FIT5230 Assignment & Milestones (50%)

Malicious AI

Assignment Description

For this assignment, you will work in **teams of 2** (or 3 with additional requirements), or you may choose to complete it individually. To ensure that every team has the chance to both attack and defend their designed frameworks, each theme will be capped at **14 teams, split evenly between 7 Light (Defense) teams and 7 Dark (Attack) teams**. This balance is essential to maintain fairness and engagement across all themes, so teams are encouraged to organize early to secure their preferred theme and side. The assignment is divided into **4 milestones**, each designed to guide you through the process step by step, from initial planning and framework design to execution, attack/defense preparation, and final presentation. Detailed instructions, deliverables, and deadlines for each milestone will be provided to help you stay on track and successfully meet the assignment objectives.

Please email your tutor with the following details:

1. Team name
2. Team member names
3. Selected theme
4. Chosen side (Light or Dark)

Key Action: *Confirmed teams will be listed in the spreadsheet [here]. Please check the list before emailing to ensure there are available slots for your preferred theme and side. Teams that do not inform the tutor and are not listed on the spreadsheet by the end of Week 4 will be randomly assigned to an available theme and/or side.*

Things to decide before emailing your tutor:

1. Choose a **side**: Light vs Dark
    - Light: your team's aim is to vary/advance existing techniques to fight/counter malicious AI
    - Dark: your team's aim is to vary/advance existing techniques to do malicious AI
2. Choose a **theme** (*You are free to choose other papers and codes apart from the provided samples as long as they belong to the chosen theme*):
    - **Theme 1: Adversarial ML on Gen AI**:
        - Light: How to counter the adversarial ML attack on gen AI?
        - Dark: Is there any other way to replace/improve the current adversarial ML attack?
        - Sample paper(s):
            - Light: WMCodec
            - Dark: AdvSticker
        - Sample code(s):

        - Light: WMCodec
        - Dark: AdvSticker

- **Theme 2: Text-to-Image (TTI)**:
    - Light: any way to counter TTI? e.g. check if something is output from TTI or not?
    - Dark: TTI is a generativeAI model, so it's considered dark (can be exploited by malicious people to generate realistic fakes)
    - Sample paper(s):
        - Light: Safe Latent Diffusion
        - Dark: InstructPix2Pix
    - Sample code(s):
        - Light: Safe Latent Diffusion
        - Dark: InstructPix2Pix

- **Theme 3: Speech-to-Face**:
    - Light: any way to counter speech2face? e.g. check if something is output from speech2face or not?
    - Dark: Speech2face is a generativeAI model, so it's considered dark
    - Sample paper(s):
        - Light: GenVidBench
        - Dark: SadTalker
    - Code(s):
        - Light: GenVidBench
        - Dark: SadTalker / Hallo

3. Choose a team **name**: Light.teamname *e.g. Light.Soothsayers*

4. **Familiarize** yourselves with the selected Colabs/GitHubs
    - Discuss among yourselves (or with yourself if you're doing this individually) how you want to **vary** the techniques in one of those reference Colabs/GitHubs

1st Milestone (2%): Throwing Down the Gauntlet

1. This milestone is a **group-based** assessment task.
2. Each team to **post** to the FIT5230 Ed Discussion forum, the following:
    - Forum heading: "[teamname] topic"
    - Your **team's name** and team members' **names** and **photos** of members
    - Provide the **reference** material (i.e., Colab/GitHub link) and the main research paper your work is based on to serve as a baseline.
    - Briefly explain why the chosen problem is interesting or challenging.
    - Share your team's Colab link (it doesn't need to significantly differ from the reference Colab/GitHub, but it must be your own version).
    - Design and describe an interactive challenge for other teams to attempt. This could involve:
        - Testing Robustness: e.g., "Can you break our defense model?" by bypassing your security measures or finding adversarial inputs.
        - Detection Tasks: e.g., "Can you detect our attacks?" by identifying crafted malicious inputs or hidden patterns in your data.
3. Each student **submits** the Ed post link to the Moodle Milestone 1 submission page.
4. **Milestone 1 due: start of Week 6 (Tuesday 11.55pm AEST)**
5. Criteria for marks:

Criteria	Marks	Description
Clarity of Problem Statement	0.5%	Is the problem well-defined and technically relevant?
Baseline Justification	0.5%	Is there a strong rationale for selecting the reference model or paper, and does it explain why the challenge is interesting or meaningful?
Challenge Design Quality	0.5%	Is the interactive task for other teams thoughtfully crafted, original, and non-trivial?
Initial Customization & Setup	0.4%	Is there evidence of modification beyond just cloning the reference code?
Clarity of the Ed Post	0.1%	Is the Ed post well-written, structured, and easy to follow?

2nd Milestone (8%): Show of Force

1. This milestone is a **group-based** assessment task.
2. **Posts** to the FIT5230 Ed Discussion forum, on either of the following:
    - Main results so far: **changes** you've made to the reference Colab incl some **demos**
    - A **functional base code** that enables other teams to perform attacks or defenses.
    - Which aspect of the other team are you **targeting**, e.g., their ideas or any demonstrations.
3. Additional requirement for a **team of 3**:
    - **Expanded testing and evaluation:** e.g., testing against more than one scenario or providing performance metrics.
    - **Broader target coverage:** e.g., analyzing and addressing multiple aspects of other teams (ideas, demos, and implementation details) rather than just one.
4. Each student **submits** the Ed post link to the Moodle Milestone 2 submission page.
5. **Milestone 2 due: end of Week 8 (Sunday 11.55pm AEST)**
6. **Criteria for marks:**

Criteria	Marks	Description
Technical depth	4%	Are the models, algorithms, or methods applied beyond basic implementation?
Quality of description	2%	Is the problem clearly stated? Are the methods, challenges, and solutions explained logically and concisely?
Documentation and Reproducibility	1%	Marks for well-organized documentation, clear instructions, and making the project easy for others to run.
Engagement with other teams	1%	Did the team have meaningful attacks/defenses? Did they respond to others' ideas thoughtfully?

