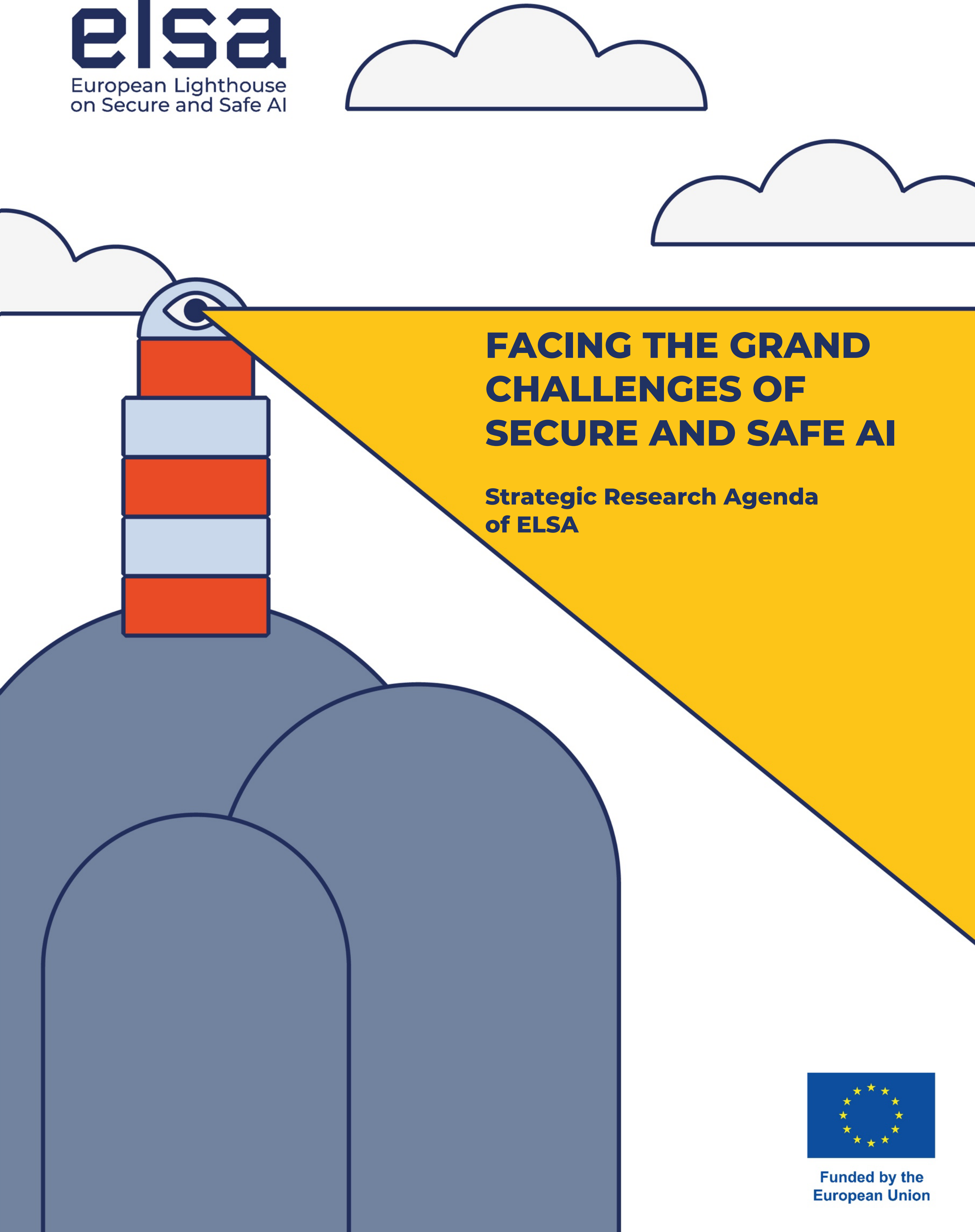
elsa

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

Content

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/

2. Scope and Problem

Contemporary developments in Artificial Intelligence (AI) affect a number of questions of human activity, which creates new risks and opportunities. While the opportunities include multiple potential and real benefits to society through the use of digital technologies and automatisation, the risks include threats to safety and security, erosion of privacy and lack of transparency as well of human agency and oversight in ethical, legal and regulatory forms. In this document, we link these fundamental problems through the approach, accepted by ELSA, to the Strategic Research Agenda (SRA).

The above mentioned problem encompasses multiple aspects:

- *Robustness, safety and security*: It is becoming increasingly evident that current deep learning systems suffer from several fundamental issues, including a lack of robustness guarantees and minimal resilience against input data perturbation, which create safety and security risks. They reinforce biases present in data. All of these technical challenges can bring substantial harm for the society

- *Privacy*: Modern AI and deep learning technologies depend on massive amounts of data, often about individuals. These data can create benefits for individuals and society, but their misuse can create digital harms

- *Transparency, human agency, oversight:* Modern AI technologies, in many cases, would only bring necessary benefits if their operation is transparent for analysis by humans. This complements the challenge of safety and security mentioned above. In addition to this, robust technical standards will not deliver safe and secure AI in Europe unless and until they are embedded within a legitimate and effective governance architectures that provide meaningful human oversight that is demonstrably in accordance with core European values: namely, respect for democracy, human rights (including the protection of safety and security) and the rule of law.

Solving these challenges is linked to a number of initiatives which are at the forefront of attention of the European Union through multiple pathways: legal, such as the AI Act; technical and scientific, such as European Networks of Excellence in AI. These aspects should be addressed through theoretical as well as empirical research. Due to these reasons, the ELSA network promotes the vision of addressing the aforementioned challenges of safe and secure AI through the link between the challenges and practical use cases, linking industry-driven practical demands and academic research.

This vision is grounded on solid theoretical and empirical foundations, accepted by the wider AI community. The NIST glossary[2] defines AI as an “Interdisciplinary

[2] https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary

field, usually regarded as a branch of computer science, dealing with models and systems for the performance of functions generally associated with human intelligence, such as reasoning and learning”. Nowadays, it is closely related to machine learning (ML) which is defined in the same glossary as “A general approach for determining models from data”. At its core are methods for inductive reasoning that aim to discover general principles from observational data. This data-driven approach has been very successful in recent years. Often algorithms are rather generic, but the resulting functionality is modulated and determined by the data. Such emphasis on the data also brings about many issues as we will discuss below. Deep learning is a particularly successful branch of ML, which leverages multi-level representations of data, providing, often through millions or billions of optimisable model parameters, a powerful mechanism to address fundamental problems of AI.

Beyond the broad scope of methodology that is grouped under the umbrella of AI and ML, the goals and objectives are also increasingly broad. Unfortunately, these are frequently not made explicit and hence very different techniques and approaches are discussed indiscriminately. One very broad categorisation that can provide guidance here is provided by Russell and Norvig[3]. AI systems can be designed and distinguished by their goal and objective along two key dimensions. First, the objective can be either to mimic humans or follow a certain specified rationale or logic. Second, the goal can be either determined and evaluated by a certain behaviour of a system or the underlying reasoning process. This categorisation provides four quadrants of prototypical AI systems:

Human-Like Reasoning	Rational Reasoning
Human-Like Behavior	Rational Behaviour

E.g. symbolic, deductive reasoning systems would fall into the category of rational/logical reasoning while many of today’s ML approaches aim at mimicking human-like behaviour. Of course hybrid approaches are being actively explored and AI/ML approaches rarely implement the extremes of this taxonomy.

Development and expansion of new deep learning techniques has made it possible to solve many decision related problems in ways unimaginable just a few years ago. However, such a fast move from fundamental and applied research into commercial products and government services has created a range of problems which can be broadly attributed to interaction between technology and society.

Important problems include:

- the need to align with ethical principles (including human rights);

[3] Russell, Stuart, and Peter Norvig. "Artificial intelligence: a modern approach." (2002).

- the need to conform with legal duties and obligations, and risk of undermining the rule of law and democracy;
- the need to provide interpretable and explainable models;
- the need to ensure causality of the decisions made by ML models;
- the need for the system to be robust in a technical or organisational sense; and
- ensuring safe and secure execution of ML models, as well as their robustness.

However, ensuring that these principles apply only to the models is not sufficient. It includes a much wider scope including provenance of data and ensuring the ethical and legal use of the developed technologies.

The recent appearance of complex foundation models trained on large amounts of data, such as large language models (LLMs) and generative models for images and video further highlighted the problems of data provenance. Despite the scale of these data collection, it is important that these data are collected in an ethical and legal way. That includes meeting the requirements of GDPR, assuring that these data lead to data reflective of the values of fairness and inclusivity.

The incorporation of these values would not be possible without an open source approach, which ensures transparency and accountability and therefore, ensures that the foundation models work for the society.

These, more generic challenges are increasingly important in the context of the recent developments of the technology as well as in the context of the European values and, more generally, the values of democracy and liberty of citizens.

3. European Vision on Trustworthy AI and the EU AI Act

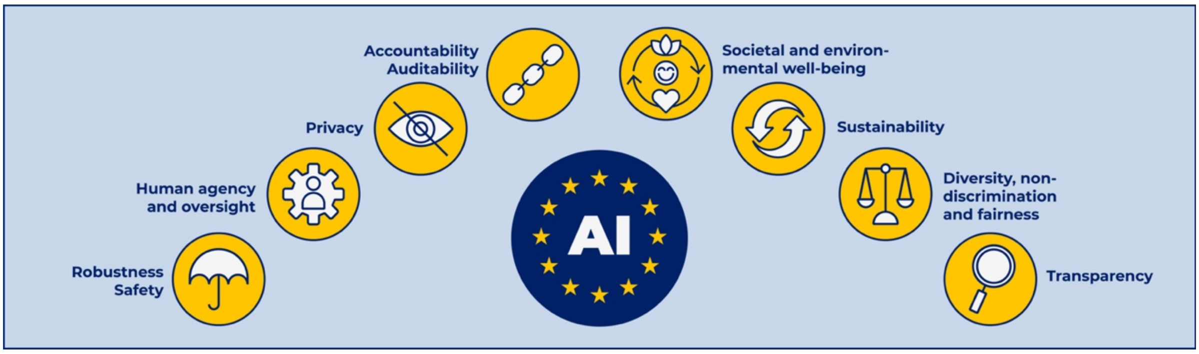

Accountability
Auditability
Societal and environmental well-being
Privacy
Sustainability
Human agency and oversight
AI
Diversity, non-discrimination and fairness
Robustness
Safety
Transparency

Figure 1 - Components of the European model of Trustworthy AI.

The High Level Expert Group in AI set up by the European Commission has laid out a **European Vision of Trustworthy AI** that should be lawful (respecting all applicable laws and regulations), ethical (respecting ethical principles and values), and robust (both from a technical perspective while taking into account its social environment). AI and ML models are being deployed widely with success and could provide a key competitive advantage to the European economy. Unfortunately, the same abilities that make AI able to bring social and economic benefits may also be used offensively, and negatively affect society. For example, the ability of AI to mimic human behaviour might induce some individuals to excessively trust its predictions, even when the model may be wrong. This may potentially cause harm to them or others. AI suffers, in fact, from several fundamental issues that can accidentally lead to unwanted system behaviours.

The European Union is interested in preserving its technological leadership and allowing its citizens to benefit from AI. However, ensuring that these technologies are developed according to human-aligned values, fundamental rights, and principles is utterly important. For these reasons, the European Union has been working on developing the **Artificial Intelligence Act** to regulate the usage of AI in different contexts. In particular, the AI Act subdivides AI-based applications into categories depending on the harm they may cause and proposes to regulate each category with a series of requirements which the system should be compliant with. AI-based applications are subdivided into: (i) unacceptable risk, (ii) high risk, (iii) low risk, and (iv) minimal or no risk.

The proposal considers the risk to be unacceptable for applications that may violate European values, for example, by violating human rights. They include applications that may alter a person's behaviour in a way that can cause harm; social scoring for general purposes done by public authorities; and the usage of "real-time" remote biometric identification systems for law enforcement in publicly accessible spaces, excluding some exceptions, e.g., border control. The usage of AI in applications with an unacceptable risk would be forbidden.

Applications are proposed to be considered high-risk when they can create a high risk to the health and safety or fundamental rights of natural persons. Some examples of high-risk AI systems include those intended to be used as safety components of products, e.g., medical devices and systems used in certain fields, including biometric identification for border control and law enforcement. The usage of AI in high-risk applications would be allowed, but they must be compliant with certain mandatory requirements and an ex-ante conformity assessment. These requirements are detailed below.

- **Equality**. Individuals with similar characteristics should receive the same response from the system regardless of their gender, ethnicity, and other characteristics that, for ethical reasons, should not affect it.
- **Transparency**. The system should provide the user with information about the process used to provide the output. This allows humans to oversee the process and spot eventual errors in the system's output.
- **Robustness, Safety, and Security**. The AI/ML-based system should be secure to deliberate attacks and preserve safety of its users and the environment in safety-critical tasks. This amounts to adopting secure-by-design countermeasures and providing robust decisions even in the presence of unexpected inputs or deliberate attempts aimed to compromise its integrity, availability, or the privacy of its users.

Applications considered at low risk are proposed to only have a transparency requirement. This means they should notify humans that they are interacting with an AI system unless this is evident, and eventually disclose that the system's output has been artificially generated or manipulated.

Finally, minimal, or no-risk applications do not have any requirements.

