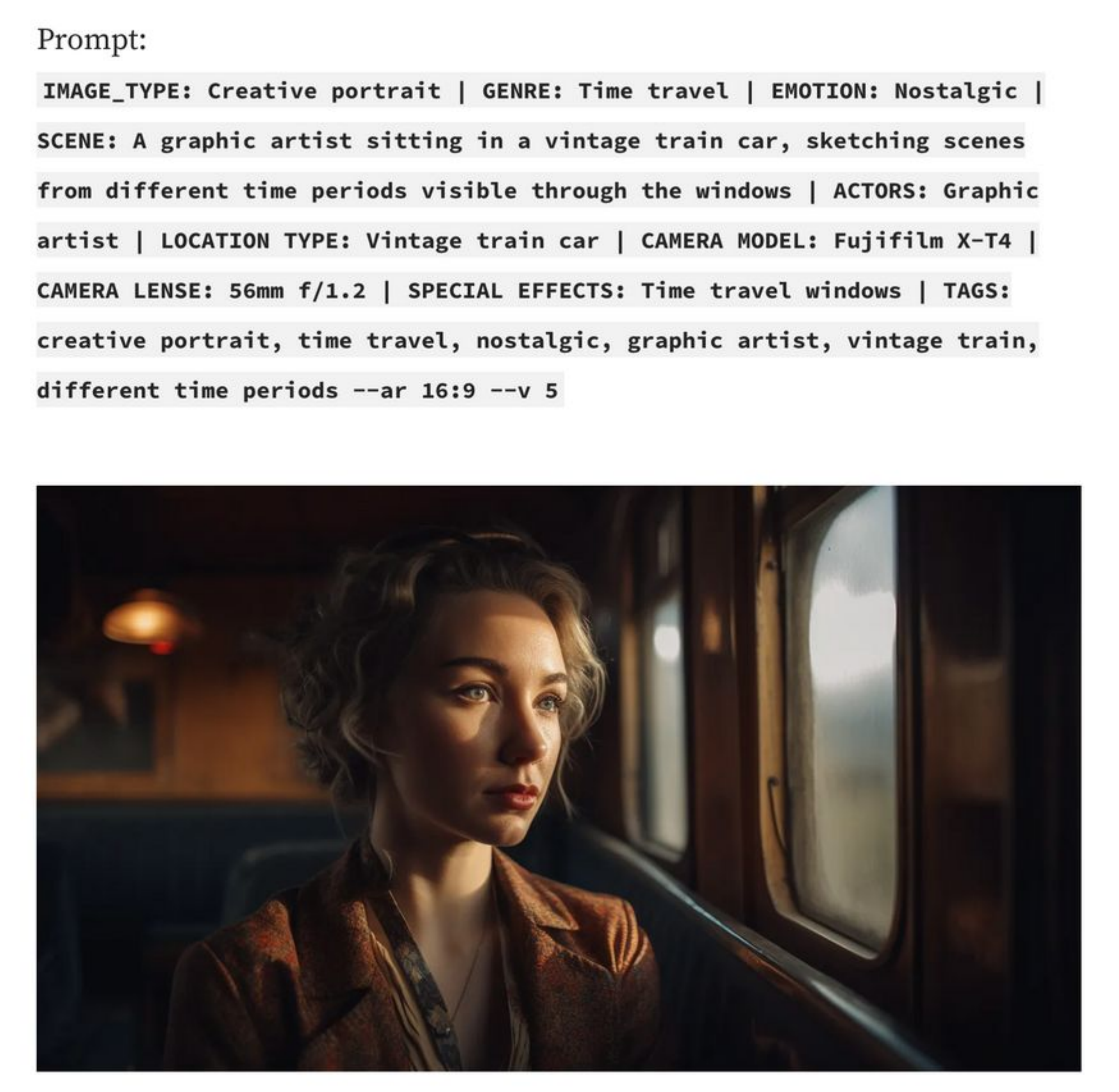
Identifying and Mitigating the Security Risks of GenAI

