Case Studies : Safety 1
Proposed Answers

Case 1 : Modeling Bad Actors in Social Media Platforms

In the programming exercise you completed this week, you worked on the content moderation system for the social media platforms “Catter” and Twitter. In this case study, we expand our exploration of social media platforms to analyze how they can be affected by bad actors’ actions.

Apply the Bad Actors Modeling Strategy to identify and analyze potential harmful actions or negative consequences that could arise from bad actors on social media platforms in general (think about Instagram or TikTok for instance) using the five motivation categories (Money, Politics, Entertainment, Ideas, Self-interest). Consider the following questions :

- ➔ What harmful actions can be taken in each category?
- ➔ How might these actions impact users and the platform?

Proposed answers :

Money :

- **Scam :** Scammers leverage social media platforms as fertile ground for their deceptive tactics aimed at financial gain. They create fake profiles and groups, posing as trustworthy entities to establish credibility. Through appealing posts, they lure users into various schemes, such as fake investment opportunities, prize giveaways, or charity scams. By manipulating emotions and trust, they persuade victims to share personal information or send money, often leading to significant financial losses.
- **Content farms :** Content farms generate revenue through a high-volume, low-quality content approach. By producing vast amounts of inexpensive content optimized for search engines, they attract significant traffic. This traffic, in turn, serves as a valuable commodity for ad placement. Content farms monetize through ad networks, earning revenue based on impressions and clicks. The more content they produce and the higher the traffic, the greater their potential for ad revenue, despite the often lower quality and relevance of the content.
- **Influencers :** Influencers wield significant power on social media platforms, often reaching large audiences. However, this influence can turn harmful when they promote misleading information, endorse harmful products, or engage in cyberbullying. Their actions can manipulate public opinion, exploit vulnerabilities, and amplify harmful trends.
- **Spoofing :** Spoofing, a deceptive practice where bad actors falsify their identities, poses a significant threat on social media platforms. By impersonating trustworthy sources, malicious individuals can spread false information and discord, and manipulate public opinion. This erodes trust in online interactions, undermines the credibility of information, and can lead to real-world consequences. Spoofing thus amplifies the potential for misinformation, division, and harm in the digital realm.

Politics :

- **Hidden propaganda by politics/ activists :** In the realm of social media, political actors and activists tactically orchestrate covert political propaganda to reach diverse audiences. Through sophisticated profiling and algorithmic targeting, they identify susceptible user groups and disseminate tailored narratives. By exploiting the mechanisms in social media platforms, these actors subtly manipulate perceptions and opinions, shaping public discourse to advance their political agendas while maintaining an appearance of legitimacy.

Entertainment :

- **Individual bullying by thrill seekers :** Thrill seekers on social media often engage in bullying by pursuing excitement through harmful behaviors. They exploit the platform's anonymity to taunt, demean, and harass others, finding satisfaction in the reactions they provoke. This behavior can escalate, causing emotional distress and even psychological harm to the victims.

- **Misinformation by trolls :** Trolls engage in the deliberate spread of misinformation on social media by crafting provocative and false content. They often exploit emotional triggers and polarizing topics to amplify their messages. Through fake accounts and coordinated efforts, trolls craft an illusion of credibility, making it challenging for users to discern fact from fiction. This deliberate manipulation of information aims to create confusion, erode trust, and manipulate public discourse for various motives.

Ideas:

- **Extreme ideas sharing by politics :** In leveraging social media, politicians may strategically frame extreme ideologies to appeal to a diverse audience. By using tailored language and emotional triggers, they can connect with different groups of people/ demographic segments. This approach exploits the platform's algorithmic features, potentially deepening societal divisions for short-term gains.

Self-interest :

- **Negative/humiliating reactions to posts for self-promotion :** In the realm of social media, self-interest often drives individuals to employ negative tactics for personal gain. This involves orchestrating degrading reactions to competitors' posts via fake accounts to undermine rivals and enhance one's own standing.

Overall debriefing: Categories of motivation overlap and it is not always clear how to classify some types of actions, but it is not really important, the goal is to identify a range of possible scenarios that could create threats for your system (security) and for your users (safety).