Risk Category	**Application Examples**	**Requirements**
Unacceptable risk	Social scoring	Prohibited
High-risk	Recruitment, medical devices	Permitted but subject to equality, transparency, robustness, safety, and security obligations, and ex-ante conformity assessment
AI with specific transparency obligations	Impersonation (bots)	Permitted but subject to information/transparency obligations
Minimal or no risk		None

The desiderata specified in the EU AI Act for all the risk levels are clear and essential to create systems that comply with European values. However, developing AI-based systems for high-risk applications that meet the aforementioned requirements presents different *technical* challenges. One of the biggest issues hampering the development of systems compliant with these requirements is the lack of secure and safe AI technologies that we can trust in real-world applications.

It is well-known that the security of AI and of current deep learning models can be easily undermined at training and test time. For example, by slightly altering the systems' input, attackers can compromise the system to behave as they desire, with unexpected potential consequences for the safety of the system users (e.g., when operating in safety-critical environments). However, even if we consider state-of-the-art defensive measures and methodologies, it is still unclear whether they really enable developing and maintaining secure and safe AI-based systems in real-world applications.

4. ELSA Approach to Secure and Safe AI

ELSA addresses important aspects of safe and secure AI. The topics of research are organised in Grand Challenges measured by Benchmarks[4] and cover the following aspects: robustness and safety, privacy and infrastructure, as well as human agency. Before we go into the details of the Grand Challenges, we define overarching principles of our approach, that we outline here first:

4.1. Threat modelling and risk analysis

In contrast to the vast part of machine learning that to a large extent is motivated by expected risks, systems that are deployed in the real-world become part of an attack surface and face adversarial attacks. In such situations, an attacker might seek to systematically exploit worst case behaviour that deviates significantly from the average behaviour of the system.

Such setting often results in more challenging learning problems and ensuring robustness in such worst case situations often require different methodology. Unfortunately, there are prominent cases where these issues are addressed by heuristic or empirical approaches. These lead then to “arms race” or “cat and mouse game” settings, where attacks and defences try to outcompete each other. While in a few cases this is unavoidable, it is neither satisfying nor sustainable, as any of such solutions might be broken in the future and therefore any promises of trustworthiness of the system will not hold.

Hence, we strive for rigorous approaches that can eliminate whole types of attacks and vulnerabilities for good – in order to allow for guarantees of the trustworthiness of the system wherever possible. This is only facilitated by following paradigms and ideas that are well established in e.g. the cybersecurity and safety research communities. There are at least two key ingredients:

Threat model: The capabilities of an attacker and defender need to be clearly formalised. Only then can we reason in a rigorous way about solutions that can rule out classes of attackers. In addition, this methodology makes assumptions explicit and provides systematic progress in these challenging domains.

Risk analysis: In many situations, we are faced with the challenge that an absolute notion of security, privacy, or safety is not achievable in any practical and meaningful way. Here, we need to provide technical means that address threats that are relevant so that we can provide appropriate measures and protection to reach our goal. This technological approach is also compliant with many legal definitions that require appropriate protection.

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

4.2. Striving for foundational research, guarantees, and insights

ELSA promotes a strong foundation for trustworthy AI that builds on foundational research and formal guarantees, whenever possible. We believe this is essential for building truly *trustworthy* AI that will keep meeting its demands even as the world changes, especially in high-risk applications. Formal guarantees are also important in avoiding arms races of attacks and defences.

Methods providing formal guarantees are available for many problems encountered in trustworthy AI. Certifiably robust methods are guaranteed to be robust to certain perturbations, meaning for example that such perturbations cannot mislead a classifier[5]. Byzantine resilient methods for distributed and federated learning can tolerate a certain number of arbitrarily misbehaving clients[6]. Differential privacy allows proving that the privacy loss from releasing some computation results is bounded, even against arbitrary future adversaries[7].

While formal guarantees are an important tool, they are not a silver bullet. All guarantees depend on a particular formal model of potential adversaries, whose realism needs to be considered in light of the threat model. Work on attacks to analyse the practical risks is an important complement to formal guarantees.

Often it may turn out to be impossible to match desired formal guarantees with the required level of performance of the system. The next best option is extensive and systematic testing, but it is important to understand its limitations: no amount of testing can cover all corner cases (such as all traffic situations encountered by an autonomous vehicle).

Deployed AI systems act as parts of a larger system. When formal guarantees are available, they provide safety in integrating the AI component. When no formal guarantees exist, it is possible and often useful to apply other safety engineering approaches such as redundancy and monitoring to ensure the reliability of the overall system, while ensuring system resiliency and fast recovery when under attack, even one that was not considered a priori.

4.3. Interdisciplinary aspect

While this SRA is at its core about AI and machine learning, the previous sections have already made clear that the scope of the methodology to achieve the goal of safe, trustworthy and secure AI is substantially broader.

Such solutions can be drawn from **core research in AI and ML**.
It is therefore essential to build upon theoretical understanding of key challenges in the domain. For example, results on convergence rates of learning algorithms can help advance our understanding of robustness and reliability of

[5] Cohen, J., Rosenfeld, E., Kolter, Z. (2019). Certified Adversarial Robustness via Randomized Smoothing. In Proceedings of the 36th International Conference on Machine Learning.

[6] Blanchard, P., Guerraoui, R., Stainer, J., et al. Machine learning with adversaries: Byzantine tolerant gradient descent. In Advances in Neural Information Processing Systems, pages 119–129, 2017.

[7] Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Proceedings of the 3rd Theory of Cryptography Conference, TCC 2006.

learning algorithms. Another important aspect is causality of the data generating processes. Such work can be furthered to discover principled ways of recovering relationships between data. Advancement of these techniques is key towards ensuring fairness in AI and ML.

Beyond the aforementioned aspects, the community has drawn and needs further integration with the **broader technical research in computer science and related fields**. It is particularly noteworthy that core methodology that is becoming key e.g. in certification, privacy and robustness has been developed in parallel or even prior to the recent advances of AI/ML. For example, formal methods provide a rigorous way to prove properties about neural networks, e.g. probabilistic or exact proofs can be given w.r.t. the invariances of a neural network. Another example is cryptography based on information theory which provides principled means to information security and confidentiality. Beyond this, it has provided the bases of rigorous privacy notions (e.g. differential privacy), that remain valid even against arbitrarily strong future adversaries.

Beyond the disciplines with a technical focus, a number of **human-centric disciplines** are indispensable for Secure and Safe AI to cover the interpretability, ethical and legal aspects. Shaping up recommendations to European policy-makers and the broader stakeholder community involves building upon current insight into regulatory governance and critical algorithm studies. To advance towards the goal of safe and secure AI, these recommendations should be guided by understanding how key stakeholders perceive risks and opportunities associated with existing and proposed legal governance frameworks for data-informed services in Europe.

Finally, applying AI methods to practical use cases requires collaboration with experts and researchers from the application area to ensure that theoretically sound research meets the practical expectations.

4.4. System view: MLTrustOps

Developing and maintaining a secure and safe AI-based system not only demands for an initial, well-designed and documented process, but also for a well-structured, continuous development and monitoring infrastructure during operation. We argue here that the modern development framework known as *Machine Learning Operations (MLOps)* can provide a viable option to implement a trustworthy AI/ML continuous development cycle (Figure 2). In particular, MLOps includes a set of practices and tools that help develop, deploy, and maintain machine learning models in a production environment, with a high degree of automation, thereby also minimising the technical debt potentially introduced by the use of AI/ML models in more complex system architectures. The MLOps cycle consists of six main steps, as detailed below.

1. **Data Preparation.** The data are collected and prepared to be processed by an AI/ML model. This step requires identifying the appropriate and reliable data

sources, cleaning and transforming the data, and ensuring they are in a format that can be used for training.

2. **Model Training.** In this step, the model is developed and trained. This requires selecting the appropriate machine learning algorithms, tuning the hyperparameters, and evaluating the performance of the models.
3. **Model Packaging.** This step creates a package to ensure the model and the datasets used to train it have all the dependencies that need to be available at runtime. Thus the model predictions made in the development environment are replicable in the production environment.
4. **Model Validation.** Once the models are trained, they need to be validated to ensure that they are accurate and reliable. This involves evaluating the performance of the models on a separate set of data, known as the validation set, and refining the model if necessary.
5. **Model Deployment.** After the models have been tested and validated, they are deployed in the production environment, where they are used to make predictions or decisions. This involves setting up the infrastructure, such as servers and storage, and configuring the machine learning models to work with the production data.
6. **Monitoring and Optimization.** Once the models are deployed, they must be monitored to ensure they perform as expected. This involves setting up the monitoring infrastructure, such as alerts and dashboards. In this way, if the models are not performing as expected, the developers can promptly intervene.

Similarly to DevOps in software engineering, the MLOps development cycle neither embeds any security and privacy testing nor any support to develop robust and trustworthy models. However, we argue that such dimensions can be easily integrated into MLOps, to enable the design of an ML*Trust*Ops framework.

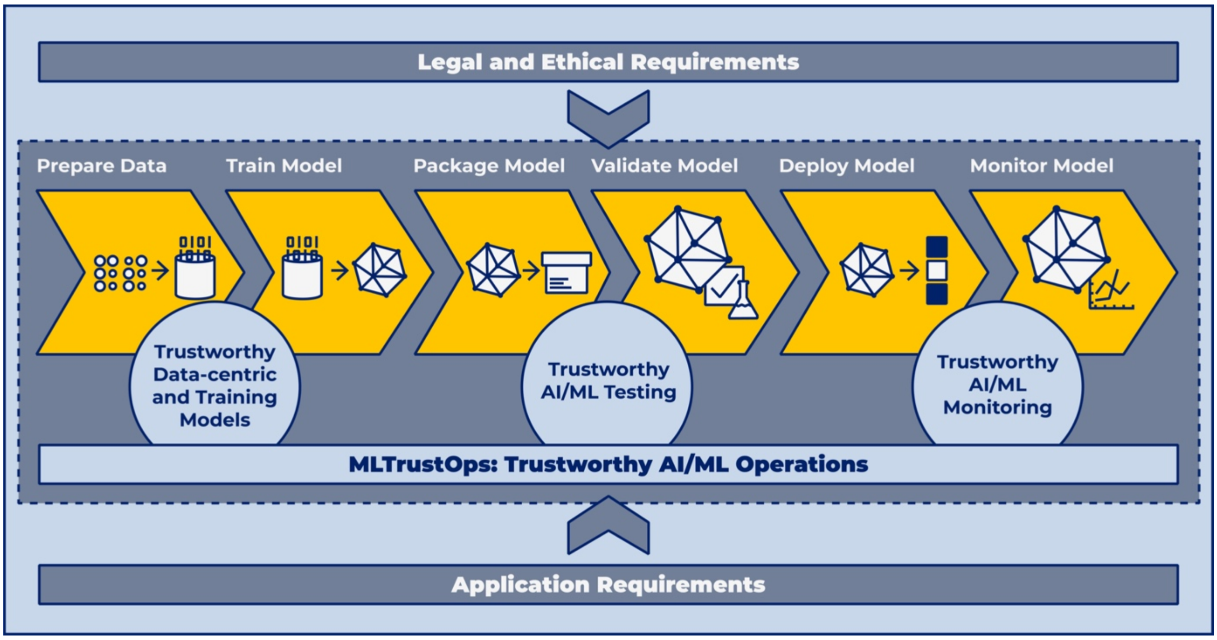

Legal and Ethical Requirements
Prepare Data
Train Model
Package Model
Validate Model
Deploy Model
Monitor Model
Trustworthy Data-centric and Training Models
Trustworthy AI/ML Testing
Trustworthy AI/ML Monitoring
MLTrustOps: Trustworthy AI/ML Operations
Application Requirements

Figure 2 - ML*Trust*Ops development cycle.

The process starts from the definition of legal, ethical, and more practical application requirements, along with an in-depth analysis of the potential threats or system misuses, as discussed in the previous sections of this document. These requirements should ensure trustworthiness and human alignment. In particular, they need then to be formalised in terms of metrics and trustworthiness dimensions that should be integrated during design and system operation. They include privacy, fairness, explainability and robustness requirements, among others. As shown in Figure 2, we advocate here that the MLOps framework can be complemented by adding three main pillars to support the different design steps and encompass the necessary trustworthiness dimensions and metrics, as detailed below.

1. **Trustworthy Data-centric and Training Methods.** This step amounts to sanitising datasets (promoting also data-centric approaches) and learning AI/ML models that naturally account for specific trustworthiness dimensions, e.g., models that are robust and certifiable against attacks and out-of-distribution samples, and also fulfil fairness and privacy requirements (e.g. according to the data minimisation principle). Explainable AI/ML methods can also play an important role here, as they may help debug what AI/ML models learn from data, and detect potential problems.
2. **Trustworthy AI/ML Testing.** This step aims to automate AI/ML testing and evaluation considering the trustworthiness requirements elicited in the previous phases, similarly to what is normally done with unit testing and integration testing in software engineering. Unit testing would amount to testing the AI/ML model in isolation against the identified threats or potential corner cases which may be incurred during operation (e.g., detecting specific biases after training), while integration testing would amount to testing the AI/ML model when operating in more complex systems and interacting with other non-ML-

based components. However, further research is needed to better understand how trustworthy AI/ML testing can be brought to a higher level of automation.

