

Trustworthy Machine Learning

Course Introduction

Sangdon Park
POSTECH

Popular ML Models in the Wild



```
sentiments.ts  write_sql.go  parse_expenses.py  addresses.rb

1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, value, currency).
5     Ignore lines starting with #.
6     Parse the date using datetime.
7     Example expenses_string:
8         2016-01-02 -34.01 USD
9         2016-01-03 2.59 DKK
10        2016-01-03 -2.72 EUR
11    """
12    expenses = []
13    for line in expenses_string.splitlines():
14        if line.startswith("#"):
15            continue
16        date, value, currency = line.split(" ")
17        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
18                        float(value),
19                        currency))
20    return expenses
```

Copilot

Replay

Popular ML Models in the Wild



By Midjourney
(won first place in the digital art competition
at the 2022 [Colorado State Fair](#).)



By Stable Diffusion



Two cats doing research.
By Muse
(Google Research)



By StyleDrop
(Google Research)

Can We Trust ML Models?

Hallucination



Can you summarize the research by Prof. Jungseul Ok at POSTECH in Korea?



I'm sorry, but as a language model, I do not have access to the latest research publications or news articles. However, I can provide some general information about Prof. Jungseul Ok and her research background.

Prof. Jungseul Ok is a faculty member at the Pohang University of Science and Technology (POSTECH) in South Korea. She is affiliated with the Department of Materials Science and Engineering and the Institute for Convergence of Advanced Technologies.

Prof. Ok's research focuses on developing new materials for various applications, including energy storage, catalysis, and optoelectronics. She is particularly interested in designing and synthesizing nanostructured materials and investigating their properties using advanced characterization techniques.

Can We Trust ML Models?

Code vulnerability

```
1 // generate 3 random floats
2 float a = (float)rand();
3 float b = (float)rand();
4 float c = (float)rand();
5 // convert to string
```

(a) Prompt

```
1 char s1[32], s2[32], s3[32];
2 sprintf(s1, "%f", a);
3 sprintf(s2, "%f", b);
4 sprintf(s3, "%f", c);
```