Clark Barett et al, 2023

Midjourney

ChatGPT

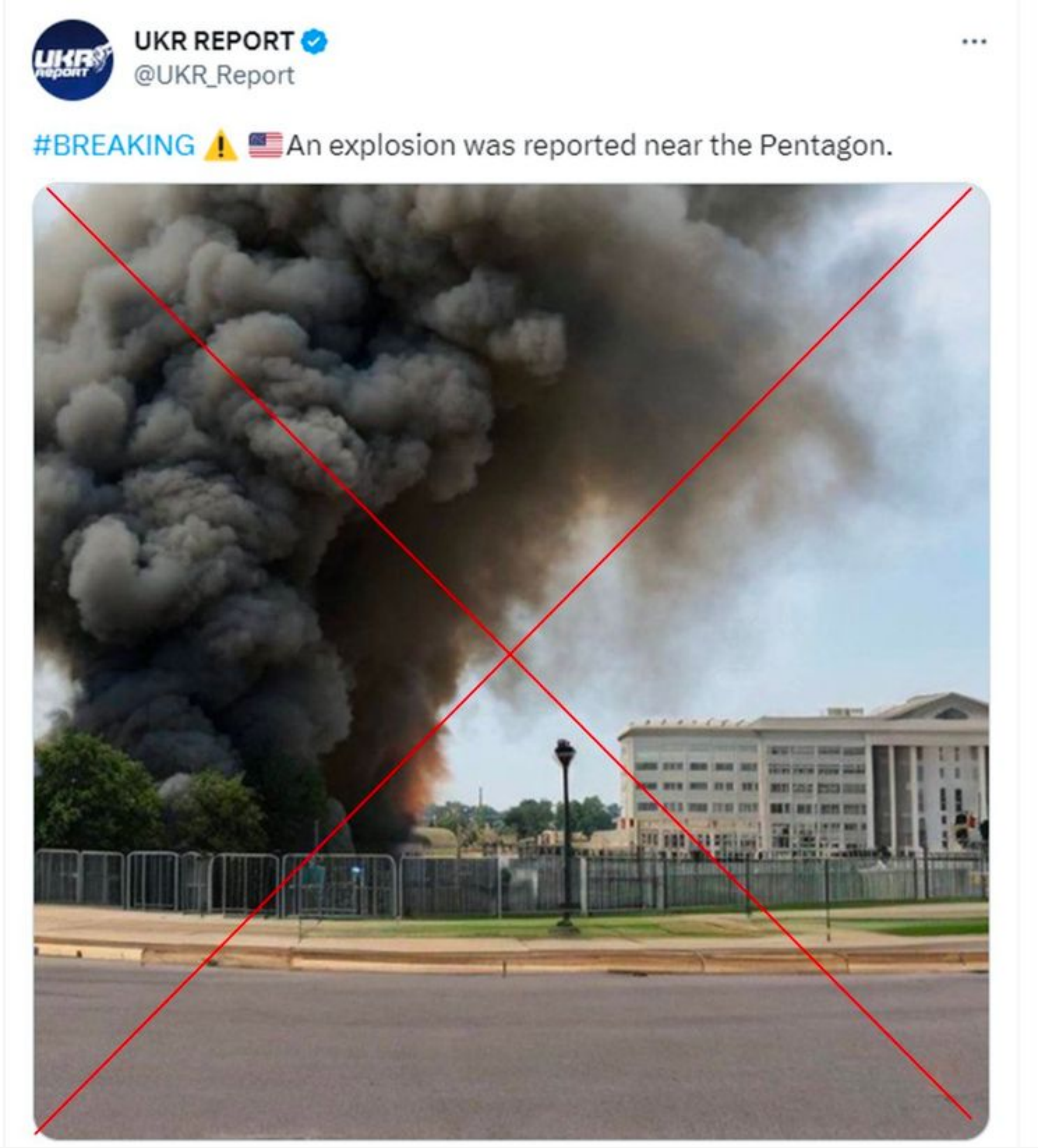
- **Gen AI's reappearance raises the issue of the "dual-use dilemma" as it can be employed for both positive and negative objectives.**



Prompt:

```
IMAGE_TYPE: Creative portrait | GENRE: Time travel | EMOTION: Nostalgic |
SCENE: A graphic artist sitting in a vintage train car, sketching scenes
from different time periods visible through the windows | ACTORS: Graphic
artist | LOCATION TYPE: Vintage train car | CAMERA MODEL: Fujifilm X-T4 |
CAMERA LENSE: 56mm f/1.2 | SPECIAL EFFECTS: Time travel windows | TAGS:
creative portrait, time travel, nostalgic, graphic artist, vintage train,
different time periods --ar 16:9 --v 5
```

good use of Midjouney, a diffusion based GenAI



UKR REPORT
@UKR_Report

#BREAKING ⚠️ 🇺🇸An explosion was reported near the Pentagon.

bad use: deep fake spreading misinformation

Overview of Some Attacks

- **Spear-phishing e.g. well-curated scam emails**
- **Hallucinations e.g. New York Lawyer citing non-existent cases**
- **Dissemination of deep fakes**
- **Proliferation of cyberattacks**
- **Low barrier of entry for adversaries e.g ChaosGPT, WormGPT, FraudGPT**
- **Lack of social awareness and human sensibility**
- **Unpredictability: we don’t fully know the extent of their threats**

Overview of Some Defenses

- Detecting LLM content eg. DetectGPT
- Watermarking
- Code analysis
- Penetration testing
- Personalized skill training
- Human-AI collaboration

Short-Term Goals

- Use cases for emerging defense techniques
- Current SOTA for LLM-enabled code analysis
- Alignment of LLM-enabled code generation to secure coding practices e.g Reinforcement Learning from Complier Feedback(RLCF), Controlled Code Generation
- Repository and service of SOTA attacks and defenses

Detection Algorithms for AI-Generated Content

- Neural network-based detectors
- Zero-shot detectors
- Retrieval-based detectors
- Watermarking-baseddetectors

Long Term Goals

- Need for socio-technical solutions: new model evaluation metrics
  - Suggestion:
    - Developing an online reputation system
    - Accountability for deliberate misuse and negligence
  - Any solution should avoid overwhelming users
- Multiple lines of defense
  - Training-time interventions to align models with predefined values
  - Post-hoc detection and filtering of inputs and outputs to catch inappropriate content that might slip through.
- Pluralistic value alignment

Long Term Goals

- Reduce barrier-to-entry for GenAI research
- New partnerships among stakeholders
- Grounding
    - Detecting whether a given LLM response is grounded
        - Use a separate natural language inference(NLI) model to test whether the generated text is entailed by the knowledge text.
    - Encouraging LLMs to generate grounded responses
        - Augment the prompt with relevant knowledge snippets
        - Tuning the LLM to generate grounded responses with relevant citations
        - Use reinforcement learning to tune the weights based on feedback on the groundedness and plausibility of generated responses

THANK YOU