3. **Trustworthy AI/ML Monitoring.** This step amounts to enabling detection of attacks which may be executed during operation and, more generally, system malfunctioning (e.g. distribution shifts that may induce unfair behaviours), allowing a timely reaction. This monitoring process aims to speed up reaction to observed threats, ensuring resilience and prompt recovery of the whole system under and after attack, while the other two pillars of the MLTrustOps framework ensure a more proactive approach against potential threats and issues that may be incurred during operation.

While this framework helps us reason on how to protect AI/ML models and implement trustworthiness dimensions within them, it is also worth remarking that there are complementary protection measures which should always be considered. In particular, AI/ML models are themselves software components and, as such, the libraries used to implement them should undergo the standard assessment and testing procedures which software is normally subject to, e.g., security scanning to detect the presence of software vulnerabilities, and functional testing to assess the scalability of the overall system. Furthermore, AI/ML models are normally used in more complex systems, including other AI/ML models as well as non-ML components. This demands not only for integration and regression testing of the system after AI/ML model updates, but for a continuous and managed validation process aimed to reduce any potential high technical debt induced by the presence of AI/ML models. Finally, it is also worth noting that AI/ML models may not be bulletproof and guaranteed to be trustworthy by design under all conditions, e.g., there may be failure modes which are impossible to encompass during system design (i.e., the so called *unknown unknowns*, to paraphrase the former United States Secretary of Defense Donald Rumsfeld). In these cases, exploiting a more holistic and traditional systems perspective, from an engineering viewpoint, may actually help us to design more reliable and safe AI/ML-based systems, e.g., by exploiting redundancy of sensors and complementary control and decision-making mechanisms, which may include both human and technical elements.

4.5. Socio-Technical View of Governance and Legal Aspects of AI Systems

Developing and maintaining integrated governance frameworks to ensure meaningful human oversight which, in turn, will secure and maintain safe and secure AI, is a serious and formidable challenge. Addressing it will require the successful integration of research from multiple disciplines: from the technical, natural and medical sciences, through to the social sciences, law and humanities, in ways that can be practically and meaningfully adopted by organisations, groups and individuals in real-world settings. The embedding of AI into complex socio-technical systems that deliver data-driven services in real-time operating at a planetary scale in ways that directly affect the safety, rights and well-being of human communities and their environment generate novel technical, legal, ethical and governance challenges.

The primary focus of the research is concerned with unresolved challenges surrounding the capacity of humans to understand, interpret and comprehend the underlying logic of an output produced by a ML model, particularly those which utilise deep learning. If advances in ML and AI are to deliver the promised benefits of enabling and enhancing human well-being and for the benefit of the many and not merely the few, then they will need to operate in real-world settings in conjunction with individuals, organisations and communities of all shapes, sizes, and capabilities, including society's most vulnerable. Unless those who are directly and indirectly affected by and interact with socio-technical systems that rely on ML models can understand why they generate particular outputs (both in general sense and in specific circumstances) and can consistently rely on those outputs being produced in accordance with their understanding and legitimate expectations of how those outputs are generated and the impact that they produce, then they will not be capable of understanding how these systems operate. Nor will they be in a position to reliably anticipate how they will operate in future. This, in turn, makes the attainment of meaningful human oversight impossible, creating serious dangers that these systems will fail consistently to operate in accordance with fundamental human values upon which democratic societies claim their allegiance and upon which they are rooted. Accordingly, until the underlying ML models can be made sufficiently interpretable and comprehensible to human users so that they acquire the level of understanding necessary for meaningful human oversight, the embedding of these models within socio-technical systems deployed in real-world settings cannot be characterised as 'safe and secure AI'.

Although much attention has been devoted to the so-called 'ethics principles' that should inform and guide the development and deployment of AI technologies, relatively little attention has been given to the governance institutions, oversight mechanisms and their legal status and interaction with other legal norms and institutions. Yet if these laudable principles are to be given concrete expression in the development and deployment of real-world AI technologies, then a vital and unresolved challenge concerns the need to ensure that we have in place practical, effective and legitimate integrated governance mechanisms and institutions that are capable of providing meaningful human oversight of these socio-technical systems. Recent legal and policy reform proposals (including the proposed EU AI Act) appear to place considerable faith in the form of technical standards, certification and assurance mechanisms established and provided by non-state actors. Yet the efficacy and legitimacy of these regulatory governance mechanisms, particularly in ensuring the 'quality' of data-driven services to ensure safe and secure AI, remains unknown.

4.6. Understanding inherent limitations and tradeoffs in Trustworthy AI

Many frameworks on Trustworthy AI, including the one defined by the EU High-Level Expert Group on AI, list a number of key requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; environmental and societal well-being; and accountability[8].

These provide a powerful list of properties we expect trustworthy AI systems to satisfy. Unfortunately these properties carry inherent limitations: given a finite collection of training data, it may be impossible to combine high utility of the system with strong guarantees of robustness or privacy, for example.

In many cases, there is a trade-off between the utility of the system, such as its accuracy, and the degree to which different dimensions of trustworthiness are satisfied. For example, requiring greater robustness or privacy will in most cases necessarily reduce the system's utility in its main task. Similarly, there is a fundamental trade-off between fairness and average utility.

To make things even worse, many of the properties are contradictory and cannot be achieved at the same time while maintaining non-trivial utility of the system. For example, robustness can be at odds with fairness[9], while privacy can be at odds with transparency[10] and fairness[11]. These examples highlight need for research actions to enable the realisation of the Guidelines for Trustworthy AI:

- Fundamental research on the theoretical limits of different dimensions and their interactions to establish what is possible.
- Collaboration between diverse technical and non-technical communities to develop an understanding of which technically feasible options are best for the society.

Understanding the limitations is also important for both avoiding false promises as well as avoiding impossible requirements.

4.7. Openness, Transparency, and Accountability

An open and transparent development of AI and machine learning is key to safe and secure technologies. Unfortunately, the current development of foundation and large language models is largely driven by industry players that did not provide the level of openness, transparency, and accountability that supports research and an overall trustworthy approach.

[8] Ethics Guidelines for Trustworthy AI by the High-Level Expert Group on AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[9] H. Xu, X. Liu, Y Li, A. Jain, J. Tang. To be Robust or to be Fair: Towards Fairness in Adversarial Training. In Proceedings of 38th International Conference on Machine Learning (ICML 2021), 2021.

[10] R. Shokri, M. Strobel, Y. Zick. On the Privacy Risks of Model Explanations. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2021), 2021.

[11] E. Bagdasaryan, O. Poursaeed, V. Shmatikov. Differential Privacy Has Disparate Impact on Model Accuracy. In Advances in Neural Information Processing Systems 33 (NeurIPS 2019), 2019.

ELSA is committed to a mission of bringing open and transparent research to the community and connecting to industrial actors in order to improve this situation. Below, we highlight key elements why such an approach is not negotiable and what is still missing to reach our objectives.

Impact on Safety and Security: Cryptography community has long established openness of algorithms and their implementations as key to security. Research has given vital contributions to safety and security of real world systems. Only a transparent and accountable approach can develop this synergistic relation. Without following this paradigm, society will be exposed to unnecessary risks and it will be increasingly difficult to establish trust in these complex systems.

Digital Sovereignty: We see an emerging ecosystem that is being built around key AI technology. Foundation and Large Language Models are a prime example of this. Unfortunately, some of these models are owned and hosted by companies and are only available via API access. The model provider will be able to monitor all use of the system, thereby compromising the privacy and confidentiality of queries fed to the system. Becoming dependent on such infrastructures can threaten digital sovereignty of the EU and its member states.

Complex Supply Chains: Like in traditional industries, we will see increasingly complex supply chains in AI and Machine Learning systems, where systems are being built out of components or even systems. Ensuring properties and behaviours of such agglomerates and compositional approaches will be hampered by opaque and inaccessible systems.

5. Grand Challenge: Technical Robustness and Safety

Data-driven AI systems based on deep learning models have recently recorded unprecedented success in many different domains, and they are also being increasingly applied in cybersecurity-related and safety-critical tasks. However, it is becoming increasingly evident that current deep learning systems suffer from several fundamental issues, including a lack of robustness guarantees, minimal resilience against input data perturbation, and a reinforcing effect on biases present in data, which could prevent their broad adoption especially in cybersecurity-related and safety-critical applications.

To overcome these issues, the first grand challenge formulated within ELSA aims to develop new methods for creating safe, robust, and resilient AI systems with robustness guarantees, and considering specific threat models that allow simulating feasible and practical attacks for the applications at hand, as described in the following research challenges.

5.1. Research Challenges

5.1.1. Security testing and robustness evaluation

While recent regulations require the development of procedures to assess the robustness of high-risk AI-based systems, it is still unclear how this should be practically implemented. On the one hand, formal evaluation methods do not scale to more realistic, complex models and application-constraints. On the other hand, empirical evaluation methods tend to provide typically an overly optimistic estimate of the actual AI/ML model robustness, and they come with no formal guarantees that the analysis is reliable.

Within ELSA, we envision a framework that can be used to assess the robustness of AI/ML models, mitigating some of the issues discussed above. In particular, we aim to assess security or robustness of a machine-learning model against different potential scenarios, following a what-if analysis.
The potential scenarios, including different threats or corner cases which may be encountered during operation, need to be identified a priori, performing an appropriate risk analysis and threat modelling of the system at hand. While it is important to evaluate the performance of the model on *in-distribution* test samples, we claim that also modelling its behaviour on out-of-distribution inputs and scenarios is highly relevant especially in security-sensitive and safety-critical applications. Indeed, after deployment, AI-based systems can face situations that are different from the ones considered at training time.

Out-of-distribution samples can be generated artificially to mislead the system (i.e., optimising the input perturbation against the target model, as done for adversarial examples), or they can be natural samples that do not belong to the same distribution as the training set. For example, for an autonomous driving car trained to recognize street signals in a city during sunny and rainy days, the images acquired in another city or while it is snowing can be considered relevant out-of-distribution samples. The main challenge in testing the performance of a system on out-of-distribution samples is that collecting them may be time-consuming. Furthermore, it is not even clear how such inputs can be detected; i.e., for any given input to a machine-learning model, it is not trivial to understand whether it is an out-of-distribution input or whether it can be reliably classified by the model.

Even testing the performance against out-of-distribution samples generated artificially presents some difficulties. Some input perturbations can be modelled mathematically (e.g., using distance functions computed between input samples). In this case, the robustness of some machine learning models can be tested using verification techniques, which provide formal guarantees about the system's robustness. However, they cannot be applied when the input perturbations can not be modelled mathematically or when the system is too complex. In these cases, the system can only be tested leveraging empirical methods (e.g., gradient-based and black-box optimizers). This is closely related to testing software systems via dedicated fuzzers, which of course have their own limitations and biases. Further research is thereby needed to devise scalable approaches that allow evaluating the security of state-of-the-art models in a reliable manner. Within ELSA, we are developing new approaches based on formal verification, empirical methods, and their combinations. Given that the current empirical methods to evaluate the security of machine learning systems sometimes fail, compromising the reliability of the security evaluation, we have also developed *debugging* tools that help identify common failures in the evaluation process (e.g., failures in the gradient-based optimization of adversarial attacks).

5.1.2. Robust and certifiable machine learning

The current solutions available to create robust and certifiable machine learning models do not scale to complex neural networks and complex natural robustness and safety properties. The main challenges are thus to overcome these limits, of which we provide a more extensive description in the following, along with the directions that can be pursued to overcome them.

Beyond Lp-norm robustness. Most certification algorithms consider bounded-norm perturbations. While these sometimes suffice as proxies for imperceptible image modifications, tasks such as object detection rely on different similarity measures, for example, cosine similarity for word embeddings and Mahalanobis distance for images. It is thus desirable to define measures and certification algorithms for semantic robustness, considering similarity measures that reflect visual or geometric aspects characteristic of the application, such as object movement or lighting conditions. More generally, robustness evaluation frameworks for more complex properties induced by the use cases are needed.

Beyond supervised robustness. Existing robustness formulations focus on the supervised learning setting. However, collecting and labelling large datasets necessary to ensure the high robustness performance is costly and may not be feasible in applications characterised by a wide variability of possible inputs. Instead, it is desirable to formulate robustness measures and evaluation frameworks directly in some appropriate semi-supervised and unsupervised settings, where the definition of robustness needs to focus on the quality of the learned representations rather than classification (prediction) because of the lack of labels. This may involve working with similarity measures such as Mahalanobis distance and will be challenging both theoretically and computationally to achieve provable robustness guarantees.

Certifiable distributional robustness. In many different applications, the system can be exposed to a distribution shift. The notion of distributional robustness will need to be adapted to such settings so that models not only learn stable and meaningful representations that are resistant to perturbations and noise but also to distribution shifts, where we anticipate causal reasoning can play a part. The challenge here will be theoretical, in how to map the requirements from the use cases into problem specification, and then methodological, which will investigate appropriate numerical and/or statistical algorithmic approaches to compute certifiable robustness guarantees.

Scalability. Robustness certification often does not scale to complex networks. We will comprehensively investigate and evaluate the scalability of robustness certification and evaluation frameworks for typical use cases (such as object detection) with respect to input dimensionality and network depth, and for a variety of activation functions. Furthermore, scalability is key to extending certification and robustness to include the training process.