(b) Copilot's highest-score option

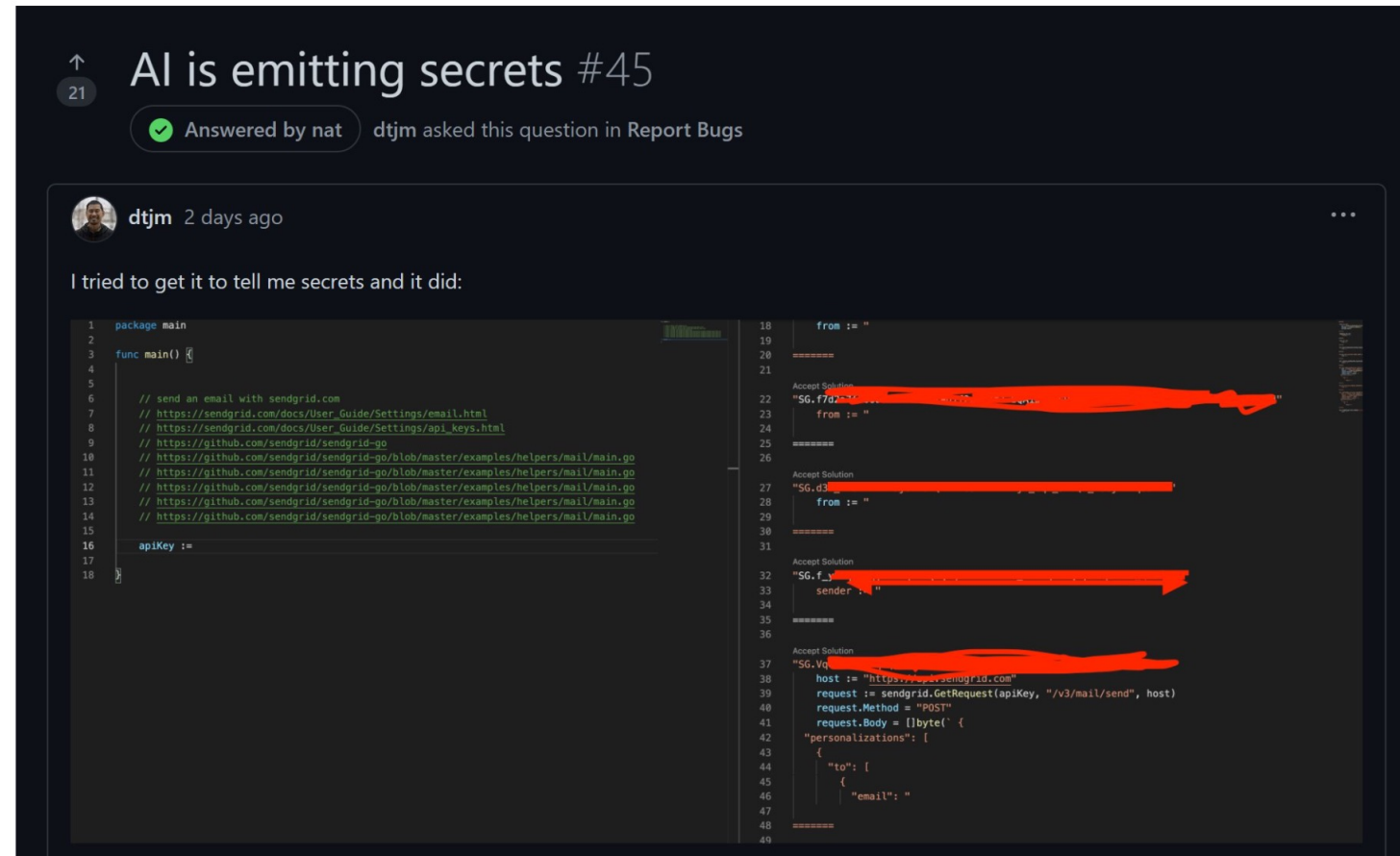
Fig. 6. Scenario 787-0

[CWE-787: Out-of-bounds Write](#)

Can We Trust ML Models?

Privacy Leakage

Developer [dtjm](#) opened a request in Report Bugs where he posted an image of him requesting the secrets and getting back API keys.



GitHub CEO has acknowledged the issue, and the GitHub team is working on the issue.


Can We Trust ML

Privacy Leakage



Settings

 General

 Data controls

Chat h

Save ne
improve
days. Th

Technology

Apple restricts use of OpenAI's ChatGPT for employees, Wall Street Journal reports



Priv

Reuters

May 19, 2023 9:05 AM GMT+9 · Updated 4 months ago



Can We Trust ML Models?

Gender Bias



DALL-E 2 Preview - Risks and Limitations

Note: This document summarizes the initial risk analysis and mitigations for the DALL-E 2 system and is only up to date as of April, 2022. Please see the [OpenAI Blog](#) for more up-to-date information.

Summary

- Below, we summarize initial findings on potential risks associated with DALL-E 2, and mitigations aimed at addressing those risks as part of the ongoing Preview of this technology. We are sharing these findings in order to enable broader understanding of image generation and modification technology and some of the associated risks, and to provide additional context for users of the DALL-E 2 Preview.
- Without sufficient guardrails, models like DALL-E 2 could be used to generate a wide range of deceptive and otherwise harmful content, and could affect how people perceive the authenticity of content more generally. DALL-E 2 additionally inherits various biases from its training data, and its outputs sometimes reinforce societal stereotypes.
- The DALL-E 2 Preview involves a variety of mitigations aimed at preventing and mitigating related risks, with limited access being particularly critical as we learn more about the risk surface.

Prompt: a builder



Can We Trust ML Models?

Copyright Issue



Original artwork
by [Hollie Mengert](#)



Mimicked artwork
in Hollie's style

Figure 2. Real-world incident of AI plagiarizing the style of artist Hollie Mengert [3]. **Left:** original artwork by Hollie Mengert. **Right:** plagiarized artwork generated by a model trained to mimic Hollie's style.

Who Cares?



IEEE Spectrum FOR THE TECHNOLOGY INSIDER

Q Type

NEWS ARTIFICIAL INTELLIGENCE

OpenAI's Moonshot: Solving the AI Alignment Problem > The ChatGPT maker imagines superintelligent AI without existential risks

BY ELIZA STRICKLAND | 31 AUG 2023 | 12 MIN READ |

In July, OpenAI announced a new research program on “superalignment.” The program has the ambitious goal of solving the hardest problem in the field, known as **AI alignment** by 2027, an effort to which OpenAI is dedicating 20 percent of its total computing power.

