THE MORALS OF ALGORITHMS

A contribution to the ethics of AI systems

NXP

1 INTRODUCTORY REMARKS

When discussing **artificial intelligence (AI)** with researchers and scientists, one can easily pick up on their enthusiastic optimism about the tremendous potential that AI-based applications carry for the whole of humankind. This includes machine learning, a subset of AI, which uses computer algorithms to automatically learn and improve from environmental data without being explicitly programmed.

From ***detecting COVID-19 in x-ray and CT scans***, to ***diagnosing skin or breast cancer*** far more reliably than any human physician could, ***preventing road fatalities***, ***detecting illegal rainforest logging*** or ***fighting wildlife poaching***: Machine-trained AI systems provide the means to tackle some of the world's most challenging health crises, social problems and environmental issues, potentially helping hundreds of millions of people in advanced, developing and emerging countries.

At the same time, there is increasing concern over potential threats that AI-powered systems pose. Not every new AI application is suitable to instill trust in people's hearts and minds. ***From racist chatbots, to discriminating algorithms or sexist recruitment tools***, there are regular reports about instances of AI "gone wrong."

Rapidly advancing forms of artificial intelligence can prevent accidents and improve transportation, health care and scientific research. Or, they can violate human rights by enabling mass surveillance, driving cyber attacks and spreading false news.

TEACHING MORALS TO MACHINES: THE BASICS

The fundamental ethical dilemma surrounding the questions of how to teach AI values and how to ensure our ethical preferences can be embedded in code will accompany us for quite some time. Casting ethical values into a set of rules for machines to follow is no trivial task.

To start with, not even humans appear to uniformly agree on what is right and what is wrong conduct. Ethical principles are – so it seems – quite culture-specific and vary considerably across different groups. In an attempt to show how divergent people's ethical values can be, MIT researchers created a platform called the ***Moral Machine***. This online tool is a variant of the classic "***trolley problem***." Study participants are asked to choose whose lives a self-driving car should prioritize in a potentially fatal accident. By asking millions of people around the globe for their solution to the dilemma, the researchers found that people's ethical judgement shows some variation across different cultures.

AI applications, building on big data and combined with the omnipresence of devices and sensors of the IoT, will eventually govern core functions of society. These applications will reach from education to health, science to business, right up to the sphere of law, security and defense, political discourse and democratic decision making. While some societies strive to leverage the potential of AI to achieve compliance with behavioral norms and regulation, other cultures are taking a more cautious approach when it comes to balancing the rights of the individual against the claims of society as a whole.

GOVERNING BODIES

The EU and U.S. are both working on policies and laws targeting artificial intelligence.

Both entities have started to draft guidelines for AI-based applications to serve as ethical frameworks for future global usage. These guidelines draw inspiration from the ***Universal Declaration of Human Rights***.

The common thread in these guidelines is a human-centric approach in which human beings maintain their unique status of primacy in civil, political, economic and social fields. The EU's ethical guidelines on AI mirror its basic ethical principles by operationalizing these in the context of AI: cultivating the constitutional commitment to protect universal and indivisible human rights, ensuring respect for rule of law, fostering democratic freedom and promoting the common good. Other legal instruments further specify this commitment; for instance, the Charter of Fundamental Rights of the European Union or specific legislative acts such as the General Data Protection Regulation (GDPR).

The European Commission has set up the High-Level Expert Group on Artificial Intelligence (AI HLEG). This group has defined respect of European values and principles as set out in the EU Treaties and Charter of Fundamental Rights as the core principle of its "human-centric" approach to AI ethics. In line with its guidelines for trustworthy AI systems, the group has laid down three principles that need to be met throughout the entire lifecycle of an AI system:

Lawful AI

Compliance with all laws and regulations, regardless of the positive (what may or should be done) or negative (what cannot be done) nature of imposed rules of conduct.