Efficiency and precision trade-off. Robustness certifications and evaluation involves a variety of methods, including exact, approximate and statistical approaches. While exact methods offer completeness, trading off exact precision for approximate bounding results in more efficient any time methods, and completeness can be recovered by combining fast approximate methods such as convex relaxation with branch-and-bound computation. Statistical methods provide estimates of robustness that may be unsound but fast and, in many cases, sufficient for the application being considered. Within ELSA, we will comprehensively investigate and evaluate the efficiency and precision trade-off of robustness certification and evaluation frameworks for typical use cases (such as object detection).

5.1.3. Uncertainty estimation and decision making

The decisions of AI/ML models, including neural networks, are unreliable when the input sample is out of the training distribution or corrupted by noise. Estimating the uncertainty of their decisions, enables probabilistic reasoning, supporting the interactions with humans in the loop and to quantify the risks associated with decisions. Conformal prediction and Bayesian approaches can estimate predictive uncertainty and their principles can be used to make deep neural networks more robust and reliable. However, their computational complexity often overshadows their advantages. Within ELSA, we aim to develop new Bayesian approaches to deep learning, assessing uncertainty when transferring models across related tasks, and reasoning about uncertainty in data-driven models informed by prior knowledge.

5.1.4. Relationships among robustness, privacy, explainability and fairness

The theoretical relations and potential tradeoffs among robustness and privacy, fairness, and explainability are still unclear. Fundamental theoretical questions are still unanswered, including: how privacy mitigation strategies (e.g., differential privacy) can affect and be affected by robustness; if robustness can help increase or reduce the fairness of a model and if new notions of fairness should be introduced; and how robustness requirements can affect, reduce, or improve the explainability of a model.

The main research challenge is thus to understand and investigate how metrics that are not technical (like accuracy and computational requirements), but that are more related to the notion of trustworthiness in machine learning and then to human-oriented metrics (robustness, privacy, fairness, and explainability), are actually related to each other. In fact, in the past, it has been observed and studied that technical metrics (e.g., accuracy) are impacted by trustworthiness metrics (e.g., fairness), creating a certain tension, but the relationship among trustworthiness metrics is much less investigated. Recently, it has been observed that when multiple trustworthiness metrics need to be optimised, other tensions arise, e.g., tension between privacy and fairness, robustness and fairness, and robustness and privacy.

Within ELSA, we will push forward fundamental research in this direction, focusing on establishing a framework able to better understand and evaluate the tensions (e.g., with trade-off bounds) and the possibility of achieving simultaneously (e.g., via consistency results) good performances in terms of trustworthiness metrics.

5.2. Further Research Directions

Building upon the ELSA approach and perspectives, we can envision two main additional research directions in the near future towards developing more safe and reliable AI/ML models.

The first is understanding how to systematise and automate the evaluation and testing of AI robustness. The techniques existing at the state of the art, developed to this end, can be applied only to simple AI models and mathematically tractable perturbation models. Complex models are evaluated with time-consuming techniques that sometimes silently fail, causing a false sense of security. Moreover, the security of AI components is usually evaluated by considering them as a standalone module, which does not interact with any other system component. New protocols and methodologies are thus also needed to evaluate the security of systems with one or more AI-based components, following a more principled engineering approach.

The second challenge is to devise models with trustworthiness guarantees, which can operate reliably also in out-of-distribution scenarios, i.e., under outlying and corner cases with respect to the data used to train the model. Although different methods have been proposed to mitigate these issues, and more will be proposed within ELSA, further effort will be needed to make these systems robust enough to be trusted in practical cases, along with systematising their design process and testing.

To conclude, incorporating AI in high-risk applications will demand further work to properly quantify the risk associated with AI-based, automated decisions, and take appropriate mitigation measures, following a well-principled engineering approach. Understanding how to satisfy the security and safety requirements specified in the AI Act is a grand challenge that has to be solved to allow the usage of AI-based systems in high-risk applications.

5.3. Use Cases: Autonomous Driving and Cybersecurity

In the following, we describe the two use cases related to the first grand challenge posed by ELSA: the autonomous driving and cybersecurity use cases.

5.3.1. Autonomous Driving Use Case

The Autonomous Driving use case aims to assess machine learning performance in the vision/perception tasks that play a crucial role in advanced driving assistance systems (ADAS) for passenger cars and autonomous vehicles (AVs) such as robo-taxis, shuttles, and delivery droids.

Unleashing these safety-critical systems onto public roads is a considerable challenge, as they must deal with diverse — sometimes hazardous — driving conditions. Furthermore, these largely data-driven systems, running on limited onboard computers, must withstand severe disturbances, ranging from sun glare to sensor blockage, and physical or digital adversarial attacks. While data growth and design advances continue to improve the raw performances of driving stacks, current machine learning methods suffer deficiencies in robustness, generalisation, transparency, and model verification. Evaluating these properties is also challenging due to a lack of tools and testbeds.

This use case aims to quantitatively compare the robustness of perception models trained on identical datasets, especially when confronting distribution shifts and perturbations. We focus on two primary tasks for vision-perception of autonomous driving: **semantic segmentation** and **object detection**. Semantic segmentation classifies every pixel within an image into a semantic class such as pedestrian or car, producing a segmented image. Object detection identifies and locates objects within an image, typically using bounding boxes, and predicts the most probable semantic class for each identified object.

For thorough benchmarking regarding robustness to perturbations and distributional shifts, we will evaluate visual models on four criteria:

1. Resistance to diverse natural perturbations like sun glare, sensor soiling, or adverse light and weather;
2. Generalisation to unfamiliar domains, such as a different city or camera;

3. Ability to detect objects of unknown categories and other out-of-training-distribution inputs;
4. Ability to characterise its own predictive uncertainty.

We will incorporate a set of baseline approaches in our benchmark, whose release, along with the analysis of the results, will foster new recommendations for safe and robust models for AV/ADAS. We are confident that this will not only accelerate the progress in this field but also attract the interest of researchers from related disciplines.

5.3.2. Cybersecurity Use Case

The Cybersecurity use case aims to evaluate machine-learning methods when they are used as a first line of defence against malicious software (malware), focusing on the Android operating system. This application is often required when a large number of Android applications must be analysed every day, demanding high levels of automation. On this task, machine learning usually performs well, learning common patterns from data and enabling detection of potentially never-before-seen malware samples, but it has been shown that those detectors:

1. tend to exhibit a rapid decay of performance over time due to the natural evolution of samples;
2. can be bypassed by even only slightly manipulating malware samples in an adversarial manner.

The practical impact of these two issues is that current learning-based malware detectors need constant updates and retraining on newly-collected and labelled data.

We propose to tackle these two issues with a benchmark that will provide tools for comparing learning-based Android malware detectors in a realistic setting and evaluate whether methods showing robustness with respect to adversarially-manipulated samples retain certain robustness properties also against real-world attacks. For this reason, we have designed three main tracks, consisting of evaluating both *adversarial* (in two different settings) and *temporal* robustness. We will include in the benchmark a number of state-of-the-art approaches as baselines, to provide an initial scoreboard to which participants of the challenge should compare.

A continuous evaluation of malware detection models based on different specific metrics - that are computed on the periodically-scheduled evaluation rounds - will allow us to understand whether improving adversarial robustness against certain perturbation models will also help improve robustness against the real, temporal drift observed from real-world data over time, or if different techniques should be developed, instead. Our ultimate goal is to understand how to build and deploy AI-based malware detectors that can be maintained with less effort and are able to react more promptly to novel threats. This competition will help advance the methodologies for security testing and robustness evaluation of AI/ML models, for designing robust and certifiable AI/ML models, and for their uncertainty estimation.

5.4. Benchmark Metrics

In our benchmarks, models can be ranked by different metrics that provide complementary information about their performance.

Autonomous Driving. For this use case, we will unify fragmented benchmarks existing in the literature, typically used for assessing performance (accuracy, robustness) on a single type of perturbation or corner-case, repurpose published datasets to accommodate new assessments, modify real scenes with elements of interest, and mine rare corner-cases. This wealth of diverse driving data will allow us to assess the models' performances in different conditions, such as light flares, sun glares, soiling, weather perturbations, and the presence of unknown objects. The considered metrics will cover different analysis perspectives, including calibration, uncertainty estimation, robustness to distribution shifts, and ability to detect out-of-distribution cases. Obtaining a model that performs well considering this mix of metrics would lead to a perception model that can be trusted to be resilient in difficult conditions (e.g., extreme weather conditions, changes of operation domains, rare but potentially catastrophic situations). Alternatively, the model may launch an alert, or fail “gracefully” in front of the so-called *unknown unknowns*. To capture these behaviours, we choose specific metrics for out-of-distribution detection, for in-domain long-tailed objects and corner cases, and for robustness, such as the Area Under the Receiver Operating Characteristic (*AUROC*) and the Precision-Recall (*AUPR*) curves, the false positive rate (*FPR*) at a fixed true positive rate, and the standard accuracy.

Cybersecurity. For this use case, the robustness of machine learning-based malware detectors can be measured by quantifying their performance decay in the presence of adversarial input perturbations (adversarial robustness) and new malware families unknown at training time (temporal robustness). The adversarial robustness evaluations will thus be conducted by relying on well-defined perturbation models, whereas for the temporal evaluation, we will apply specific data sampling rules in order to reproduce a real-world setting. As the specific task performed by malware detectors is binary classification, and the test sets are unbalanced, we can select metrics such as *Precision*, *Recall*, *F1 score, and true positive and false positive rates*. All these metrics can be used also to evaluate robustness by measuring them on adversarially-manipulated samples, within a given perturbation model (e.g. injecting a maximum number of API calls without compromising functionality of the malicious samples). Finally, we will aggregate the metrics computed for each submitted model on different temporal data splits, in order to estimate their temporal trend by applying specific metrics such as the Area Under Time (*AUT*).

Grand Challenge: Main Summary and Value Proposition

Research Challenges

- Security Testing and Robustness Evaluation
- Robust and certifiable machine learning
- Uncertainty estimation and decision making
- Relationships among robustness, privacy, explainability and fairness

Further Challenges

- Systematize/automate AI/ML security testing
- Development of AI/ML models with trustworthiness guarantees

Use Cases

- **Autonomous Driving**: out-of-distribution image segmentation
- **Cybersecurity:** malware detection under adversarial/temporal drift

Value Proposition: ELSA provides a unique perspective and approach to developing and testing safe and secure AI methods in the context of cybersecurity-related and safety-critical applications, considering proper threat models and practical attacks.

6. Grand Challenge: Robust Private Collaborative Learning

Modern machine learning depends on ever larger data sets. Many of the most interesting data sets are about or touch upon people – their behaviour and properties. Collecting and processing such data sets can be at odds with privacy of the data subjects.

The vision of this grand challenge is to develop learning systems that can use large distributed data sets while guaranteeing data subject privacy. As collecting the data to a central database increases privacy risks, we focus on a distributed setting, similar to private federated learning. In contrast to federated learning where the process is orchestrated by a single entity and the outcome is typically a single model, we focus on collaborative learning, which considers more equal settings where multiple parties may have different goals. Formal guarantees for privacy can be obtained through differential privacy and for robustness through Byzantine robustness.

6.1. Research challenges

6.1.1. Robust collaborative learning with heterogeneous data

Byzantine-robust learning. A severe downside of the current distributed systems is the lack of robustness, in the sense that malicious participants can sabotage the ML system by feeding it wrong data intentionally, known as data poisoning. A strong theoretical model for this situation is given by Byzantine-robust training, which refers to the setting where a fraction of all participants can exhibit arbitrary malicious behaviour (such as providing wrong data and/or not following the training protocol). In this sense, this assumes the strongest possible type of adversary, as there are no bounds assumed as to how much an adversary might alter or “poison” their contributed data.

Byzantine-robust learning can be achieved with relative ease when the data is iid, participants collaborate on the same single learning task, and privacy is not a concern. Challenges start to arise, the more these assumptions are violated, as benign non-iid data can be difficult to distinguish from adversarial data, and many robust aggregation algorithms are incompatible with secure aggregation commonly used in private learning. Developing efficient algorithms for settings where many are violated simultaneously is an important area for future research.

Online private multi-agent learning. The online learning paradigm is a crisp mathematical model within which multi-agent learning with non-i.i.d. data sources can be studied with rigorous performance guarantees.

In multi-agent online learning, the learning agents are nodes which directly communicate only with their neighbours. At each time step, each active agent makes a prediction on the next element of their local data stream, incurs a corresponding loss and observes some feedback information (e.g., the loss gradient). By sharing the feedback with their neighbours, the agents can learn faster than by operating independently.

Homogeneous and heterogeneous learning correspond, respectively, to single and multi-task learning. In single-task, the agents compete against the best global model over the union of the local data streams. In multi-task, each agent competes against the best model over their own local data stream.

The introduction of differential privacy (DP) requirements in multi-agent online learning creates a number of interesting challenges. The fundamental question is to provide sharp characterizations of the trade-off between privacy and utility in online learning, where utility is measured in terms of regret[12]. It will be useful to consider agents with personalised levels of privacy as well as user-level DP in addition to the more common item-level DP. Other important issues that need to be addressed concern the impact caused by the communication constraints and the impact on the memory footprint of the online algorithm caused by the implementation of the DP mechanism.

6.1.2. Privacy, utility and incentives in collaborative and federated learning

An important challenge in collaborative and federated learning is the development of mechanisms that protect the privacy of participants while preserving the utility of the learned models, i.e., the accuracy of their predictions.

