

Still just a machine

Security issues in the use of artificial intelligence

Workshop #7



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Agenda

1. Welcome and Introduction
2. Introduction to AI and Security Issues
3. Why AI Should Not Be Trusted Unconditionally
4. Common Security Threats in AI
5. Online Tools for Securing AI
6. Hands-On Session: Playing Prompt Injection Games
7. Hands-On Session: Exploring Safe AI Sandbox Environments
8. Real-World Applications
9. Q&A and Discussion



**Funded by
the European Union**

**WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)**

Introduction to AI and Security Issues



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Understanding and mitigating the risks

AI is powerful but can be risky if not properly secured

- AI systems making mistakes due to tampered data
- Breaches of private information
- Presenting wrong answers in a convincing way



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Why AI Should Not Be Trusted Unconditionally

Limits and Risks of AI

- **Knowledge Cutoff:** AI may not have the latest information
- **Bad Actors:** Malicious manipulation of AI outputs
- **Bias:** AI can inherit and amplify biases from training data
- **Training Data Poisoning:** Corruption of data used to teach AI
- **Open Source vs. Proprietary Models:**
 - Open Source: More transparency, but potentially more vulnerabilities
 - Proprietary: Less transparency, but potentially better security controls



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Common Security Threats in AI



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Common Security Threats in AI

Common Problems in AI Security

- **Adversarial Attacks:** Tricks AI into making errors
- **Data Poisoning:** Giving AI bad information
- **Model Inversion:** Extracting sensitive data
- **Privacy Breaches:** Unauthorized access to data



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Hands-on Activity: Playing Prompt Injection Games



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Playing Prompt Injection Games

Gandalf's Spellbook

- Visit a prompt injection game site: Gandalf's Spellbook
- Follow the prompts and intentionally interact with the AI to uncover vulnerabilities
- Compare results with peers and discuss findings



<https://gandalf.lakera.ai>



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Playing Prompt Injection Games

TensorTrust AI

- Visit a prompt injection game site: TensorTrust AI
- Follow the prompts and intentionally interact with the AI to uncover vulnerabilities
- Compare results with peers and discuss findings



<https://tensortrust.ai>



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Hands-on Activity: Exploring Safe AI Sandbox Environments



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Exploring Safe AI Sandbox Environments

AI Dungeon

- Visit a prompt injection game site: AI Dungeon
- Start a session
- Interact with the AI to see how it handles different inputs
- Discuss the AI's strengths and weaknesses in a secure environment



<https://play.aidungeon.com>



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Real-World Applications



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Real-World Applications

Real-World Examples and Risks

- **Healthcare:**

- **Risk:** Exposure of patient data
- **Fix:** Use strong access controls

- **Finance:**

- **Risk:** Fraud through AI mistakes
- **Fix:** Monitor transactions carefully

- **Self-Driving Cars:**

- **Risk:** Incorrect reactions to trick signals
- **Fix:** Thorough testing of AI systems



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)

Discussion points

- How can we stay ahead of AI security problems?
- Which fields are most at risk with AI?
- Why is it important to check AI systems regularly?
- What can happen if AI security is ignored?
- Which tools or practices are best for AI safety?
- How do we balance new AI tech with security needs?



**Funded by
the European Union**

**WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)**

Thank you



**Funded by
the European Union**

WORKSHOPS FOR YOUNG PEOPLE “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people” within the framework of the small-scale cooperation partnership project in the Youth sector of the Erasmus+ Program “Technologies of tomorrow - combating disinformation and building security capacity in the use of artificial intelligence by young people.”
(2023-1-ES02-KA210-YOU-000164824)