FIT5230 Assignment & Milestones (50%)

Malicious AI

GIST

- Get into **teams** of 2 (*or 3 if you provide strong justifications, or you're free to do it individually if you'd like to*)
- 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
- Choose a **theme**:
    - AudioSeal:
        - Light: AudioSeal is a countermeasure against generativeAI
        - Dark: adversarial ML attacks on AudioSeal
        - Paper(s):
            - https://doi.org/10.48550/arXiv.2401.17264
        - Code(s):
            - https://github.com/facebookresearch/audioseal
        - Note: *this theme is based on ZiQian's PhD focus, students who have good ideas on how to vary the baselines may post their ideas to the FIT5230 Ed discussion forum or email ZiQian for comments. Note that tutors/PhD students are not expected to help debug code, as it is the students' responsibility to build on/vary the baseline codes as part of their assignment*

    - 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
        - Paper(s):
            - InstructPix2Pix: Learning to Follow Image Editing Instructions
            - Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models
        - Code(s):
            - https://github.com/timothybrooks/instruct-pix2pix
            - https://github.com/ml-research/safe-latent-diffusion

    - Note: *this theme is based on Prof Raphael's student Julia's PhD focus, students who have good ideas on how to vary the baselines may post their ideas to the FIT5230 Ed discussion forum or email Julia for comments.*

  - Speech-to-Face:
    - Light: any way to counter speech2face? e.g. check if something is output from speec2face or not?
    - Dark: Speech2face is a generativeAI model, so it's considered dark
    - Paper(s):
      - SadTalker: Learning Realistic 3D motion coefficients for stylized Audio-Driven Single Image Talking Face Animation
    - Code(s):
      - https://github.com/fudan-generative-vision/hallo
      - https://github.com/OpenTalker/SadTalker
    - Note: *this theme is based on Prof Raphael's student AiFang's PhD focus, students who have good ideas on how to vary the baselines may post their ideas to the FIT5230 Ed discussion forum or email AiFang for comments.*

  - 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?
    - Paper(s):
      - ACE: https://arxiv.org/abs/2303.09962
      - AdvSticker:https://ieeexplore.ieee.org/abstract/document/9779913
    - Code(s):
      - https://github.com/guillaumejs2403/ACE
      - https://github.com/jinyugy21/Adv-Stickers_RHDE
    - Note: *this theme is based on ZiQian's and Prof Raphael's students PeiSze's & YinYin's PhD focus, students who have good ideas on how to vary the baselines may post their ideas to the FIT5230 Ed discussion forum or email ZiQian, PeiSze or YinYin for comments.*

- Choose a team **name**: Light.teamname *e.g. Light.Soothsayers*
- You may feel free to additionally **shortlist** a few more Colabs/GitHubs implementing research papers achieving your team's theme, besides the baselines' codes listed above

e.g.
https://colab.research.google.com/github/smartgeometry-ucl/dl4g/blob/master/gan.ipynb
implements this paper: https://arxiv.org/pdf/1511.06434.pdf

- **Familiarize** yourselves with the shortlisted 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

- Each team to post to the FIT5230 Ed Discussion forum, the following:
    - Forum heading: "[teamname] topic"
    - Your **team's name**
    - Team members' **names** and **photos** of members
    - The **reference** Colab/GitHub & research paper that you will mainly focus on
    - Describe briefly **why** it is an interesting/challenging problem
    - Your team's **Colab**(s) link: no requirement on how different it should be from the reference Colab/GitHub, as long as it's your Colab link
    - Your **challenge** to the other teams: e.g. detect your attacks, break your defences, etc
    - *When*? Between now & end of Week 5 (Sunday 11.55pm AEST)
- *Criteria for marks*:
    - Quality of description (2%)
    - Uniqueness, Creativity, Ambition (*Bonus marks*)

2nd Milestone (8%): Show of Force

- Teams decide when & how many times to give concise updates on progress, via:
    - **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**
        - What aspect of the other team that you are **targeting**, ideas, demos
    - Give brief **presentations** during the lab/tutorial session on the following:
        - **Demos** of main results you've obtained so far
        - Demos of **ideas** you've tried on the other teams' aspects that you're targeting, or on the Colab/techniques that they reference
    - *When*? Between Week 6 to Week 10 (Sunday 11.55pm AEST)
- *Criteria for marks*:
    - Technical depth (5%)
    - Quality of description / presentation (3%)

    - Uniqueness, Creativity, Ambition (*Bonus marks*)
    - How many other teams' aspects for which you have ideas (*Bonus marks*)
    - Whether & number of the other teams which is/are targeting your team's ideas (*Bonus marks*)

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

- Teams to post a **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
- Please also post the video clip to YouTube, & a social media platform of your choice e.g. Instagram, Twitter, FB, ...
- Submit your **Colab** as the final report
- *When*? Friday 11.55pm (AEST) of Week 12
- *Criteria for marks*:
    - Novelty / technical contribution (11%)
    - Ambition (1%)
    - Uniqueness / attractiveness (2%)
    - Quality of description / Clarity of Colab (11%)
    - The number of Likes/Shares/Views your post gets (*Bonus marks*)

4th Milestone (15%): Adversarial Gameplay

- Each student (not team) to submit a 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

- This should concisely describe different strategies the student has used throughout the semester in order to maximize his/her marks & advantage over other students & teams including his/her own teammates

- This part is considered as the flexible section for which the student is free to think of how to contribute to, and to strategize to get maximal marks from. Generally, marks will be awarded, including bonus marks beyond the limit for this 15% component, for the following unique forms of adversarial gameplay:
    - Team based
        - the student's contribution to successful attacks/circumventions/bypassing of other teams' techniques, or the Colab/GitHub/techniques that they reference

        - the student’s contribution to successful deceptions of other teams e.g. seemingly targeting one aspect of the other team when you’re focussing on another, misleading other teams/players
    - Individual based
        - **Blindsides**, incl double blindsides: colluding with any players of other teams without being discovered
        - **Deception**: describe strategies the student has used to disguise his/her intent, fool other players whether in other teams or in his/her own team
        - **Powers of observation**: being able to see through other players’ deceptions, successfully uncovering another player/team’s plots
    - *When*? Sunday 11.55pm (AEST) of Week 12
- *Criteria for marks*:
    - Contribution to team’s gameplaying strategies against other teams (5%)
    - Individual gameplaying strategies (10%)
    - Unique and/or unconventional gameplaying strategies not used by other students *(Bonus marks up to 5%)*
    - Powers of observation against strategies of other students *(Bonus up to 2%)*