Algorithms for differentially private federated learning. Differential privacy (DP) is most often used to guarantee privacy in federated and collaborative learning. Formally analysing the privacy in these cases is more difficult than in the centralised setting. A classic model for DP in a federated context is local DP, where each entity does not trust anyone else and protects its data by adding local noise to its contributions before sharing them with others. This strong model unfortunately leads to poor utility[13]. Intermediates between local DP and trusted curator model (a central party which gathers all raw data) can be built using secure shuffling and secure aggregation. Fully understanding the theoretical properties of these models (including compositionality) and developing algorithms that are as close to the trusted curator model as possible, while adhering to the limitations of these models, such as the finite domain of secure aggregation, are interesting open problems. The notion of metric differential privacy[14], a framework where privacy is defined with respect to an underlying metric, can be useful for achieving a good privacy-utility trade-off in situations where local datasets are highly heterogeneous, as commonly is the case in federated learning.

Privacy without noise. Another interesting approach is to consider methods which do not necessarily add noise to updates (as DP does), but exploit other

[12] Jain, P., Kothari, P., and Thakurta, A. (2012). Differentially private online learning. In Proceedings of the 25th Conference on Learning Theory. Agarwal, N., and Singh, K. (2017). The price of differential privacy for online learning. In Proceedings of the 34th International Conference on Machine Learning.

[13] Jayaraman, B., and Evans, D. (2019). Evaluating differentially private machine learning in practice. In 28th USENIX Security Symposium, USENIX Security 2019 (pp. 1895–1912).

[14] Chatzikokolakis, K., Andrés, M. E., Bordenabe, N. E., and Palamidessi, C. (2013). Broadening the scope of differential privacy using metrics. In Privacy Enhancing Technologies - 13th International Symposium, PETS 2013.

sources of randomness, like the batch sampling procedure or the “mixup” procedure, or rely on different approaches like parameter pruning or quantisation. Pruning and quantization can lead to models with better generalisation and be less vulnerable to attacks, but also reduce the computation/communication requirements and therefore potentially the energy consumption. These alternative techniques could also be combined with more traditional DP approaches, potentially leading to improved privacy–utility trade-offs.

Incentives and private federated learning. On top of privacy and utility trade-offs, it is important to study incentive strategies to motivate the data owners to cooperate by compensating them for their privacy loss in the interest of global utility. This is especially important in a collaborative environment, where not only the model owner but also the intention of the data owner must be considered, as data owners can opt out if they do not agree to the privacy–utility trade-off desired by the model owner.

The goal of the incentive mechanism in federated learning is to find the equilibrium point for both the data owner and model owner to be satisfied. The main challenge in achieving this goal is to develop a robust formal framework for the pricing of data in order to incentivize the participation of data owners by compensating their loss of privacy. Only when we estimate each data owner's contribution accurately, can we provide an appropriate incentive to them. In federated learning, it is more difficult because there are multi-dimensional evaluation criteria such as model accuracy, privacy, fairness, communication, etc. Several techniques, including Shapley value, are studied for this purpose. In addition to that, agents might have personalised privacy levels, which may affect utilities. In this environment, it is natural to study the efficiency of stable equilibria and envy-free equilibria compared to the socially optimal solution (also known as price of stability and price of fairness).

This setting defines a data market that can be viewed as a supply chain where data owners are suppliers and data buyers are retailers. From a game-theoretic viewpoint, one is interested in studying natural solution concepts arising from this market, such as Stackelberg equilibria.

Repeated interactions at a data market can be modelled as a repeated game. The typical measure of performance in these repeated games is the regret, which compares the cumulative utility of the learner’s decisions with that of the single optimal decision on the same sequence of games. Finding regret minimising solutions in various realistic settings is an interesting area of research.

6.1.3. Communication and computation efficiency for scalable learning

A critical challenge in realising this promise of collaborative learning is to develop efficient methods for communicating and coordinating information between distributed devices, in the most communication and computation-efficient way possible. On most distributed systems, the communication of information between devices is vastly more expensive than reading data from main memory and performing local computation. Moreover, the optimal trade-off between communication and computation can vary widely depending on the dataset being processed, the available system resources being used, and the training objective being optimised. While extensive work has gone into communication-efficient training paradigms in recent years, the main challenge that needs to be addressed is to combine such efficient learning algorithms with the additional crucial aspects of privacy and robustness as well as personalised collaborative learning.

6.2. Further research directions

Theory and practice of private synthetic data. Being able to generate anonymised synthetic data that could be freely shared and analysed as real data would solve many privacy problems. Unfortunately, this is difficult to realise in practice: generating meaningful data with strong anonymity guarantees (strong DP) is hard and analysing such data as if they were real can lead to biases[15]. Developing methods that produce higher-fidelity private synthetic data under strong DP would open new opportunities in many applications. The cited work presents the first method that allows consistent downstream analysis from synthetic data. This should be generalised to more general settings. Finally, there is very little theoretical research on DP synthetic data, especially beyond discrete tabular data. More theory could help understand the fundamental limitations of the approach.

Verification of DP properties and anonymity. A popular approach to privacy-preserving learning is described as “algorithms go to data”, indicating that data holders would run analyses on their data on behalf of analyses who would have no direct access to the data. This poses important questions on how to verify that the algorithm is not attempting to steal the data. One way to solve this is to prove that the algorithm is DP with sufficiently strong privacy parameters. Developing methods that allow data users to prove to data holders that the algorithms they are intending to use are indeed DP is an important challenge for future research.

DP beyond tabular data. DP provides privacy protection to individuals when the contribution of each individual can be cleanly separated to define adjacent data sets that differ in the contribution of a single individual. In many real-life applications and data sets this may be difficult: the contributions of different individuals are intertwined, and some contributed elements may appear several times. Developing a generalised DP-like approach to handle such situations more flexibly than current methods would be useful in many applications.

Privacy with using public data and pre-trained models. Research and practice of machine learning is increasingly moving to a “foundation model era” where the solutions build upon large existing models pre-trained on large data sets. Fine-tuning a large pre-trained model on a sensitive data set rather than training from scratch has led to significant improvements in accuracy of various DP computer vision[16] and natural language processing[17] tasks. More generally, this highlights

[15] Räisä, O., Jälkö, J., Kaski, S., and Honkela, A. (2023). Noise-aware statistical inference with differentially private synthetic data. In International Conference on Artificial Intelligence and Statistics (pp. 3620-3643). PMLR.

[16] Luo, Z., Wu, D. J., Adeli, E., and Fei-Fei, L. (2021). Scalable differential privacy with sparse network finetuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021).

[17] Li, X., Tramer, F., Liang, P., and Hashimoto, T. (2022). Large language models can be strong differentially private learners. In The Tenth International Conference on Learning Representations, ICLR 2022.

the utility of using public data or pre-trained models to help DP learning. Developing better methods for this as well as theory to study the best approaches would be very useful. As with foundation models in general, it can be difficult to ensure that the test data have not already been included in the training set. This could be a major problem in privacy, as we might seriously misjudge a privacy-preserving method's capabilities in this setting.

6.3. Use cases: Health, Robotics and Document Intelligence

ELSA considers three use cases as examples of those requiring privacy: health, robotics, and document intelligence.

6.3.1. Health

Health is clearly high among sectors expecting major advances from AI. Broader use of health data holds promise for more effective and efficient treatments. However, health data are highly sensitive and need to be handled with care. New models for managing secondary use of health data, i.e. its use for purposes not related to patient care such as research and development, are being developed. While some countries have implemented their national models, a joint European approach is being developed as the European Health Data Space (EHDS).

The consensus model for secondary use of health data is based on analysing only lightly modified data in secure processing environments, while making sure that any published results are strongly anonymised. This is a compromise between information security and usability: researchers can work with mostly pristine data inside a protected sandbox, but identifiable sensitive data can never leave that sandbox.

Large-scale use of health data often requires combining data from many sources. Moving everything to a single secure processing environment raises risks and can run into regulatory barriers. This has created increasing interest in distributed learning technologies, such as federated learning and swarm learning. These need to be coupled with strong privacy technologies such as differential privacy to ensure privacy of the data.

Getting access to health data for research typically requires a time-consuming application process. New privacy-enhancing technologies such as private synthetic data promise to make the process smoother by allowing easier sharing of anonymous synthetic data. Again, strong privacy technologies such as differential privacy are needed to ensure the privacy of the data. Strong privacy requirements can reduce the accuracy of the synthetic data, highlighting the need for balancing privacy and utility for different tasks and new research to improve the trade-off.

6.3.2. Robotics

In the robotics use case, we consider a setting where autonomous robots operate alongside humans in complex environments. While the robots could improve their behaviour, i.e. their policies, by locally updating their model based on the experiences obtained, the learning would most likely benefit from samples outside the local environment.

Therefore our task becomes to learn the policies for the robots across multiple parties. Not only does this allow to reduce the computational cost for a single party, but also allows the model to be learned with more examples than each of the parties hold.

In settings where the training data contains sensitive information, privacy must be accounted for. This could be the case for example for robots that operate in private homes and use sensors such as cameras to collect the training data. Therefore, we need to make sure that the training data is only handled by parties that we can trust. The federated learning paradigm aims to solve this by keeping the training data on the local sites, and only communicating the model updates to a trusted server that aggregates the updates and sends the aggregate back to the clients.

However, while this approach solves the clear privacy threat of directly leaking the sensitive training data, we also need to make sure that the learned policies do not leak any sensitive information learned from the sensitive features. This could be achieved by adding differential privacy as an additional layer of security to the FL framework.

Besides the privacy concerns, the robots need to also operate safely. This is of utmost importance when the robots operate alongside humans. However when these robots are learned collaboratively there is a risk that a malicious party could try to affect the learned policies by poisoning the training data with adversarial examples. To guarantee that the learned policies are safe, we need to make sure that the learning procedures used are robust to adversarial examples or any other type of a malicious attack.

The most common privacy-preserving learning methods, aimed at addressing the first goal, are based on adding noise to the learning process in order to mask out any individual's contribution. This added noise can possibly degrade the accuracy of the learned policies, and thus can also cause concerns with regards to safety. Therefore in order to guarantee both safety and privacy, we cannot simply use the robust and private federated learning tools independently, but we need to develop methods to combine the two.

6.3.3. Document intelligence

The last decades have seen an increased digitisation of documents for practical and environmental reasons. However, the manual management of these digitised documents is becoming tedious and therefore requires an automation of the process.

Fortunately, many approaches and tools have been proposed to automatically process document images, collectively referred to as Document Intelligence. Document intelligence is the research area at the intersection of computer vision and natural language processing, focusing on techniques and methods for extracting, interpreting and inferring information from documents. It is a rapidly growing field of research with applications in many sectors where the processing of large volumes of documents is essential, such as finance, insurance, public administration, business or personal document management. Document intelligence includes several subfields such as optical character recognition (OCR), information extraction and document visual question answering (DocVQA). DocVQA focuses specifically on the task of answering questions about the content of a document image, which can involve high-level analysis and reasoning about both text and visual information.

Therefore, more and more companies want to automate their document-based processes. Current practices focus on training generic DocVQA models that can then be applied on various downstream document understanding tasks.

However, training an accurate machine learning model requires an amount of data that a single company may not have.

One possible solution to deal with this problem is to train this model collaboratively by aggregating and centralising data from all companies. Unfortunately, documents are very sensitive and can leak valuable information. For example, some of the information available in these documents could be directly related to the company's business and could therefore lead competitors to want to infer this type of information.

Another solution is to consider federated learning. However, even though federated learning is more private than the centralised approach, many attacks have shown that a lot of information can still be inferred from the updates/models shared between the clients and the server.

The privacy of federated learning can be guaranteed by additional privacy-enhancing technologies such as differential privacy. Differential privacy is based on requiring the results of a computation to remain almost unchanged when the data for one user is changed. The document intelligence application highlights an interesting challenge for practical application of differential privacy: how to define one user. Invariance to changing a single document is likely too weak while changing one huge provider might make learning very difficult, suggesting a need for new formulations that are better aligned with application requirements.

6.4. Benchmarks and metrics

Evaluating privacy-preserving machine learning is much more difficult than general machine learning challenges, because of the inherent tradeoff between privacy and utility. For any approach, it is possible to build a curve representing different tradeoffs.

For easy evaluation, we need to somehow obtain comparable results from this curve for the different methods. This could be done either by fixing the utility and evaluating privacy at that point, or fixing privacy and evaluating utility at that point. From the privacy perspective, fixing privacy seems more reasonable as legal privacy requirements can be seen as a hard constraint.

After selecting the operating point, the next subsections discuss challenges in concrete evaluation of privacy of a model.

6.4.1.Evaluating formal differential privacy guarantees

DP is a mathematical property over all adjacent data sets, usually differing by the contributions of a single individual. Because the number of possible adjacent data set pairs is usually very large, it cannot be verified empirically but with a mathematical proof. Any DP algorithm therefore has to come with such a proof and verifying this proof is the primary method of verifying that an algorithm satisfies DP.

A pen-and-paper proof of DP still leaves open the translation of the abstract algorithm analysed into computer code. Possible problems can range from simple bugs or overlooked properties of the computing environment, into problems stemming from different abstractions (e.g. real numbers vs. floating point numbers). There is some very recent work in so-called privacy auditing that attempts to construct examples to disprove the privacy claims that can be used to catch some such bugs. These can help create confidence that the implementation is correct, although they cannot formally prove it.