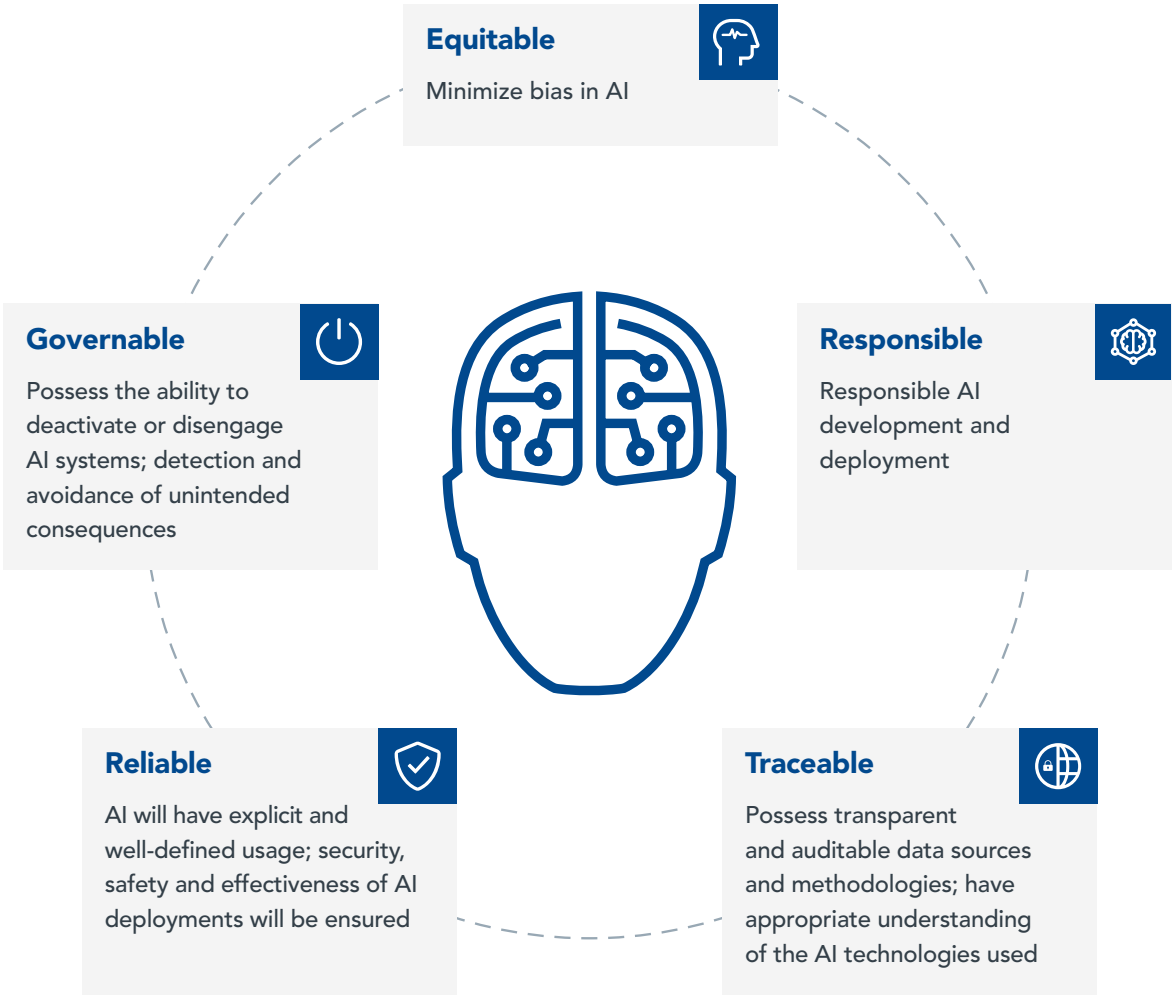
Robust AI

Systems should be able to perform in a safe and secure manner, where safeguards are carefully installed in order to avoid unintended adverse impacts. Therefore, it is necessary to ensure that AI systems are robust, not only from a technological, but also from a social perspective.

Ethical AI

Adherence to ethical principles and values. Cultivating the societal commitment to protect universal and indivisible human rights, ensuring respect for rule of law, fostering democratic freedom and promoting the common good. Holding up a human-centric approach, in which human beings maintain their unique status of primacy in civil, political, economic and social fields.