Case 2 : Analyzing Safety Implications of Autonomous Vehicle Software Using STRIDE Strategy

As a software engineer working on the development of an autonomous vehicle system, your task is to analyze potential safety implications using the STRIDE strategy.

Part 1 : For each situation, identify the associated threat from the STRIDE strategy

Part 2 : Provide a technical countermeasure to prevent or mitigate each issue

Answers :

1. **threat : Spoofing**
   a. A malicious actor broadcasts fake satellite signals that override legitimate signals, confusing the receiver, causing the in-car systems to incorrectly position the vehicle.
   b. **countermeasure : Implement secure communication protocols and cryptographic mechanisms to ensure the authenticity and integrity of sensor data. Use redundancy and cross-validation techniques to detect and mitigate the effects of spoofed data.**
2. **threat : Tampering (and elevation of privilege)**
   a. An unauthorized individual gains access to the vehicle's software system and modifies the decision-making algorithms, causing the vehicle to behave unpredictably or dangerously.
   b. **countermeasure : Implement code integrity checks to detect any unauthorized modifications to the software. Apply access control mechanisms to restrict unauthorized access to the software system. Another option could be to have the software audited by an external organization and require security experts to identify vulnerabilities.**
3. **threat : Denial of Service**
   a. A malicious attacker floods the vehicle's communication channels with excessive data or requests, causing the software system to become overwhelmed and unresponsive, potentially leading to a safety-critical failure.

   b. **Countermeasure: Implement input validation and filtering mechanisms to handle malicious or excessive data effectively. Utilize rate limiting and traffic monitoring techniques to detect and mitigate denial-of-service attacks. Design the system to have fail-safe mechanisms that allow the vehicle to operate safely even under abnormal communication conditions**
4. **threat : Information Disclosure**
   a. A cybercriminal intercepts the communication between the autonomous vehicle and the cloud-based control system, gaining access to sensitive information about the vehicle's operations, routes, or passengers.
   b. **Countermeasure: Employ strong encryption mechanisms to protect the confidentiality of communication channels between the vehicle and the control system. Implement secure authentication and access control mechanisms to ensure that only authorized entities can access sensitive information.**
5. **threat : Repudiation**
   a. A passenger claims that the autonomous vehicle caused an accident due to a software malfunction, but there is no way to prove or disprove the claim.
   b. **Countermeasure: Implement comprehensive logging and audit mechanisms in the software system, capturing relevant data such as sensor readings, system states, and user interactions. Ensure that the logs are tamper-proof and securely stored to provide accurate information for forensic analysis in case of incidents.**

Case 3 : Harm Modeling Strategy

The goal of this exercise is to identify and assess potential harms associated with the technologies described in provided scenarios. These scenarios are voluntarily futuristic for practice purposes. The overall goal is to create awareness about the different types of harm technology can cause and make you realize that we tend to underestimate the number or impact of those harms in real life.

Read the provided scenarios, then apply the harm modeling strategy to assess the ethical implications and potential consequences of this technology. In your analysis, don't forget to consider four categories of use: malfunction, misuse/abuse, unintended use and intended use.

You can use the table from the strategy, reproduced below.

First scenario : Affect-Display Textile Garment (Sleeve)

Read the scenario here: Sleeve - VSD Lab (vsdesign.org)

Proposed Answer :

Category	Type of harm	Description of harms
Humans	Physical injury	The display of some emotions might lead to physical altercations
	Emotional or psychological injury	People with unstable feelings or with depression could be excluded because of the public display of their emotions Children could try to generate emotions in others, e.g. through bullying, just to see Sleeve display their emotions Errors in the emotions displayed could lead to misunderstandings and conflicts
Resource allocation	Opportunity loss	Some persons could be denied jobs based on Sleeve's output
	Economic loss	
Human Rights	Dignity loss	People have no way to disprove what Sleeve 'says' about them Sleeve leads to the oversimplification of emotions, displaying only one color when emotions are more complex