6.4.2. Evaluating empirical guarantees with attacks

Formal guarantees are important, but they only give a theoretical upper bound on the vulnerability of a particular method. For a complete understanding, it is important to complement these with empirical lower bounds on the vulnerability obtained from attacks, such as membership inference attack and reconstruction attack.

A widely used scenario for evaluating vulnerability to attacks is a blue team - red team competition, where blue teams submit methods that protect the privacy, while red teams seek to attack and break the protections. The strongest attacks against a federated learning system would operate online based on messages exchanged during learning, possibly even as an active participant of the learning process, but this is difficult to organise in practice due to the amount of data often exchanged. A much easier alternative is to deploy attacks such as membership inference against the final model, either in white box mode (adversary has full access to the internals of the model) or in black box mode (adversary only has access via some restricted API).

Grand Challenge: Main Summary and Value Proposition

Research Challenges

- Robust collaborative learning with heterogeneous data
- Privacy, utility and incentives in collaborative and federated learning
- Communication and computation efficiency for scalable learning

Further Challenges

- Privacy with pre-trained foundation models and public data
- Extending formal privacy to more complex and realistic settings

Use Cases

- **Health**: secure access to distributed data; private synthetic data
- **Robotics**: privacy-preserving learning from deployed devices
- **Document Intelligence**: multimodal private federated learning

Value Proposition: ELSA combines unique expertise on differentially private and federated learning as well as private synthetic data, from core theory to practical applications.

7. Grand Challenge: Human Agency and Oversight

7.1. Research challenges

Research challenges of the human agency grand challenge are multifaceted and include technical questions, as well as social impact, including ethical, legal and regulatory challenges.

The primary focus of the technical research of this grand challenge is concerned with unresolved challenges surrounding the capacity of humans to understand, interpret and comprehend the underlying logic of an output produced by an ML model, particularly those which utilise deep learning. If advances in ML and AI are to deliver the promised benefits of enabling and enhancing human well-being and for the benefit of the many and not merely the few, then they will need to operate in real-world settings in conjunction with individuals, organisations and communities of all shapes, sizes, and capabilities, including society's most vulnerable. It precipitates the need for addressing multiple ethical, legal and regulatory challenges described below.

7.1.1. Interpretability

The question of interpretability is multifaceted and can take different forms. It includes explaining existing, non-transparent models, including foundational models such as vision transformers; it also includes creating new, interpretable-by-design models, which are created with a specific purpose of being interpretable.

Within ELSA, we leverage the collaboration between the technical partners to produce new methods aiming for interpretability, as well as looking into the overlap between this grand challenge and the two other ones outlined above, to identify the trade-offs between privacy, transparency and robustness.

7.1.2. Disentangled learning

More relaxed than interpretability is the notion of disentangled learning. This refers to the way in which we want to factorise our decision making into a set of disentangled decisions, each having a desired meaning for the humans. For example, we may want to program a robot in a way that we can disentangle its actions into a combination of straight walking and jumping, or generate plausible images of a human face which contains certain features: hair length, facial hair, smile, frowning, amongst others. While this does not go as far as interpretability, and the model may not be transparent in its working, it still allows agency for a human to define the desirable behaviour of the model by shaping up the features to their liking.

Within ELSA, we consider the possibilities of disentangled learning both in conjunction with interpretability and as a tool to identify interpretability.

7.1.3. Datastreams and adapting to shifts in the data pattern

Many existing ML methods are designed to operate in scenarios where the training data is fixed; a change in the data set, consequently, means that the whole system needs to be optimised from scratch. The concept of lifelong learning challenges this setting: the data could be provided through an ever changing data stream, and the system is elastic enough to adapt to this data, ideally in real time. These data streams are masterminded by humans, external to the training system, and therefore can also be viewed as a type of human-in-the-loop ML.

In ELSA, we consider lifelong learning to be an important aspect of human-in-the-loop ML, and see it as a complementary problem to the problem of transparency. Interpretable-by-design methods including prototype-based machine learning can be advantageous in both tasks and help improve the performance.

7.1.4. Federated and multiagent scenarios

In many cases, the problem is complicated by models working in a complex system, involving distributed access to the data. It may take different forms, including sharing parts of the input data, producing part of the decision making, or working within a multiagent system. Such a problem statement means taking into account the problems of security and differential privacy which are an integral part of solving the challenge of federated learning, linking this grand challenge with the two previous ones.

In ELSA, we are exploring the trade-offs between, on one hand, federated and multiagent scenarios' safety and security, and on the other hand, their transparency.

7.1.5. Meaningful human oversight

In the development and deployment of AI, it is vital that these machines remain always subject to meaningful human control, thereby ensuring that they remain our servants and not our masters.

There are a number of properties attributable to AI technologies that make ensuring meaningful human control a serious and hitherto unresolved challenge. These include the technical properties of many of these technologies (for example, their automaticity, lack of functional interpretability, opacity and stochastic properties (for some forms of ML), speed, scale, dynamism, lack of assurance concerning the provenance, integrity, legality and quality of the underlying data and so forth) but is also attributable in no small measure to the complexity of the socio-technical systems in which they are embedded, which is exacerbated by the multiplicity of actors, organisations and components which contribute to the supply chain through which real-world technologies are brought into being. Accordingly, there remain acute challenges associated with establishing and maintaining meaningful human oversight over the development and operation of these systems.

ELSA contributions are grounded in existing and proposed regulatory governance frameworks that apply to data-informed services, as a means for generating insight on the adequacy of existing frameworks and the proposed approach of new legal frameworks (including the EU's proposed AI Act). By engaging with regulators, stakeholders and drawing on and integrating insight from the voluminous literature on regulatory governance and critical algorithm studies, while engaging in primary research to understand how key stakeholders understand the risks and opportunities associated with existing and proposed legal governance frameworks for data-informed services in Europe, ELSA can provide recommendations to European policy-makers and the broader stakeholder community.

7.1.6. The rule of law

AI technologies may be deployed in ways that directly threaten the integrity of democracy, respect for human rights and the rule of law.

The recent release of large language models (LLMs) such as ChatGPT and others can be readily exploited by malicious actors in ways that seriously threaten epistemic trust, and thereby the foundations for peaceful social cooperation among strangers upon which civilisation depends. This is one of the motivating concerns underpinning the on-going efforts at international level by national representatives, convened under the auspices of the Council of Europe, to establish an international convention which is underpinned by the express purpose of establishing 'certain fundamental principles, rules and rights aimed at ensuring that design, development and application of AI systems is fully consistent with respect for human rights, the functioning of democracy and the observance of rule of law' [Council of Europe, proposed International Convention on AI, Human Rights, Democracy and the Rule of Law 2023, Article 1[18]]. Yet the capacity for socio-technical systems in which AI technologies are embedded to operate remotely in a highly opaque manner, in real-time and at scale, makes the ability to ensure that these systems operate are brought under the rule of law, and are not employed in ways that seek to manipulate, exploit or otherwise interfere with the rights and freedoms of natural persons remains an open challenge.

ELSA partners are involved in the creation of legal and regulatory frameworks to contribute towards addressing the challenge of human agency. This includes, for example, involvement of ELSA participants in the first IEEE standard on Explainable AI [XAI, P2976[19]] which holds a working group and is a process that takes years bringing together academia, industry and other interested stakeholders. However, whether these mechanisms are effective vehicles for securing safe and secure AI including meaningful human oversight is unknown and unproven. Not only is the effectiveness of these techniques uncertain, but various commentators have drawn attention to the 'private' nature of the underlying standards and assurance processes, which appear contrary to basic principles of democratic accountability, transparency and participation.

7.2. Further research directions

Interpretability of highly-parameterised foundational models remains a big challenge. It is not only so because of the opaque nature of these models which can be (at least to a certain extent) addressed using the standard explanation models, but also due to the problems such as opaque data collection. In many cases, such as for GPT-4 or Vision Transformers, for both text and image models, the data is neither public nor is it documented; furthermore, it is not clear how those data impact the decision making.

Attribution and erasure of the trained data are also important aspects of this problem: when some of the data in an already trained model are shown to be

[18] https://ai-regulation.com/council-of-europe-draft-convention-on-ai-human-rights-democracy-and-rule-of-law/#:~:text=The%20Draft%20%5BFramework%5D%20Convention%20on,by%20November%2015th%2C%202023.

[19] https://sagroups.ieee.org/2976/

problematic (e.g. due to the ethical or legal reasons), there needs to be a way to erase those, a problem studied as machine unlearning.

Quantification and characterisation of model transparency presents another formidable research challenge. It is possible to address this problem through different viewpoints such as quantification of disentanglement, or through qualitative analysis of the transparency, in a way similar to the characterisation of the risk in the AI act. However, these measures do not encompass all possible scenarios of characterising transparency, and new ways need to be found.

7.3. Use cases: Robotics and Multimedia

7.3.1. Robotics

Robots that autonomously carry out complex tasks have a great potential to solve major societal challenges, e.g. enabling sustainable food production, helping in disaster situations, or assisting people with limitations. In many of these applications, robots will have to operate with and around humans to solve desired tasks. These settings, however, put additional design requirements on the robots' operation with humans-in-the-loop, e.g. strong safety requirements or interpretability of the robots' decisions. Enforcing such design requirements is particularly difficult for autonomous robots that employ modern ML techniques for decision making. On one hand, these robots can learn highly complex behaviours and adapt to new tasks through interaction with their environment. On the other hand, many ML models are practically black boxes and often do not allow robot developers to use classical verification techniques to ensure desired properties of the robot during the robot operation.

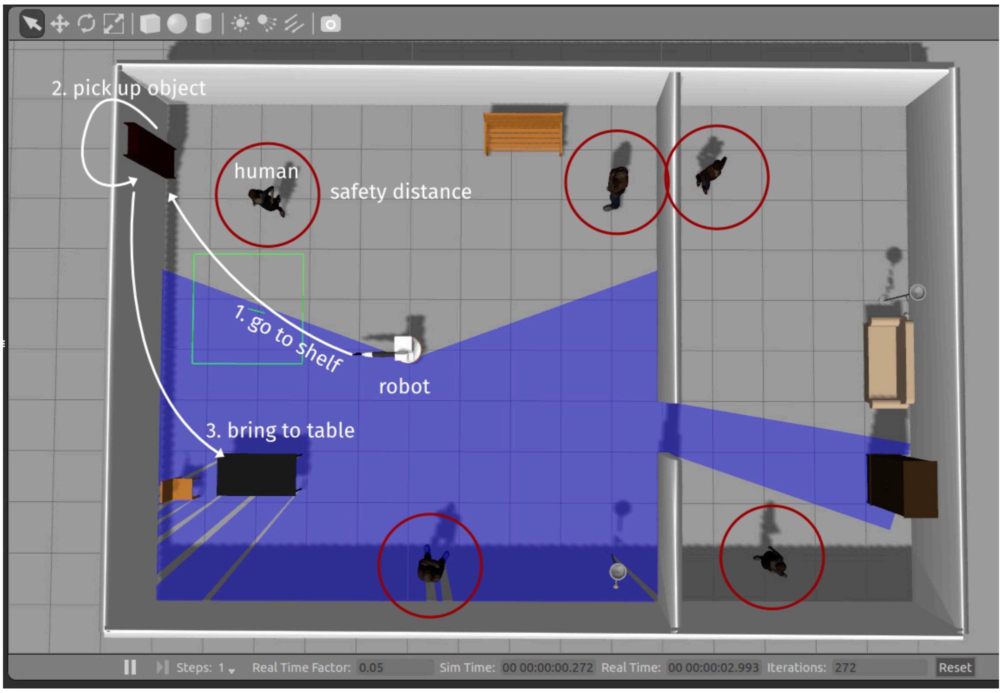


Figure 2: Motivation of necessity for autonomous robots to operate safely and as intended around humans.

It is of utmost importance that autonomous robots operate safely and as intended around humans. For instance, when robots perform kitchen tasks in a home-like environment (see Figure 2), the policy of a robot must always maintain a sufficient safety distance (see red circles in Figure 2) and low velocities around humans. To learn such robot policies, robots need to anticipate the human's motions during learning to not cause undesired situations. Besides safe operation, autonomous robots must further incorporate and obey human instructions, such as task goals and how the task should be completed. To develop learning-based autonomous robots that adhere to human instructions, new research advances are required in robot learning with humans-in-the-loop. Users of the robots should be able to instruct the robot about the intended task to perform, e.g. "get the object from the shelf and put it on the table", and how the robot should perform the task, e.g. ''place the object on the side of the table where the user normally sits". This task specification needs to be understood by the robot to learn the desired behaviours. To oversee the learning process, we need to make the robot's learned behaviour interpretable for humans. The robot should explain to humans what it has learned so far and how it will perform the task. For instance, the robot in Figure 2 may signal to the user that it first goes to the shelf in the corner of the room, then picks up the object, and finally places it on the table next to the chair (see Figure 2). The advances in addressing this grand challenge will help to create interpretable models of what the robot has learned and how it will perform a given task.

By taking into account human oversight, one can enable new generations of autonomous robots that can safely execute tasks with and around humans. The learned policies will minimise residual risks and help ensure that robots operate in ways that ensure that they do not unduly threaten human health and safety. With the help of new interpretability and explainability techniques, these robots will further be able to explain what they have learned and how they will execute given tasks, always giving users the oversight of the robot. In this way, autonomous robots will operate as intended by users and may, in turn, contribute towards addressing societal needs for physical assistance in real world environments and offer the potential to enhance economic productivity, health and safety of human workers.