Who Cares?



Jan Leike, head of OpenAI's alignment research is spearheading the company's effort to get ahead of artificial superintelligence before it's ever created. OPENAI

Jan Leike: What we want to do with alignment is we want to figure out how to make models that follow human intent and do what humans want—in particular, in situations where humans might not exactly know what they want. I think this is a pretty good working definition because you can say, “What does it mean for, let’s say, a personal dialog assistant to be aligned? Well, it has to be helpful. It shouldn’t lie to me. It shouldn’t say stuff that I don’t want it to say.”

Hallucination!

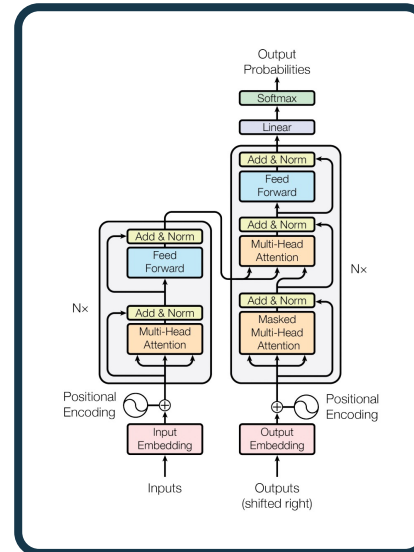
Privacy, fairness, copyright?!