3rd Milestone (25%) @Week 12: Champions of the World

1. This milestone is a **group-based** assessment task.
2. Teams to post a **10-minute video clip** (*including your facial videos when describing*), to the FIT5230 Ed Discussion forum, describing your main results including a demo of what your team has achieved. Submit your **Colab** as the final report on Ed Discussion forum.
3. Each student **submits** the Ed post link to the Moodle Milestone 3 submission page.
4. Give a **15-minute presentation** during the lab/tutorial session covering the following topics, with all team members taking turns to present:
    - **Showcase** the key outcomes you've achieved so far.
    - Present **ideas** you've tested on other teams' targeted aspects, or experiments based on the Colab/techniques they referenced.
    - Describe the **strategies** used to successfully defend or attack other teams' work (e.g., blindsiding, deception, or keen observation).
    - Explain how the team **collaborated** (or, for solo projects, how the individual effectively managed the workload).
5. **Milestone 3 due: start of Week 12 (Tuesday 11.55pm AEST)**
6. **Criteria for 10-minute video (5%):**

Criteria	Marks	Description
Description of concept	0.5%	Evaluate your understanding of the technical topic your project is based on.
Demonstration of code	2%	Explain the core concepts and methodologies your team employed, providing a clear demonstration of your code's functionality and the results you've achieved.
Analysis of results	1.5%	Critically analyze these results, i.e., discussing the project's strengths, limitations, and potential areas for future work etc.
Quality of video	1%	High video quality with clear audio, visual fidelity, and the inclusion of members' facial videos.

7. **Criteria for 15-minute presentation (8%):**

Criteria	Marks	Description
Description of concept	0.5%	Evaluate your understanding of the technical topic your project is based on.
Novelty of contribution	2%	Assess your team's creativity and initiative. It's about what you contributed that went

		beyond the basic requirements or the standard approaches.
Results presentation	2%	How effectively you showcase your achievements.
Peer Targeting Impact	2%	Discuss the specific strategies you used to successfully defend against or attack other teams' work, detailing how you applied techniques like blindsiding, deception, or keen observation.
Team collaboration	1%	Explain how your team collaborated and managed the workload to achieve these results.
Presentation quality	0.5%	Number and sophistication of interactions with other teams. Was the team a target and how did they defend?

8. **Criteria for Colab submission (12%):**

Criteria	Marks	Description
Organization and structure of Colab	3%	Evaluates the professionalism and clarity of your report.
Functionality of code	3%	Ensure the code is fully functional, well-commented, and runs without errors to demonstrate your project's achievements.
Clarity of methodology	3%	Provide a detailed explanation of your methodology, precisely describing the techniques and experimental setup.
Results and analysis	3%	Present your results with visualizations and an insightful analysis that discusses the significance of your findings, the project's limitations, and potential avenues for future research.

9. **Bonus marks (up to 4%)**
    - The video goes viral in academic/ML circles (more than 50 likes/shares/views).
    - Released as a GitHub repo with community adoption or stars.
    - Inclusion of ethical considerations, e.g., misuse mitigation plan or red teaming report.
    - Successfully impersonated tactics from another team to confuse attribution / Covertly mislead another team into attacking an irrelevant information

4th Milestone (15%): Adversarial Gameplay

1. This milestone is an **individual-based** assessment task.
2. Each student to submit an individual report written with the Overleaf [www.overleaf.com] using the IEEE Transactions journal template: https://www.overleaf.com/latex/templates/ieee-latex-template-for-transactions-on-magnetics/hncvmwqcydfn
3. The report should consist of the following unique forms of adversarial gameplay:
    - Discuss what you would do differently to **enhance** the system if there were no time or resource constraints. Include a proof of concept or demo to show that your idea is feasible.
    - Share your **personal reflection**, including your background, and explain how it influenced your strategies or provided an advantage in tackling the selected theme and coding tasks in this assignment.
    - If you were to switch to the **opposite side** of your current Google Colab project, outline the strategies you could use to attack (if you initially defended) or defend (if you initially attacked). Provide a proof of concept or demo to demonstrate the concept.
4. Each student **submits** the report to the Milestone 4 Moodle submission page.
5. **Milestone 4 due: end of Week 12 (Sunday 11.55pm AEST)**
6. **Criteria for marks:**

Criteria	Marks	Description
Strategic Depth & Feasibility	6%	Describe a specific idea for a new feature, architecture, or mechanism. Provide a functional demo (e.g., code snippet, Colab notebook, video) to show the idea's feasibility. Explain the technical details and why it's a significant improvement.
Personal Reflection & Self-Awareness	1.5%	Share relevant personal or professional experiences. Explain how these experiences influenced your strategies or provided a unique advantage. Reflect on what you learned about yourself and your skills during the project
Adversarial/Defensive Creativity	6%	Outline specific strategies for attacking (if you were a defender) or defending (if you were an attacker). Provide a functional demo (e.g., code snippet, Colab notebook, video) to showcase a key part of your new strategy. Justify why this strategy would be effective, based on your knowledge from the original project.

Analytical Rigor	1.5%	Use data, logical arguments, and project findings to support every assertion. Provide a critical analysis of your work, other teams' approaches, and the system's vulnerabilities. Show a deep understanding by making non-obvious observations and drawing meaningful conclusions.