7.3.2. Multimedia

The ability to generate highly realistic images and videos using generative deep learning models has created new challenges in the domain of multimedia. In particular, it has become increasingly difficult to distinguish between real and fake visual content, which has important implications for applications such as content moderation, forensics, and journalism. This task is particularly challenging because deep learning models are highly effective at generating images that are visually indistinguishable from real images. At the same time, these models can have distinct statistical properties that can be exploited to differentiate them from real images.

One of the big challenges is detection of deep fakes generated using deep learning models. For that purpose, it is possible to generate a large dataset of deep fakes, which will be used to train and evaluate algorithms designed to detect and classify fake images.

Another aspect of the use case is developing baseline algorithms for detecting deep fakes and manipulated media. These models will be used to evaluate the performance of new approaches developed by the research community, and will serve as a benchmark for measuring progress in this area. Metrics such as precision, recall, and F1 score will be used to evaluate their performance, with higher scores indicating better performance. Other metrics used to evaluate deep fake detection models include the area under the receiver operating characteristic curve (AUC-ROC), which measures the model's ability to distinguish between real and fake images as its discrimination threshold is varied.

A potential benchmark set-up involves comparing the deep fake detection model's performance metrics against those collected by human classification. By comparing the accuracy, precision, recall, and other relevant metrics of the model with those of human classification, it is possible to get insights into the strengths and weaknesses of each approach. This information can then be used to refine the model and improve its performance, as well as to develop a more effective combination of human and algorithmic approaches to combating the threat of deep fakes.

7.4. Benchmarks

7.4.1. Benchmark Metrics

As interpretability is defined in the terms of (human) understanding, it may not be possible to provide universal metrics which do not explicitly take humans-in-the-loop into account.

Instead, a number of options are available to monitor such progress, which include:

- quantification of representation learning transparency; and
- developing qualitative monitoring tools.

Quantification of representation learning transparency concerns the problem of understanding which aspects of the model contribute towards the decision making. For this purpose, a number of approaches have been proposed including measures of disentanglement [Do and Tran, 2023[20]] [Sepliarskaia et al, 2019[21]]; metrics based on attention-based post-hoc explanations [Bibal et al, 2022[22]]. However, this does not solve the problems of quantification of transparency as current explanations have limitations. For example, attention-based models may merely register correlations between explanations and the inputs [Wiegreffe and Pinter,

[20] Do, Kien, and Truyen Tran. "Theory and Evaluation Metrics for Learning Disentangled Representations." International Conference on Learning Representations, 2023.

[21] Sepliarskaia, Anna, Julia Kiseleva, and Maarten de Rijke. "How to not measure disentanglement." arXiv preprint arXiv:1910.05587 (2019).

[22] Bibal, Adrien, Cardon, Rémi, Alfter, David, et al. Is attention explanation? an introduction to the debate. In : Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022. p. 3889-3900.

2020[23]] and it is well known that correlation is not (necessarily) a causation. Therefore, the methodology of evaluation is an open theoretical and practical question.

Qualitative characteristics can be centred around such aspects as number and nature of model parameters; possibility for and the nature of visual and linguistic interpretability.

To produce meaningful oversight over decision making, it is necessary to:

- report quantitative and qualitative metrics;
- develop new techniques to derive quantitative metrics for continuous assessment of interpretability; and
- engage with the use cases to include the scenarios of transparent decision making into the practical challenges.

From the perspective of use cases, there is a need for customised metrics which reflects their needs. The autonomous driving use case, for example, is centred around robustness and reliability of the models, and therefore, the question of transparency should be solved jointly with it. The robotics use case is centred around the notion of safety, with some of the metrics for safety evaluation also being capable of shedding light on the intrinsically linked question of transparency. The multimedia use case concerns distinguishing generated data, and therefore, the focus would be on the clues which would help give away the salient features of the particular algorithm, generating these data.

7.4.2. Addressing the challenge of integrated governance to secure meaningful human oversight

To address the challenge of integrated governance to secure meaningful human oversight, the ELSA team will undertake primary and secondary research to critically interrogate potential mechanisms through which the ‘quality‘ of data-driven services might be secured, and produce analytical materials, peer-reviewed and policy-oriented content to disseminate and draw attention to our findings and recommendations.

The goal of our research will be to not only study how to set up effective governance structures for algorithmic decision-making systems but also to understand the needs, perceptions, and hopes of policy makers, who may govern such systems. We hope our interaction with policy makers would usher in a new series of technical methods that help provide decision-makers with control over system outcomes.

[23] Wiegreffe, Sarah et Pinter, Yuval. Attention is not not explanation. In : 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019. Association for Computational Linguistics, 2020. p. 11-20.

Grand Challenge: Main Summary and Value Proposition

Research Challenges

- Transparency and explainability in human-readable terms
- Legal and ethical challenges of safe and secure AI
- Governance architectures ensuring meaningful human oversight

Further Challenges

- Interpretability of highly-parameterised foundational models
- Attribution and erasure of the trained data
- Quantification and characterisation of model transparency

Use Cases

- **Robotics**: interpretable models for safe robots
- **Multimedia**: detection of deepfakes

Value Proposition: ELSA combines unique expertise in human-in-the-loop decision making, encompassing technical aspects such as interpretability and core ML, as well as ethical, legal and regulatory knowledge.

8. Case Study: Large Language Models, ChatBots, Intelligent Assistants

Challenges to Secure and Safe AI by LLMs

Development and expansion of new deep learning techniques has made it possible to solve many decision related problems in ways unimaginable just a few years ago. However, such a fast move from fundamental and applied research into commercial products and government services has created a range of problems which can be broadly attributed to interaction between technology and society.

We suggest that concrete examples are given of these new problems which involve:

- The need for alignment with ethical principles (including human rights).
- The need to conform with legal duties and obligations, and potentially risk of undermining the rule of law and democracy.
- The need for the system to be robust in a technical or organisational sense.
- The need for a human (in the loop) to understand and evaluate the machine output/answer.

These problems are, in many ways, intertwined: in the recent works, one can see that the computational aspects of the state-of-the-art models are intertwined with the legal and ethical challenges. This can be illustrated by the following example. A large language model (LLM) must not, in accordance with the GDPR, disseminate private data. But if human's understanding of such LLM is poor enough that one cannot formally prove that the system would not output these personal data, in full or partially, in any possible scenarios, this may lead to falling short of GDPR requirements. This hypothesis is supported by formal complaints on ChatGPT LLM to the French personal data regulator, CNIL, which claim the lack of data protection. Similar complaints to the Italian regulators resulted in a temporary ban of the ChatGPT software[24].

LLMs and other big models (such as image generators) include proprietary data as well as complex black-box architectures. The problems related to the proprietary data, in this context, include the fact that it is not publicly known which specific data set the model is trained on. This creates the possibility of non-ethical data collection and leads to a lack of reproducibility. The technical tasks to address

[24] https://www.bbc.com/news/technology-65139406

these issues may include the removal of impact of undesirable, non-ethically collected data (e.g. personal data, intellectual property) from the model.

The black box nature of many of the existing AI models, especially based on deep learning, necessitates the development of the methods and tools for *transparent* machine learning. At the heart of the need for making LLMs transparent lie multiple legal and ethical issues that require input from multiple domain-experts and stakeholders. Such questions include those around intellectual property permeating from the training data to model outputs, eroding consumer trust, and how to ensure LLMs can be verified for factual correctness, data provenance and adherence to ethical principles.

In some sense, the role of how generative tools affect human agency is at the heart of the current LLM revolution. How does an agent like ChatGPT affect decisions that individuals make in the real world? We expect a surge in work, led by consortia like our own, to not only study the effects of how LLMs empower decision-makers but also understand how these powerful systems are governed on a regulatory, context-specific basis. The ethics of deploying such technologies requires a precise characterization of their shortcomings from interpretability to robustness (or lack thereof). Many countries, including those in the European Union, will need to decide how to govern augmented decisions, where humans observe AI system suggestions in their decisions.

Technical perspective of Security and Safety of LLMs

The disruptive progress of Large Language Models (LLMs). LLMs have recently shown surprising capabilities. This technology bears great potential and is of a disruptive nature to the economy as well as society as a whole. The use of this technology should be critically reflected and used/deployed/integrated in compliance with legal boundaries and our societal values.

Deployment and Application-Integrated LLMs. We are currently seeing a rapid deployment at scale of this technology. Given severe security and safety concerns such as lack of distinction between trusted and untrusted inputs – even openly acknowledged by the involved companies[25] [26] – the consequences are difficult to foresee. From our scientific perspective and our information/experience with the technology so far, the users are at risk and compliance is unclear to say the least. Nevertheless, deployment has already happened with integration in Bing and an announced deployment in Microsoft Office. Companies foresee a trillion dollar business with millions of users already today. This is somewhat in contrast to the exploratory approach portrayed by OpenAI. Integration with plugins will lead to

[25] GPT-4 System Card. OpenAI 2023. https://cdn.openai.com/papers/gpt-4-system-card.pdf

[26] Introducing Google's Secure AI Framework. Google 2023. https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/

even more capable, compositional systems, for which assessing and ensuring their trustworthiness is an open challenge.

Task-open. LLMs are advertised and have shown strong capabilities at a range of tasks. In particular, the task is given by the user as a prompt. Hence the functionality is modulated at runtime. This gives a wide scope of potential application scenarios.

General Compute Platform. The current development point in the direction of LLM-type models becoming a general compute platform which is not task constraint. This raises severe concerns about cybersecurity and safety as described below. There is a larger debate about AGI, AI alignment, and other societal risks, which also deserve attention. Here, we focus on the imminent risks of the current technology.

Data vs Instructions - Trusted vs Untrusted Sources. We like to highlight one very central technical challenge. The current models are trained to be "instruction following". As a consequence the current generation of technology is not able at its core to distinguish between data and instructions as well as information from trusted and untrusted sources. Decades of research in cybersecurity have identified these ingredients as root causes of unsecure systems. Consequently, we are seeing the current generation of LLMs being an insecure and unsafe platform. While the efforts from OpenAI show an awareness of many of these risks, the fact that the already deployed models are not resilient and still vulnerable to jailbreaks, re-programming and other prompt injection attacks is further evidence that no principled solution is available right now. We outlined some of the core cybersecurity risks in a technical report that we also brought to the attention of the involved companies[27].

Cybersecurity Perspective of LLMs

While the elaborations above have already pointed out key issues with this technology in its current form, we further elaborate on the potential cybersecurity risks that are associated in particular with application-integrated LLMs and the outlined issues with intermingling instructing-following models with untrusted data. This emphasised the point of bringing together lessons learnt from cybersecurity in terms of threat modelling, separating data from instructions, and recognizing untrusted input, to the latest developments in AI/ML.

This analysis led to a range of Prompt injection (PI) attacks that pose a significant threat to the security of LLMs. While PI attacks have been primarily limited to individuals attacking their own LLM instances (or a public model such as ChatGPT),

[27] Sahar Abdelnabi*, Kai Greshake*, Shailesh Mishra, Christoph Endres, Thorsten Holz, Mario Fritz. "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection." *The 16th ACM Workshop on Artificial Intelligence and Security Workshop* (2023).

integrating LLMs with other applications might make them susceptible to untrusted data ingestion where malicious prompts have been placed. We call this new threat indirect prompt injections and demonstrate how such injections could be used to deliver targeted payloads. This technique might allow attackers to gain control of LLMs by crossing crucial security boundaries with a single search query.

Recent LLMs may behave like computers executing programs[28]. Thus, one can draw insights from the classical computer security domain to design a new set of attack techniques. A high-level overview of the threat model, covering the possible injection delivery methods, the different threats, and the possible affected individuals or systems, can be found in the paper[29]:

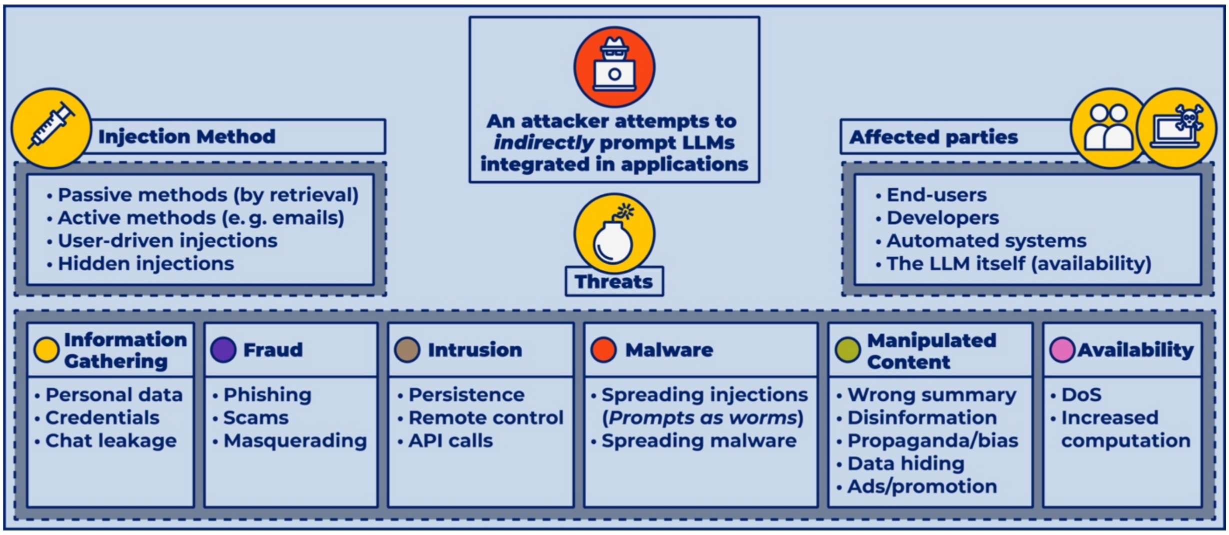


Figure 3 - Application integrated LLM threat landscape.