HOLISTIC APPROACH TO ETHICAL AND TRUSTWORTHY AI



In spite of the fact that each of these principles is fundamental for a successful deployment of AI systems, people should not regard them as comprehensive or self-sufficient. It is up to us as a society to ensure their harmonization and alignment.

In addition to these guidelines and principles the HLEG has issued an ***assessment list*** to aide stakeholders to check their policies and processes against their requirements for trustworthy AI.

PRIVATE SECTOR

Governments are not the only entities advancing ethical AI frameworks. As companies look to maintain a competitive edge in a continually evolving marketplace, businesses leading the AI sector, such as **IBM**, **Google**, **Microsoft** and **Bosch**, have published their own AI ethical principles.

Other companies, including Facebook and Amazon, have joined consortiums such as the Partnership on AI (PAI) and the Information and Technology Council (ITI). These companies also support standardization bodies, such as the IEEE®, that are driving ethical principles for the design of AI systems. These latter consortiums have developed and published their own ethical codes.

In a nutshell, these principles are based on:

Transparency

The decision-making process of AI systems should be explainable in such terms that people are able to understand the AI's conclusions and recommendations. In the latter case, it is important to explain the nature and functionality of these algorithms. Users should be aware at all times that they are interacting with an AI system. We are aware that, by nature, certain algorithms do not offer capability to explain the entirety of the decision making. In this case, it is important to make clear the nature and functionality of these algorithms.

Fairness

Minimize algorithmic bias through ongoing research and data collection which is representative of a diverse population aligning with local regulations designed to avoid discrimination.

Safety

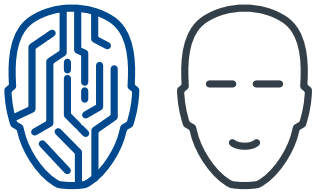
Develop and apply strong safety and security practices to ensure reliable performance and prevent tampering.

Privacy

Provide appropriate transparency and control over the use of data. Users should always maintain control over what data is being used and in what context. They must be able to deny access to personal data that they may find compromising or unfit for an AI to know or use.

2 OUR METHODOLOGY – NXP AI PRINCIPLES

Designing trustworthy AI/ML requires secure solutions for a smarter world that reflect ethical principles deeply rooted in fundamental NXP values. We focus on design, development and deployment of artificial intelligence systems that learn from and collaborate with humans in a deep, meaningful way.



1. The Principle of Non-Maleficence: "Be Good"

AI systems should not harm human beings. By design, AI building blocks should protect the dignity, integrity, liberty, privacy, safety and security of human beings in society and at work. Human well-being should be the desired outcome in all system designs.

Algorithmic bias in relation to AI systems that work with personal data should be minimized through ongoing research and data collection. AI systems should reflect a diverse population and prevent unjust impacts on people, particularly those impacts related to characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, disability, political affiliation or religious belief.

2. The Principle of Human Autonomy: "Human-Centric AI"

AI systems should preserve the autonomy of human beings and warrant freedom from subordination to — or coercion by — AI systems. The conscious act to employ AI and its smart agency, while ceding some of our decision-making power to machines, should always remain under human control, so as to achieve a balance between the decision-making power we retain for ourselves and that which we delegate to artificial agents as well as ensure compliance with privacy principles.

3. The Principle of Explainability: "Operate Transparently"

At NXP, we encourage explainability and transparency of AI-decision-making processes in order to build and maintain trust in AI systems. Users need to be aware that they are interacting with an AI system and they need the ability to retrace that AI system's decisions. Explainable AI models will propel AI usage in medical diagnoses, where explainability is an ethical practice often required by medical standards, as well as by EU regulations such as the GDPR.

Additionally, an AI system needs to be interpretable. The goal of interpretability is to describe the internals of the system in a way that is understandable to humans. The system should be capable of producing descriptions that are simple enough for a person to understand. It should also use a vocabulary that is meaningful for the user and will enable the user to understand how a decision is made. This might also reveal issues on inadequate or insufficient data used for model training.

