SPARTA Roadmap Template - Example

**Table 1.** General information

|  |
| --- |
| **Title:** Trustworthy AI |
| **Problem description:**  The use of AI is spreading among many industry sectors and organizations. While it offers new functionality and decision support, there is a lack of development in the direction of security and trust in AI. |
| **Final goal:** Defined concepts and implemented processes for trustworthy AI systems and their use by the citizens and businesses. |
| **Status Quo:**   * **Europe:** Multiple universities in Europe performing research in adversarial and secure machine learning, ALOHA 2020 Project * **International:** DARPA Robust AI Programs |
| **Estimated year of completion:** 2029 |
| **Research aspect:** We need to investigate possible threats for AI systems, make AI more secure and in the end develop methods to enable AI systems to be trustworthy. |
| **Industrial demand:** Industry is increasingly adopting AI, including the security-sensitive areas, such as medical, automotive or e-government. A trustworthy AI approach would further increase this adoption and enable strengthening of the industry. |
| **Social aspect:** Users of AI technology are concerned with the possibility of manipulation of AI decisions and introducing bias, which erodes their trust in AI systems. We need to develop trustworthy AI to increase the benefits that AI can provide for the society. |
| **Benefit for EU:** AI industry in EU would gain a competitive edge with this technology, as it would enable them to offer better products to market. |
| **SWOT Analysis**   * **Strengths:** Many research institutions in EU are working on this problem * **Weaknesses:** Lack of knowledge and expertise in the industry in making AI secure * **Opportunities:** Strengthening the industry through providing new concepts for increasing product value using trustworthy AI * **Threats**: Lack of straightforward path to the solution |
| **Domain (JRC Taxonomy):** Theoretical Foundation, Human Aspects, Legal Aspects, Data Security |
| **Sector (JRC Taxonomy):** Health, Energy, Financial, Government, etc. |
| **Relation to Emerging Technologies:** Autonomous vehicles, Artificial General Intelligence, Brain-computer Interfaces |

**Table 2.** Timeline (including dependencies):

(State subgoals as stages until final goal and their dependencies. Describe stages in Table 3)

Dep: set of completed subgoals as precondition

ID: the ID of the subgoal

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dimension** | **Stage 1** | | **Stage 2** | | **Stage 3** | | **Stage 4** | | **Stage 5** | |
|  | Dep | ID | Dep | ID | Dep | ID | Dep | ID | Dep | ID |
| **Certification (1.\*)** |  |  |  |  |  |  |  |  | 3.5 | 1.1 |
| **Education (2.\*)** |  |  | 3.1 | 2.1 |  |  |  |  |  |  |
| **Technology (3.\*)** |  | 3.1 | 3.1 | 3.2 |  | 3.3 | 3.2 | 3.4 | 3,1, 3.3, 3.4 | 3.5 |
| **Exp Duration** | 1 year | | 1 year | | 1 year | | 2 years | | 5 years | |

**Table 3.** Detailed description

|  |  |  |  |
| --- | --- | --- | --- |
| **Stage/Dimension** | **Sector (JRC)** | **Domain (JRC)** | **Regulation** |
| **1.1** | Health, Energy, Financial, Government... | Data Security,  Certification | Legal framework for certification |
| **Description (incl. obstacles):**  Certified AI  Development of methodology for certification of AI systems. | | |
| **1.2** |  |  |  |
| **Description (incl. obstacles):** | | |
| **1.3** |  |  |  |
| **Description (incl. obstacles):** | | |
| **1.4** |  |  |  |
| **Description (incl. obstacles):** | | |
| **1.5** |  |  |  |
| **Description (incl. obstacles):** | | |
| **2.1** | Health, Energy, Financial, Government... | Education and Training |  |
| **Description (incl. obstacles):**  Education programs on AI security threats for AI engineers and decision makers | | |
| **2.2** |  |  |  |
| **Description (incl. obstacles):** | | |
| **2.3** |  |  |  |
| **Description (incl. obstacles):** | | |
| **2.4** |  |  |  |
| **Description (incl. obstacles):** | | |
| **2.5** |  |  |  |
| **Description (incl. obstacles):** | | |
| **3.1** | Health, Energy, Financial, Government... | Data Security |  |
| **Description (incl. obstacles):**  Threat modeling for AI  Development of a comprehensive approach to detect threats and assess severity. | | |
| **3.2** | Health, Energy, Financial, Government... | Data Security |  |
| **Description (incl. obstacles):**  Defensive methods for robust AI  Robust AI methods and development of detection of a wide range of adversarial attacks. | | |
| **3.3** | Health, Energy, Financial, Government... | Data Security,  Human Aspects,  Legal Aspects | Influence of biased AI on legal processes. |
| **Description (incl. obstacles):**  Explainable and fair AI  Development of methods that include human-readable explanation of AI functionality and decision making. Furthermore, this decision making should exclude unfair bias. | | |
| **3.4** | Health, Energy, Financial, Government... | Data Security,  Theoretical Foundation |  |
| **Description (incl. obstacles):**  Proactive AI protection  Approaches to model the possible activities of the adversary and proactively protect against future attacks. | | |
| **3.5** | Health, Energy, Financial, Government... | Theoretical Foundation, Human Aspects, Legal Aspects, Data Security | Legal framework for development and use of trustworthy AI systems |
| **Description (incl. obstacles):**  Trustworthy AI systems  Provable security mechanisms for AI systems, relying on formal verification methods. | | |