	Liberty loss	Conformity might be reinforced towards neutral emotions
	Privacy loss	All emotions become public, it becomes impossible for people to choose which emotions they want to display
	Environmental impact	The large scale manufacturing of Sleeve consumes enormous quantities of materials The garment to make Sleeve is not recyclable and results in huge dumps Buildings are abandoned because of bad mood colors, which results in an increase in construction needs
Social systems	Manipulation	Advertisement systems exploit the colors displayed by Sleeve to lead people to buy more Political actors could use Sleeve to increase their power/influence (e.g. exploiting anger)
	Social detriment	People over-rely on Sleeve and lose their social/emotional intelligence and critical thinking

Second scenario : Smart home technologies

In the era of technological advancements, a suite of interconnected smart home technologies has emerged, promising unparalleled convenience, safety, and comfort. The suite includes smart doorbells, connected refrigerators, adaptive robot vacuum cleaners, automatic lights and blinds and a voice-activated assistant like Google Home.
Emily and Mark are two individuals leading busy lives in a bustling city. Emily is a young tech-savvy professional who relies on smart home technology to streamline her daily routine. She installed a smart doorbell with facial recognition to enhance her home security. Her refrigerator automatically replenishes groceries through online orders, adapting to her tastes. Her robot vacuum cleaner keeps her home spotless with minimal effort, while her voice assistant controls her lighting, entertainment, and even her morning coffee.
On the other hand, Mark is skeptical of these technologies. He is a family-oriented parent juggling work and household responsibilities, who prefers manual control over his home environment and values his privacy. His refrigerator is a traditional one, and he cleans his home the old-fashioned way. He believes in limiting the data that smart devices collect about him and his children.
Over time, these smart home technologies become deeply integrated into society, revolutionizing individual lifestyles as well as the economy and urban planning as data collected from smart devices inform city infrastructure investments.
However, not everyone opts for these technologies. Individuals like Mark, who are concerned about data privacy and security vulnerabilities, choose to stick with traditional, non-connected household items. Others cannot afford to equip their home.

Proposed Answer :

Category	**Type of harm**	**Description of harms**
Humans	Physical injury	A smart doorbell could wrongfully let an aggressor get into the house An electricity blackout could lead to a failure of the door and window systems, imprisoning all house inhabitants inside To satisfy the client's tastes, the smart fridge could recommend unhealthy food habits
	Emotional or psychological injury	A stalker could gain unauthorized access to the devices and their data, e.g. access intimate pictures and post them online, or have the devices perform disturbing actions (e.g. noise)
Resource allocation	Opportunity loss	An insurance company could use the data from the smart devices to base their decisions and prices on daily habits (e.g. diet related to risk of cancer) Having all the devices working smoothly might require a good connection or wifi, people living in remote areas might not be able to benefit from it. The voice assistant might not recognize some accents or some voice types
	Economic loss	A company could adapt their pricing when the smart devices (e.g. smart fridge) order online People who cannot afford the technology see their property lose value

Human Rights	Dignity loss	
	Liberty loss	In case of sanitary crisis (e.g. pandemic), the system could be used to monitor people's movement and apply fines or even restrict movements
	Privacy loss	The system relies on a constant surveillance of private life Very intimate pictures taken by a robot could accidentally leak on the web (ex: Roomba Une femme a été photographiée sur les toilettes par son aspirateur-robot Roomba (futura-sciences.com))
	Environmental impact	Manufacturing the devices at large scale consumes large amounts of resources Daily life becomes even more dependent on electricity The end of life of the devices generates mounting quantities of electronic waste
Social systems	Manipulation	A smart fridge could manipulate people in consuming specific products proposed by partnering companies
	Social detriment	People relying on their smart fridge loose the competences for planning meals and managing their diet

Overall debriefing: Not all types of harm are relevant in a given scenario. Some types of harm might fall into several categories, in which case you simply choose to put it in one. It may sometimes feel that the harms identified are unlikely, but that's ok: in a second step, you would need to assess the severity, scale, probability and frequency of the harms.