At the same time, we are aware that measures to protect IP and to provide safety and security features will require maintaining secrecy about certain operating principles. Otherwise, knowledge of internal operating principles could be misused to stage an adversarial attack on the system.

4. The Principle of Continued Attention and Vigilance: "High Standards and Ecosystems"

We aspire to the highest standards of scientific excellence as we work to progress AI development. Drawing on rigorous and multidisciplinary scientific approaches, we promote thought leadership in this area in close cooperation with a wide range of stakeholders. We will continue to share what we've learned to improve AI technologies and practices. Thus, in order to promote cross-industrial approaches to AI risk mitigation, we foster multi-stakeholder networks to share new insights, best practices and information about incidents. We also foster these networks to identify potential risks beyond today's practices, especially when related to human physical integrity or the protection of critical infrastructure.

As designers and developers of AI systems, it is an imperative to understand the ethical considerations of our work. A technology-centric focus that solely revolves around improving the capabilities of an intelligent system doesn't sufficiently consider human needs. By empowering our designers and developers to make ethical decisions throughout all development stages, we ensure that they never work in a vacuum and always stay in tune with users' needs and concerns.

5. The Principle of Privacy and Security by Design: "Trusted AI Systems"

AI must rely on two basic principles: security by design and privacy by design. Security and privacy must be taken into account at the very beginning of a new system architecture; they cannot be added only as an afterthought. We must adopt the highest appropriate level of security and data protection to all hardware and software, ensuring that it is pre-configured into the design, functionalities, processes, technologies, operations, architectures and business models. This also requires establishing risk-based methodology and verification to be implemented as baseline requirements for the entire supply chain. In our view, the ***Charter of Trust initiative for cybersecurity*** in the IoT has already provided an excellent template for this. When it comes to the attribution of liability in case of damages caused by AI-enabled products or services, the implementation of state-of-the art privacy and security technology can also serve as a key criterion to be assessed.

Privacy by design is enabled through secure management of user identities and data security. Traditional software attack vectors must still be addressed, but they do not provide sufficient coverage in the AI/ML threat landscape. The tech industry should consider not approaching next-gen issues with last-gen solutions. Instead, it should strive to build new frameworks and adopt new approaches which address gaps in the design and operation of AI/ML-based services.

NXP considers privacy to be a pivotal human right. We are committed to the concept of privacy by design and to incorporating an appropriate level of security and data protection in hardware and software, within the realm of our control.

But once corporate principles are set up, how do we ensure that AI systems are designed and – more importantly – deployed in an ethical way? To that end, ethics officers and review boards should be empowered to oversee these principles and warrant their compliance with ethical frameworks.

ROOTING MORALS ON THE PHYSICAL LEVEL

This leads us to the underlying question: once we have established a basic set of ethical rules for AI, how do we ensure that AI-based systems cannot become compromised? What is often overlooked is the fact that, in order to implement ethical imperatives, we first have to trust AI on a more fundamental level. Can we?

While progress in AI has unfortunately coincided with the development of new cyber threats, machine learning (ML) can help ensure that AI-based systems aren't compromised by them. In fact, ML security has been routinely applied to ***SPAM and malware detection***, where it can expand or enhance anomaly detection in the data stream.

However, ML can be a double-edged sword. While ML enables industry-grade malware detection programs to work more effectively, it will be misused by criminal masterminds to enhance the offensive capabilities of their attacks. In order to prevent or defend against these attacks, we must focus on how to leverage hardware security to improve overall system security and data privacy.

At NXP, we have built some of the most sophisticated secure devices in the world. We equip them with countermeasures that repel a broad range of logical and physical attacks, such as side-channel or remote attacks. It's only a matter of time until hackers will employ AI to extract secrets and critical information from secure systems. We must anticipate and shore up defense mechanisms against upcoming challenges like these.

Thus, the very first steps towards trustworthy AI, capable of complying with our ethical guidelines, are the integrity of the system, the integrity of its data and the security of inference. Just as it is in functional and data security of an IoT system, AI systems must follow principles of security and privacy by design in order to uphold appropriate security and privacy-preserving features.