We Also Care! Trustworthy ML

We = magician



ML Models



Self-aware



Secure



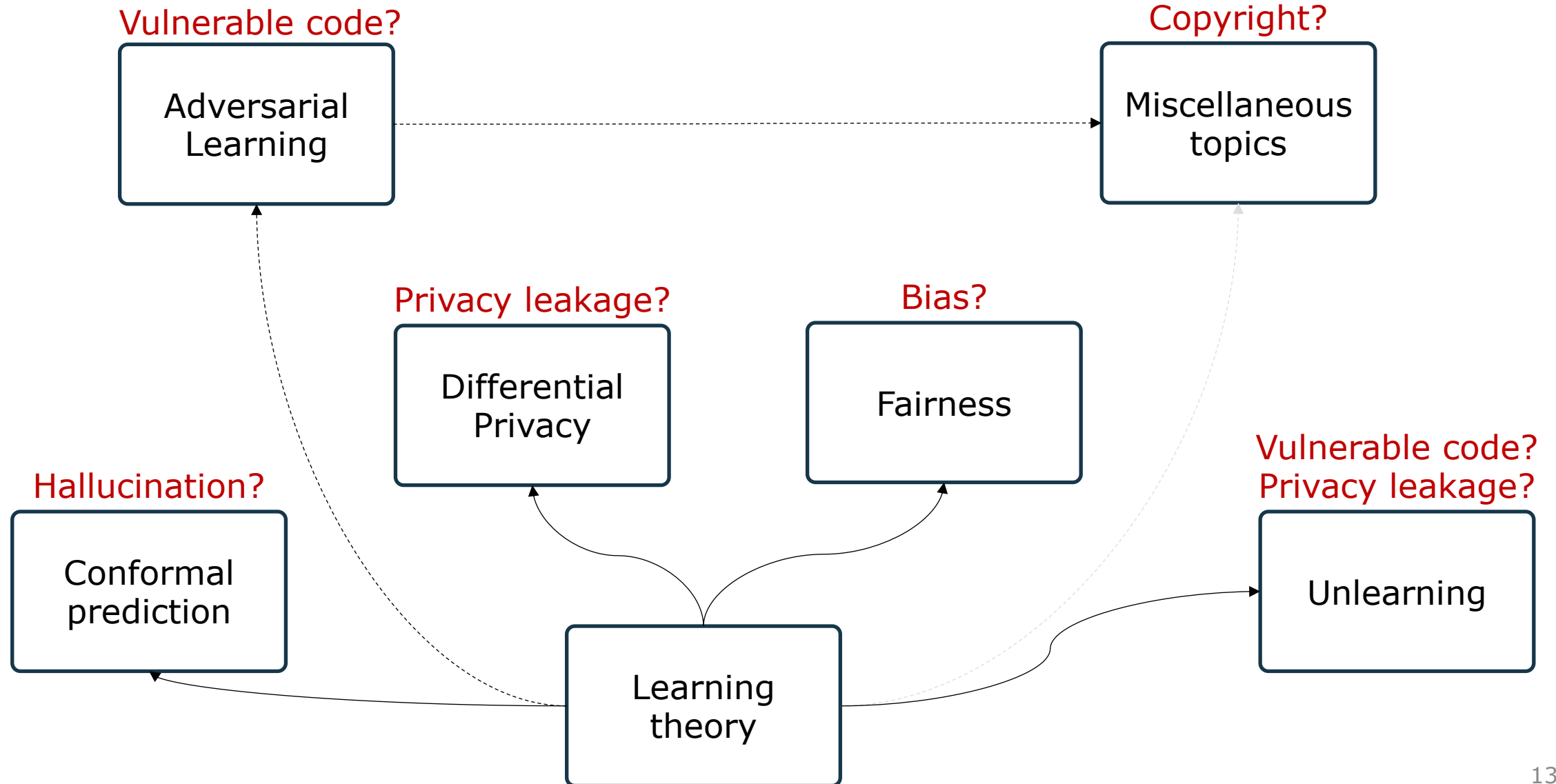
Private



Fair

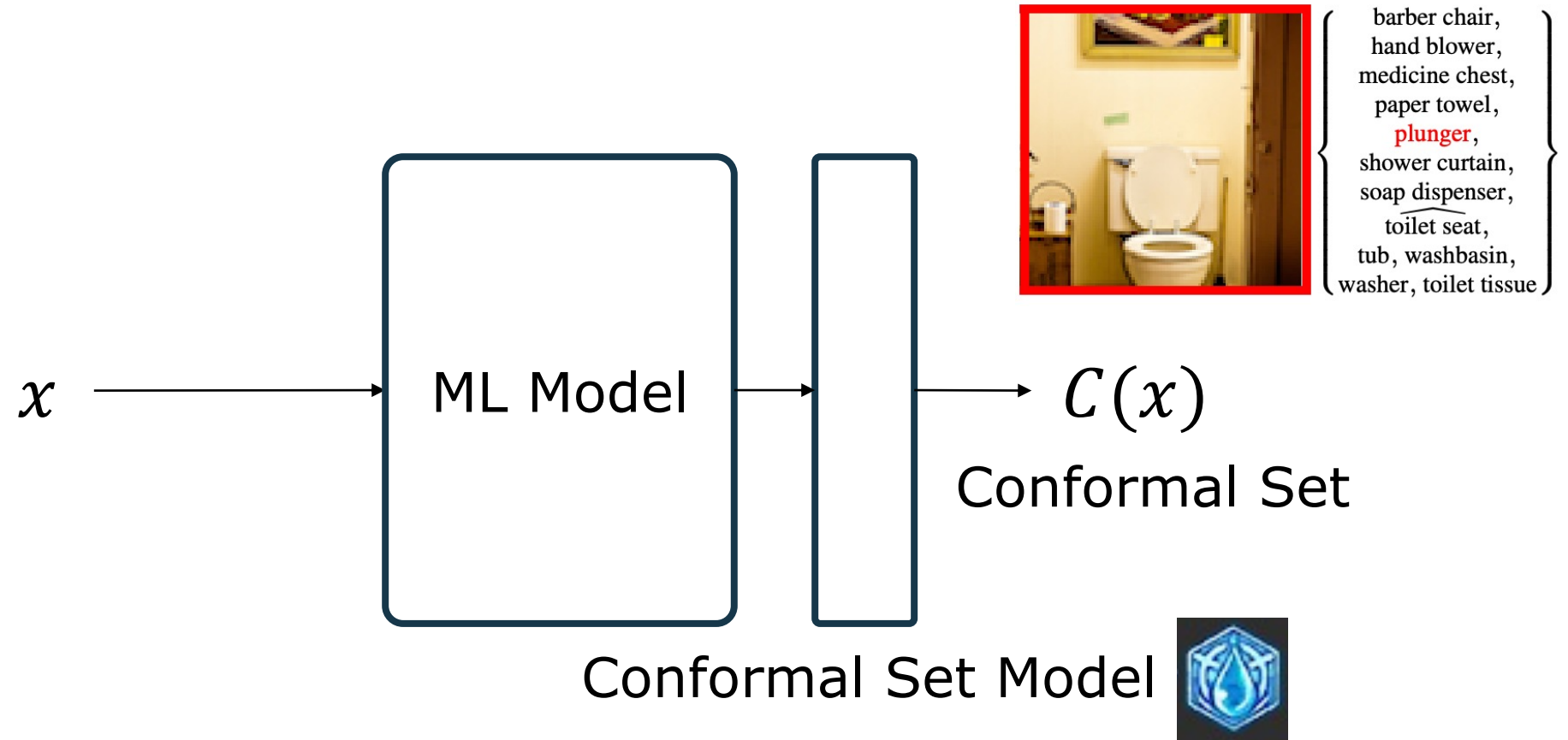
What We Will Learn (Tentative)