Privacy of LLMs

LLMs can ingest personal data both as part of the training data and in the prompt. These raise different privacy concerns.

LLMs have been shown to be highly prone to memorising chunks of their training data verbatim, and reproducing that in response to a suitable prompt. As a first approximation, LLMs should therefore be assumed to memorise all of their training data. This memorisation could in principle be avoided by training with differential privacy (DP), but pre-training very large models with DP with acceptable accuracy has so far proven impossible. On the other hand, it is quite possible to

[28] Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. "Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks." arXiv (2023).

[29] Sahar Abdelnabi*, Kai Greshake*, Shailesh Mishra, Christoph Endres, Thorsten Holz, Mario Fritz. "Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection." The 16th ACM Workshop on Artificial Intelligence and Security Workshop (2023).

perform private fine-tuning with a smaller collection of sensitive data with strong DP guarantees.

Sensitive data in prompts poses other, even less studied problems. First and foremost, the provider of a hosted LLM can observe everything submitted to the LLM. There have already been cases of sensitive company documents being potentially leaked after an employee fed them to ChatGPT. First work addressing more complicated scenarios, such as distilling sensitive data used as a prompt into an anonymous equivalent prompt has recently started appearing as pre-prints.

9. Mitigations

GPT-4 was trained with intervention to reduce jailbreaks, such as safety-relevant reinforcement learning with human feedback (RLHF)—our work[30] and several other jailbreak attacks[31] [32] show that it is possible to adversarially prompt the model even in real-world applications. While some jailbreaks are later fixed, the defensive approach seems to follow a “Whack-A-Mole” style. The extent of how RLHF can mitigate attacks is still unclear. Some recent theoretical work[33] shows the impossibility of defending against all undesired behaviours by alignment or RLHF. Empirical evidence of inverse scaling in RLHF models was also reported[34]. Nevertheless, understanding the practical dynamics between attacks and defences and their feasibility and implications (ideally in a less obscured setting) are still open questions.

Besides RLHF, deployed real-world applications can be equipped with additional defences; since they are typically undisclosed, we could not integrate them into our synthetic applications. However, our attacks succeed on Bing Chat, which seems to employ additional filtering on the input-output channels without considering the model’s external input. Even if applied, it remains unclear whether filtering can be evaded by stronger forms of obfuscation or encoding[35], which future models might further enable.

Other potential defences might include processing the retrieved inputs to filter out instructions. However, this might create another dilemma. On the one hand, to prevent the rescuer from falling into the same trap, we might need to use a less

[30] Sahar Abdelnabi*, Kai Greshake*, Shailesh Mishra, Christoph Endres, Thorsten Holz, Mario Fritz. "Not what you’ve signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection." The 16th ACM Workshop on Artificial Intelligence and Security Workshop (2023).

[31] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, Yang Liu. "Jailbreaking chatgpt via prompt engineering: An empirical study." arXiv (2023).

[32] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. "Jailbroken: How Does LLM Safety Training Fail?." arXiv (2023).

[33] Yoav Levine Amnon Shashua Yotam Wolf, Noam Wies. 2023. "Fundamental Limitations of Alignment in Large Language Models." arXiv (2023).

[34] Perez et al., "Discovering Language Model Behaviors with Model-Written Evaluations." arXiv (2022).

[35] Daniel Kang, Xuechen Li, Ion Stoica, Carlos Guestrin, Matei Zaharia, and Tatsunori Hashimoto. "Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks." arXiv (2023).

general model that was not trained with instruction tuning. On the other hand, this less capable model might not detect complex encoded input. For example, we show that by using an encoding (e.g. Base64), we needed to explicitly provide instructions for the model to decode the prompt. However, future models might perform such decoding automatically, e.g., when using self-encoded prompts[36] to compress the input and save the context window.

Another solution might be to use an LLM supervisor or moderator that, without digesting the input, specifically detects the attacks beyond the mere filtering of clearly harmful outputs. This might help to detect some attacks whose purpose does not depend on the retrieved sources (e.g., some scams) but might fail to detect disinformation and other manipulation attacks. Verifying against retrieved sources will induce a similar dilemma to the one explained above. A final promising solution is to rely on interpretability-based solutions that perform outlier detection of prediction trajectories[37]. Unfortunately, it is currently hard to imagine a foolproof solution for the adversarial prompting vulnerability, and the efficacy and robustness of these defences against obfuscation and evasion still need to be thoroughly investigated in future work.

[36] Noah Goodman Jesse Mu, Xiang Lisa Li. "Learning to Compress Prompts with Gist Tokens." arXiv (2023)

[37] Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt."Eliciting Latent Predictions from Transformers with the Tuned Lens." arXiv (2023).

10. Looking ahead

Europe has understood the strategic role of AI technology and has made strategic investments into the whole ecosystem. While there is immense momentum and potential, a joint, major effort is needed in order to achieve a balance of power on an international level that lays the foundations for the decades to come.

AI technology is becoming an integral part of IT systems and will drive innovation and the economy at large in many - if not all sectors. Such a key technology will have a significant effect on the prosperity of the EU and is interconnected with aspects of digital sovereignty.

While we are in a decade of disruptive breakthroughs in AI technologies and systems, we equally see challenges and short-comings becoming more evident and apparent. The rapid transition into practice and deployment to millions of users, puts a magnifying glass on open challenges of safety, security as well as trustworthiness of these systems in general.

Basic, foundational research is the source of innovation chains that drive the AI ecosystem in compliance with our expectation on these shared societal values. Fundamental grand challenges demand fundamental solutions that will stand the test of time. Europe is traditionally strong in this regard. ELLIS has established an excellence driven, European AI network and ELSA is building on its momentum to spearhead the developments on safe and secure AI.

Strong support of independent and open academic research is vital to balance huge industrial investments to make sure the technology develops in a way that benefits all, not just narrow industrial interests. This would support open source, transparent data provenance and models motivated by the issues of robustness, security and safety.

With all its challenges and complications, the EU has pushed for a vision on how to regulate AI – a contribution that could not be more timely. It should be understood as an opportunity to unite behind common goals also providing certainty and direction, rather than hampering innovation in this space.

ELSA is connecting European experts on safe and secure AI, who serve as a think tank on this key, strategic topic. In this capacity, a closer dialogue with policy makers and industry is invited in order to set Europe on the right track for the coming decades.

As technology is developing and deploying at such a rapid pace, there is a substantial risk to cause a disconnect and divide between society and technology. Therefore, a public debate should be fostered in order to have an informed discussion, informed opinions, and informed decisions that ultimately empower people. In particular, when safety, security and trustworthiness is concerned, technology can only be part of the solution and solutions need to be human-centric.

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 to shape this technology in a European understanding.

11. Credits

EDITORS:

- Plamen Angelov, Lancaster University
- Battista Biggio, University of Cagliari
- Mario Fritz, CISPA Helmholtz Center for Information Security
- Antti Honkela, University of Helsinki
- Dimosthenis Karatzas, Computer Vision Center

CONTRIBUTORS:

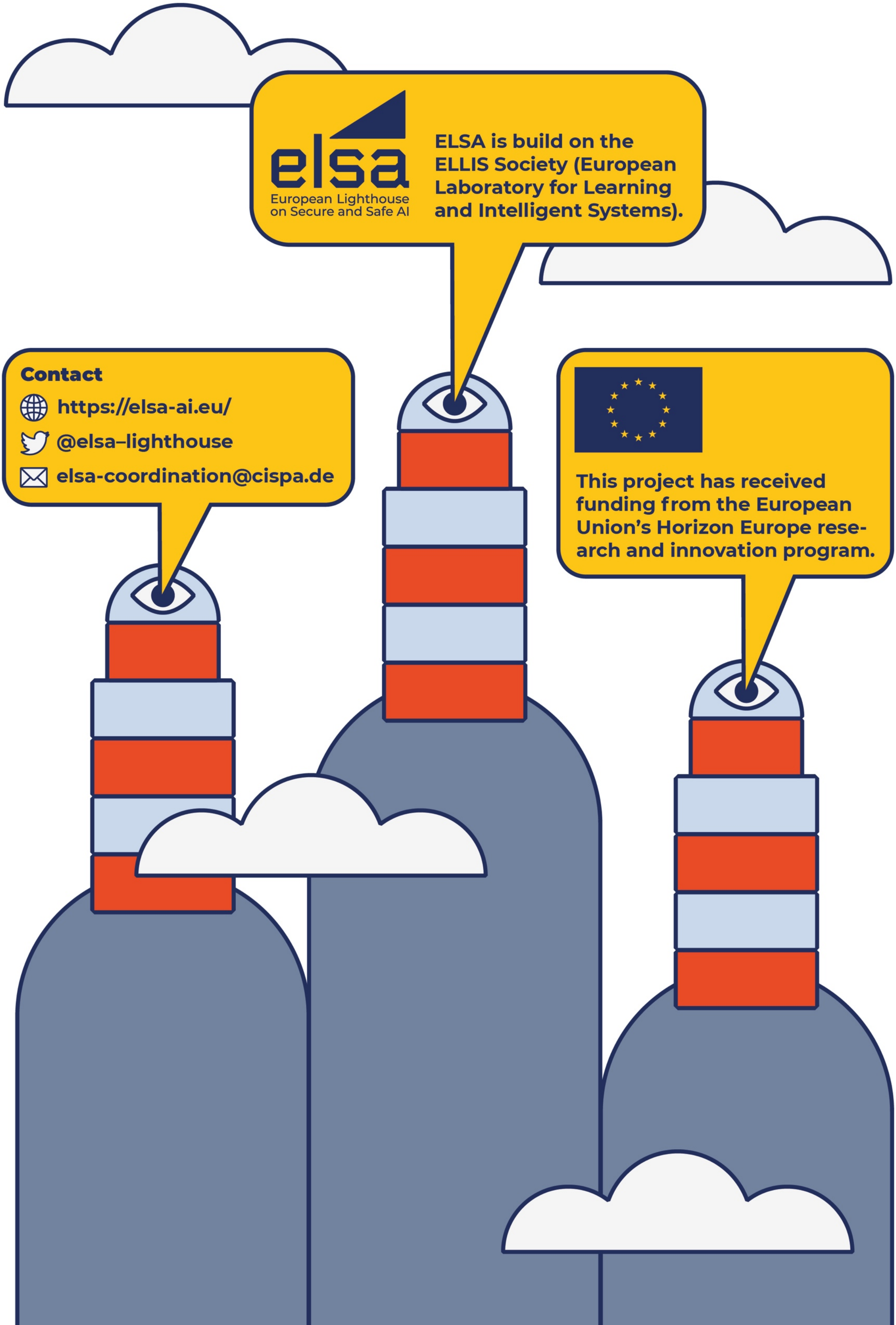
- Plamen Angelov, Lancaster University
- Battista Biggio, University of Calgliari
- Mario Fritz, CISPA Helmholtz Center for Information Security
- Antti Honkela, University of Helsinki
- Dimosthenis Karatzas, Computer Vision Center
- Dmitry Kangin, Lancaster University
- Mark Niklas Müller, ETH Zürich
- Ambra Demontis, University of Cagliari
- Maura Pintor, University of Cagliari
- Angelo Sotgiu, University of Cagliari
- Josep Lladós Canet, Computer Vision Center
- Thorsten Holz, CISPA Helmholtz Center for Information Security

This SRA is built on the following 3 deliverables of the ELSA network:

- **D1.1 Grand Challenge Report: Technical Robustness and Safety Grand Challenge and Benchmarking Metrics**
  Lead authors: Ambra Demontis, Angelo Sotgiu, Maura Pintor, Battista Biggio; *Contributors:* Marta Kwiatkowska, Xiyue Zhang, Luca Oneto, Fabio Roli, Mario Fritz, Tatiana tommasi

- **D2.1 Report on Privacy and Infrastructures Grand Challenge and Benchmarking Metrics**
  Lead author: Antti Honkela; *Contributors:* Nicolò Cesa-Bianchi, Martin Jaggi, Catuscia Palamidessi, Massimiliano Pontil, Juliette Achdou, Joonas Jälkö, Raouf Kerkouche, Kangsoo Jung, Mario Fritz

- **D3.1 Grand Challenge Report: Human-in-the-loop decision making: integrated governance to ensure meaningful oversight**
  Lead Authors: Plamen Angelov, Dmitry Kangin, Umang Bhatt, Karen Yeung, Rotem Medzin; *Contributors:* Mario Fritz, Dimosthenis Karatzas, Ruben Perez Tito, Mohamed Soubgui, Vincent d’Andecy, Patrick Perez, Battista Biggio



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.