Security by design

Security by design relies on well-known system properties:

- Capacity to ensure service availability
- Confidentiality for keeping secrets secret
- Integrity to guarantee unmodified data transport and software execution
- Authenticity by means of identity verification for trusted operations

We strive to combine these four properties to:

- Offer end-to-end security
- Comply with future security standards
- Counter potential attacks
- Act as cost-effective and safety-oriented security protection mechanisms

The difficulty is that the ones making the decisions on the appropriate security levels and features, as well as the ones implementing them, are not the ones who will bear the losses. In this context, global security certification provides a solid framework for the trustworthiness of systems and components and creates a level playing field for all participants in the market.

Privacy by design

- Offers secure storage of keys and secure management of user identities while respecting privacy settings
- Provides individual device identities in a privacy preserving way
- Enables the deployment of confidential communication

A privacy-by-design approach means storing privacy-sensitive information only when it is absolutely necessary and never computing unencrypted privacy-sensitive information.

Privacy preservation adoption will happen globally, driven by new legislation and regulations such as the General Data Protection Regulation (GDPR) in Europe. These new regulations restrict the data collection that is allowed to be performed by devices and services. For example, EU data protection not only restricts solely automated individual decision-making, i.e. without any human involvement, but also profiling, i.e., automated processing of personal data to evaluate certain things about an individual. As per the GDPR, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, among other stipulations.

3 SECURITY IN AI: BUILDING TRUST

The biggest challenge for AI is that there are many independent actors within the ecosystem. Added to this, the share of liability and the levels of trust between these actors are not clear-cut.

If system security is not adequately contemplated during design, overall security of deployed devices and services might potentially collapse. Encouraging security and privacy by design as guiding principles and prerequisites for the deployment of safe AI systems will add value for all players inside the ecosystem.

AI industry should aspire to introduce elaborate security management across entire product lifecycles. All stakeholders involved in developing and operating AI ecosystems should work together towards interoperable and assessable security. Relevant stakeholders span device and component manufacturers, service providers, governments, standardization bodies and educational institutions.

So, what is the most important consideration when discussing the security foundations of an AI system? As already mentioned, we strive toward incorporating security throughout each stage of product lifecycle – it is our goal to make it multi-faceted and deeply embedded in everything from edge devices that use neural-network processing system-on-chips, to the applications that run on them, and cloud communication and storage.

Security is a matter of getting the details right. Even the smallest of errors, made at any point in the implementation, can eventually create weaknesses and put the overall design at risk. Effective security solutions are the result of a strict development process with clearly defined design rules, multiple iterations of careful review, and full control over the many sub-components involved from design to system deployment and maintenance.

A few foundational functions for enabling AI security must be considered in order to protect the product during all phases of its operation: when offline, during power up, at runtime and during communication with other devices or the cloud. As a guiding principle, ensuring permanent integrity of the system is essential to creating trust in its behavior.

Secure Boot

Secure boot ensures system integrity by using a cryptographic signature on the firmware to guarantee that, when the system comes out of reset, it does exactly what it's intended to do first. More specifically, by having a secure boot system in place that relies on a hardware root of trust, the root public key is protected, ensuring that it cannot be altered. Protecting the public key in hardware ensures that the identity of the root of trust can be determined, and that it cannot be forged.

Key Management

The secret key material must be kept inside a hardware root of trust, since that would allow only application-layer clients to manage the keys indirectly through application programming interfaces (APIs). To ensure continued protection of these secret keys, it's imperative to authenticate the importing of keys and to wrap exported keys.

Secure Updates

As AI applications get more sophisticated, the data and models need to be updated continuously. These updates must be traceable to ensure that former weaknesses of the system are documented. The process of securely distributing new models needs to be protected with end-to-end security. Hence, it is essential that products can be updated to fix bugs, close vulnerabilities, and evolve product functionality in a trusted way. A flexible, secure update function can even be used to allow post-consumer enablement of optional hardware or firmware features.