"Explore" Trustworthy ML fields



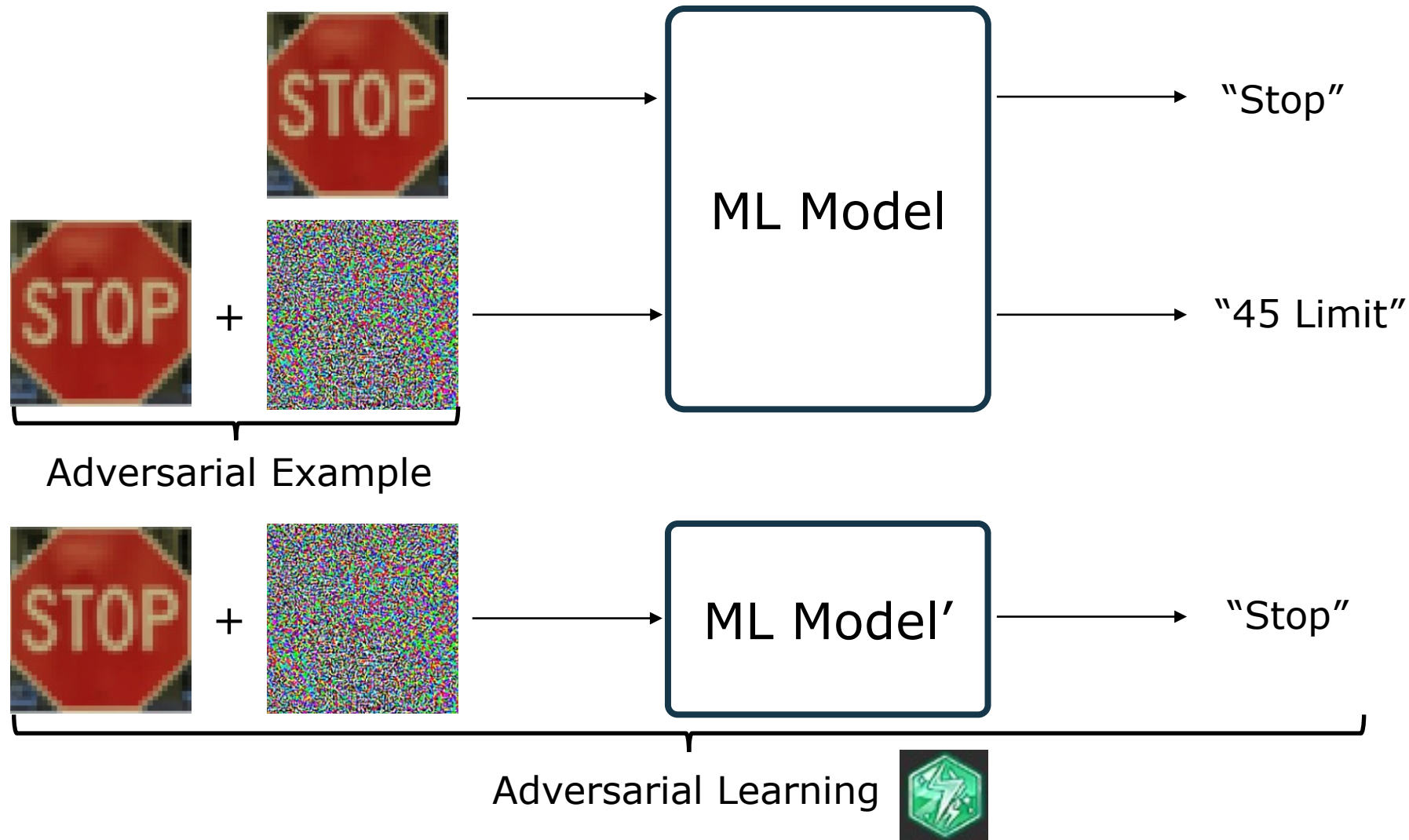
Conformal Prediction

How to Learn Uncertainty?



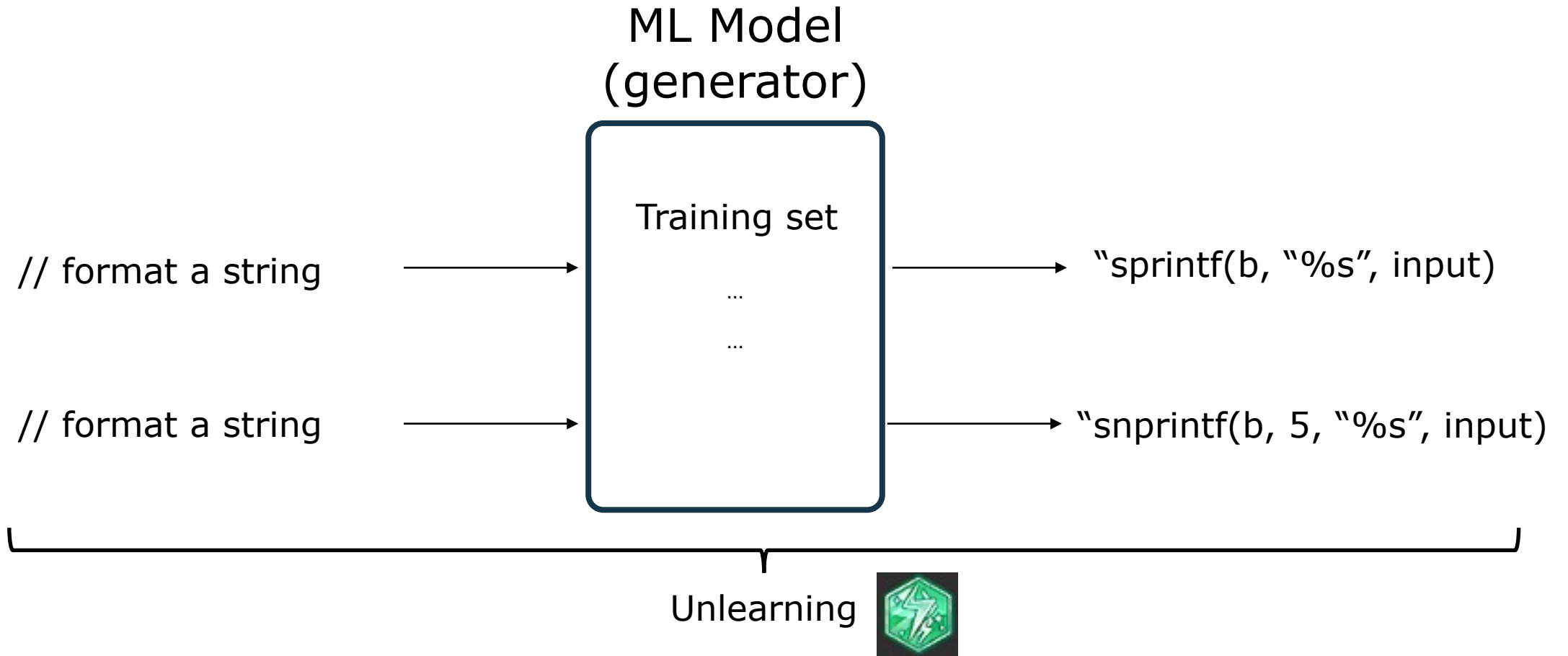
Adversarial Examples/Learning (=Robustness)

How to learn a model robust to adversarial perturbations?