Secure Communications

To ensure that communications between edge devices and the cloud are secured and authentic, designers use protocols that incorporate mutual identification and authentication. This, on the other hand, ensures that only secure and authenticated data is allowed to flow between the systems, preventing the inputs to the neural network from being altered. A hardware root of trust will enforce security of credentials used in establishing identification and authentication, as well as confidentiality and authenticity of the data itself.

Protecting the Integrity of the ML Model

As already mentioned, there are serious privacy concerns when it comes to protecting data rooted in working memory, or stored locally on disk or flash memory systems. To mitigate these concerns, one requires high-bandwidth memory encryption, backed by strong key-management solutions. Similarly, ML models should be protected through encryption and authentication, as well as backed by strong key-management systems, which in turn are enabled by a hardware root of trust.

Protection Against Adversarial Attacks and Misuse

Also known as adversarial examples, these intentional attacks are inputs to ML models designed to cause them to make a mistake. They're like optical illusions for machines. This is an important problem in AI security that requires our constant vigilance. At NXP, we are working on improving resilience against such attacks by providing adversarial training to explainable and interpretable AI models that will flag uncertainties to the user before making a decision. In a similar context, NXP is exploring the use of digital watermarks to prevent unauthorized use of ML models.

Protection Against Data Poisoning

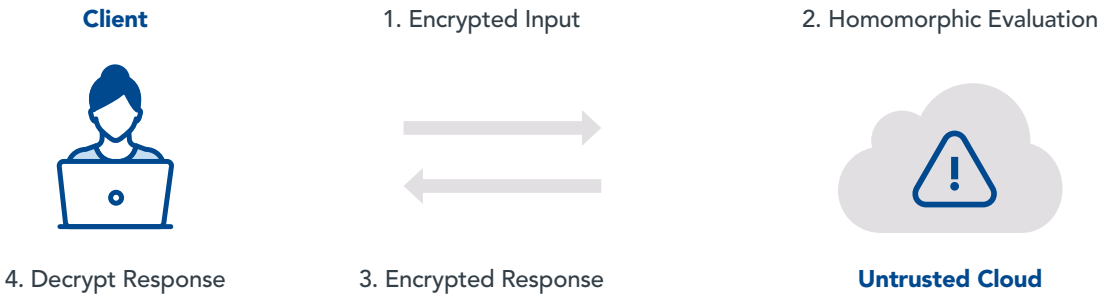
ML systems trained on user-provided data are susceptible also to data poisoning attacks, whereby malicious users inject false training data with the aim of corrupting the learned model. To prevent this, careful control of the data that is used to train any AI model is paramount. How much the attacker knows about the system matters. Detection of malicious users must be implemented with high accuracy from the very beginning. Simple human negligence could pose a similar problem. Therefore, training of AI systems should be entrusted to expert data scientists exclusively.

ENCRYPTION – THE BUILDING BLOCK FOR SECURE COMMUNICATIONS AND DATA PROTECTION

One of the biggest privacy challenges for AI systems is the capacity of processing input data, while still respecting user data privacy.

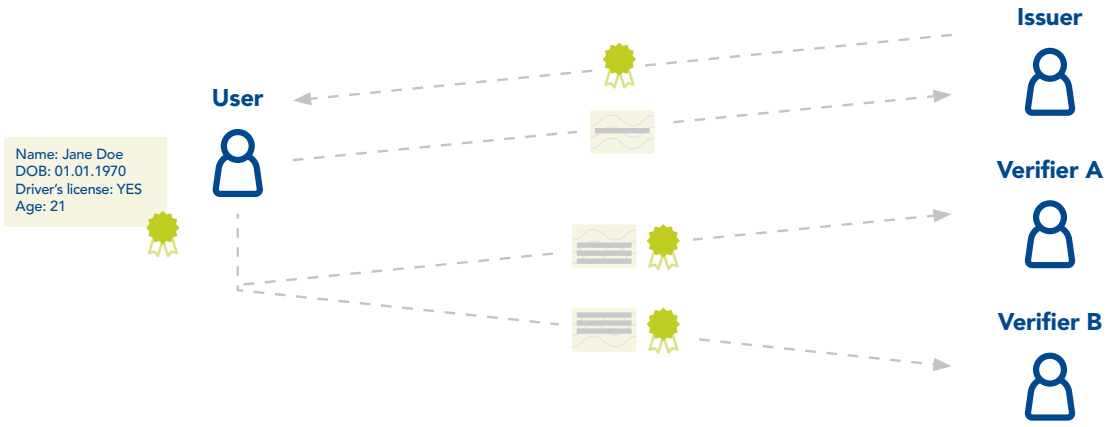
Therefore, it is becoming increasingly important for all communications to be encrypted. By implementing encryption schemes, key security benefits, such as confidentiality and the integrity of data and authentication, are achieved.

While traditional cryptography techniques such as symmetric or asymmetric encryption can be useful in decentralized AI scenarios, they fall short of enabling many key requirements for these types of systems. To overcome this handicap, decentralized AI systems have started to embrace some of the most advanced cryptography techniques coming out of academic research. Specifically, there are two security methods that are increasingly important for decentralized AI architectures: homomorphic encryption and attribute-based encryption, and more recently secure multi-party computation (MPC).



Homomorphic encryption

Homomorphic encryption is privacy-enhancing technology that encrypts data into computable cipher text. Any data being used in the computation remains encrypted and only becomes visible for the intended user. The result of the computation, once decrypted, matches the result of the same computation applied to the plain text. NXP continues to do pioneering work in this area.



Attribute-based cryptography

Attribute-based cryptography is an encryption scheme in which decryption is conditioned by the user's specific values of attributes in order to support, among others, anonymous operations, which are part of the toolbox available to enhance user privacy while employing AI products. Attribute-based authentication allows for strong authentication and privacy at the same time. It relies on a combination of flexible public keys (pseudonyms) and flexible credentials that allows a user to share only the information required for a certain transaction, without revealing any other attributes.

CERTIFICATION

Yet, even if security protection techniques are combined to achieve appropriate system security, it is still essential to convince stakeholders that they can trust the system. This is especially important in the context of user privacy and data integrity.

As consumers and service providers seek greater assurance that AI products are adequately protected, it becomes increasingly important to turn to standardized test and validation methods. The best way to achieve this is through aligned and trusted certification standards. This provides third-party validation of security claims made for all parts of the chain of trust including ICs, components, devices and services, all based on agreed testing and verification procedures and processes. Unfortunately, the existing certification schemes developed for other ecosystems, e.g., Common Criteria, FIPS 140-2, etc., prove inadequate to face the challenges of AI. They are too rigid and tailored to their respective segments, especially with respect to the constantly evolving system that AI brings. NXP is actively participating in the definition of a European Certification Framework for AI.

OUR COMMITMENT

AI opens entirely new opportunities across industries to help solve global challenges. Working in close collaboration with leading academic institutions, research organizations and pioneering technology firms, NXP is at the forefront in the development of AI solutions, driving this transformation with secure, connected processing solutions, and helping enable a boundless multitude of future applications. Given our responsibility for secure and trustworthy AI, we as NXP aspire to uphold the above principles and encourage our customers, partners and stakeholders to support us in this endeavor.

Above all, in an ever-moving world of technology excellence, we depend on strong relationships with our customers and with the end users of our technologies. It is only together that we can refine test methods, match products to applications and ensure correct application conditions. Ongoing customer feedback regarding quality levels achieved on their assembly lines and in service is a vital part of this collaboration.

Given this, let's remind ourselves that the future is nothing to be afraid of, but for us to shape.

NXP offers a comprehensive portfolio of MCUs and processors optimized for machine learning applications in automotive, smart industrial and IoT industries. Our eIQ™ machine learning software development environments include inference engines, neural network compilers, optimized libraries and deep learning toolkits designed for easier and more secure system-level application development, ML algorithm enablement and auto quality ML enablement. NXP EdgeReady solutions provide out-of-the box edge compute intelligence to rapidly take voice control and facial recognition capabilities from concept to product launch. Explore NXP solutions at nxp.com/eIQ.

www.nxp.com

Document Number: AIETHICSWP REV 3