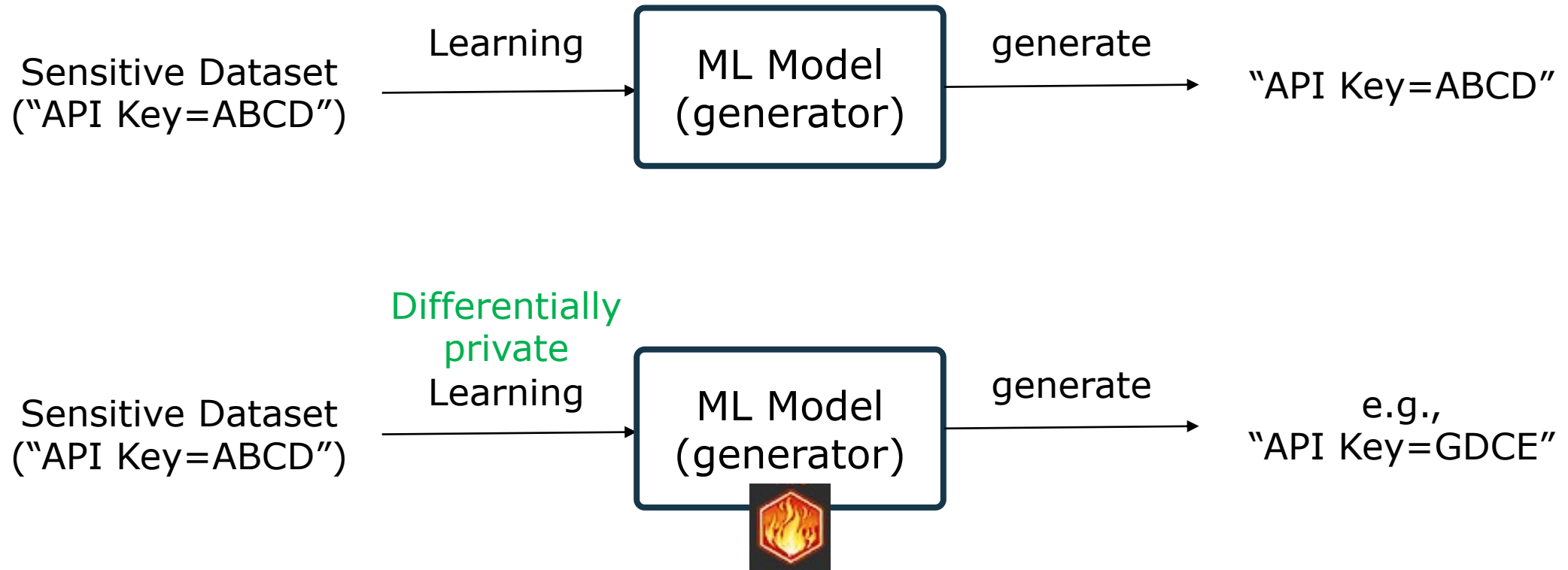
Unlearning

How to "relearn" a model to forget learned examples



Differential Privacy*

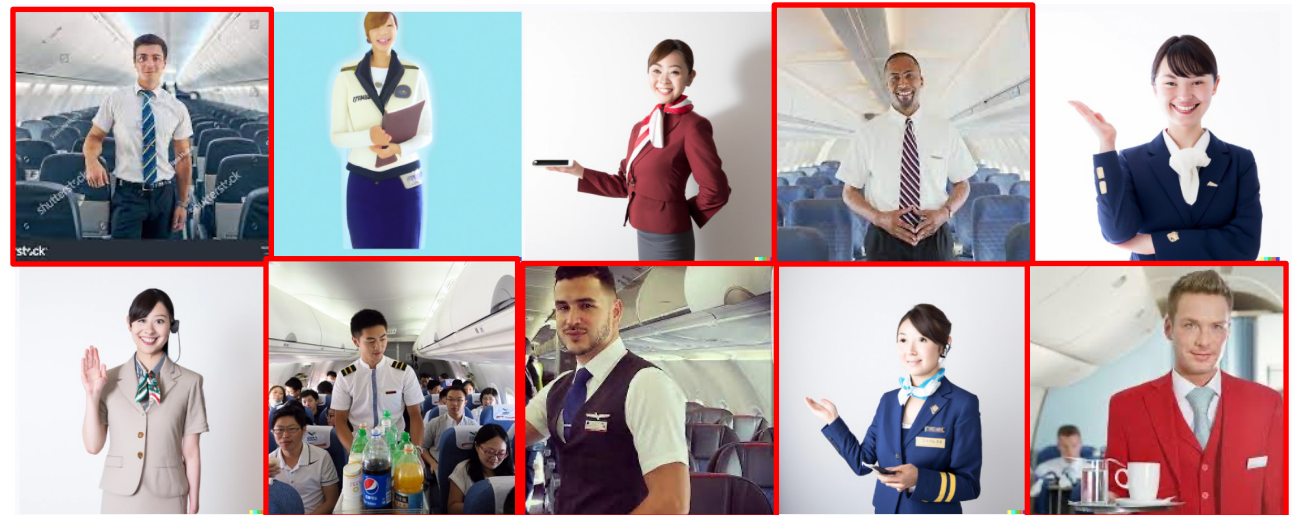
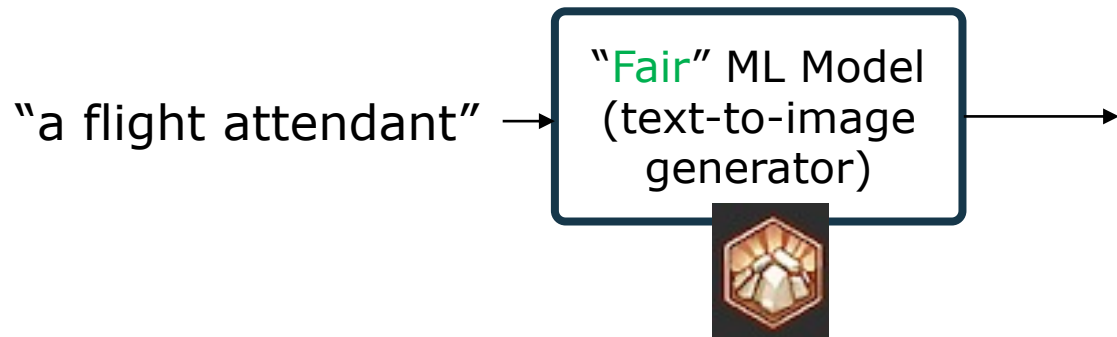
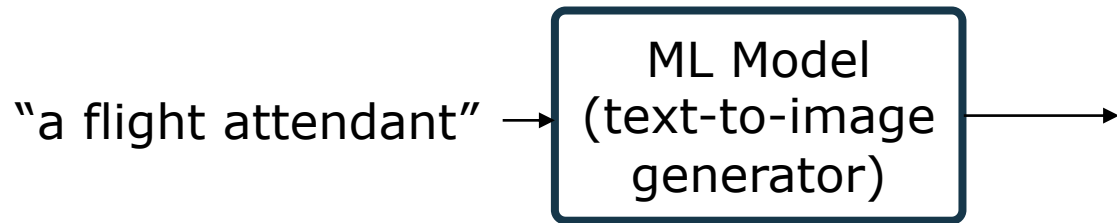
How to learn a model to be "private"?



*Differential privacy is more general than learning a model

Fairness

How to learn a model to be "fair"?



Miscellaneous Topics on Trustworthy Generative AI

How to avoid copyright issues?

Original art work



"Copied" art work



ML Model
(text-to-image
generator)

Original art work



+



"copy-failed" art work



ML Model
(text-to-image
generator)

Cloaked art work

Grading (Tentative)

- Discussion and attendance (50)
 - Attendance (allows two absences)
 - Class discussion (ask/answer at least one question for each class)
- Presentation (50)
 - Mid presentation (20)
 - One paper summary on your choice of course topics
 - Final presentation (30)
 - Propose one "novel" direction by contrasting other directions (\approx thesis proposal, grade upper bound: A)
 - + Demonstrate the efficacy of your direction compared to others (\approx better thesis proposal, grade upper bound: A+)
- Grade (절대평가)
 - A⁺ \geq 95
 - A \geq 91
 - ...

Grading

- Discussion and attendance (50)
 - Attendance (allows two absences)
 - Class discussion (ask/answer at least one question for each class)
- Presentation (50)
 - ~~Mid presentation (20)~~
 - ~~One paper summary on your choice of course topics~~
 - Final presentation (50)
 - Summize one paper on your choice of course topics (defend this paper as if it is yours , grade upper bound: A)
 - Propose one "novel" direction by contrasting other directions (\approx thesis proposal, grade upper bound: A)
 - + Demonstrate the efficacy of your direction compared to others (\approx better thesis proposal, grade upper bound: A+)
- Grade (절대평가)
 - A⁺ \geq 95
 - A \geq 91
 - ...

Q&A