LAW ENFORCEMENT CHATBOTS

CASE STUDY: 4

The development of artificial intelligence (AI) systems and their deployment in society gives rise to ethical dilemmas and hard questions. This is one of a set of fictional case studies that are designed to elucidate and prompt discussion about issues in the intersection of AI and Ethics. As educational materials, the case studies were developed out of an interdisciplinary workshop series at Princeton University that began in 2017-18. They are the product of a research collaboration between the University Center for Human Values (UCHV) and the Center for Information Technology Policy (CITP) at Princeton. For more information, see **http://www.aiethics.princeton.edu**

PRINCETON

DIALOGUES ON
AI AND ETHICS

The democratically elected leaders of Panoptica, a medium-sized European nation with a strong welfare state and a liberal legal tradition, recently noticed an uptick in crimes conducted over the internet and through other digital means (i.e., "cybercrimes"). The constitutional guarantees of due process and respect for individual rights had often hindered Panoptica's law enforcement efforts to prevent cybercrimes in the past. Once cybercrimes had been committed, the complexity of the technologies used by cybercriminals and the international nature of their crimes made them extremely difficult to prosecute.

After a high-profile string of identity thefts targeting elderly Panopticans, national police uncovered an international, online marketplace in which cybercriminals advertised and traded stolen identities across state lines. The forum was based on a website on the "dark web" that was not accessible through regular browsers or search engines but instead required special software to engage in real-time conversation and trading. The hidden nature of these forums frustrated law enforcement efforts, and the lack of progress in bringing cybercriminal to justice produced anger among police, politicians and the general population. A public debate eventually erupted concerning the nation's responsibility to protect its citizens from cybercrimes. This debate featured heavily in the parliamentary election with politicians, citizens and the media all demanding that the state take action to protect the nation's most vulnerable members. In the end, voters endorsed the idea of increasing law enforcement capacity and action to combat online identity theft.

With this democratic mandate in hand, Panoptica's law enforcement agencies teamed up with researchers at the University of Panoptica to develop a chatbot that could be used to identify cybercriminals. The chatbot, named JEREMY, worked by using natural language processing to learn how cybercriminals communicate in their online chat forums. JEREMY's machine learning algorithm was then able to use collected data to convincingly engage in conversation with individuals suspected of committing or trading in identity theft. Focusing first on smaller "stepping stone" crimes (e.g. hacking), the system would eventually learn of major upcoming plots. In the process, JEREMY would also be assembling a dossier of evidence (including evidence of intent to commit cybercrimes) that could then be used in any future criminal proceedings. Unencumbered by human time and energy constraints, JEREMY could perform these functions efficiently and at a large scale. Even better, JEREMY was free of the human biases that undermine the fairness of some policework.

JEREMY's developers and institutional supporters truly believed this was the best way to protect citizens against the "bad guys," but they were not naïve to the challenges that might be raised against the use of an automated chatbot in law enforcement. First, in keeping with the nation's preference for strong encryption without any backdoors, they put great pains into making JEREMY as secure as possible. Second, in order to minimize invasions of privacy and dignitary harms to individual citizens, the developers limited JEREMY's use to individuals who were already the subjects of open investigations. Even though developers had the technological capacity to identify virtually anyone at risk of perpetrating identity theft online, they decided that respect for individual liberty should outweigh the public safety benefits of a large-scale rollout.

Almost immediately, JEREMY began to identify online identity thieves, and information from its chats was used to successfully send dozens of perpetrators to prison. The citizens of Panoptica breathed easy, trusting that JEREMY was watching out for their security online.

Discussion Question #1:

Democratic citizens are often asked to choose between liberty and security regarding their government's actions. As a law-abiding citizen of Panoptica, how would you react to the news that your government was deploying a chatbot to protect your cybersecurity? Is it only non-law-abiding citizens that should beware?

An international incident that occurred soon after JEREMY's deployment eventually compromised this national feeling of tranquility. One of the individuals implicated in the first wave of prosecutions made possible by JEREMY was a young man from the neighboring illiberal nation of Hedonia, who was arrested while travelling on vacation to Panoptica. This man had never been a citizen of Panoptica, nor had any of his online dealings taken place while on Panoptican soil, and so he disputed the nation's standing to prosecute him under its laws. In Hedonia, he pointed out, the trading of stolen identities is not explicitly unlawful. This man further alleged that he had been the victim of entrapment by Panoptica law enforcement, as he had not originally intended to sell the identities in his possession. He blamed JEREMY for egging him on and inducing the sale. Transcripts of their conversation show that JEREMY repeatedly offered an increasing amount for the identities possessed by the man. His arrest sparked a diplomatic conflict between the two nations in which Hedonic leaders agreed with its citizen that Panoptica had infringed upon its national sovereignty. Even if a crime had been committed (which they disputed), they argued that JEREMY had unlawfully entrapped their citizen and encouraged him to act when he might not have without such prompting.

Having witnessed cybercrime rates plummet and cybercriminals finally being held accountable for their wrongdoings under JEREMY, Panopticans were generally satisfied with the results of the pilot program. However, with the diplomatic row making national headlines, many citizens of Panoptica also began to wonder whether JEREMY had overstepped. A national debate ensued, in which citizens of Panoptica addressed several concerns to their leaders about JEREMY and its use by law enforcement.

Ethical Dilemma #1: Responsibility in Cases of Entrapment

Most prominently, some citizens shared the concern raised by Hedonia that JEREMY was engaging in wrong and unlawful entrapment. These citizens feared that, rather than stopping crime from occurring, JEREMY was actually enticing potential wrongdoers into committing crimes they otherwise would not have committed. Police have traditionally been allowed to engage in investigative questioning of suspects as long as they are under reasonable suspicion; however, these practices must stop short of coercing suspects into committing crimes. In the case of JEREMY, which initiated conversation and used natural language processing to craft precise responses that occasionally led to uncovering an intent to commit a crime, it was not always clear that the system passed these standards. And in the event that JEREMY was contributing to the likelihood of a crime's being committed, its intervention seemed to detract from the moral responsibility of the criminal who eventually acted.

Ethical Dilemma #2: Accountability

In the case of identity theft, Panopticans were generally willing to cede some of their individual liberties in order to promote security. However, because JEREMY's algorithms needed to remain secret in order to function effectively on the "dark web," citizens were not informed about the system's architecture and programming. Specifically, information about how JEREMY chooses to target one individual for intervention rather than another was not made publicly available. This meant there was no feasible way to alert suspects or offer means of redress to those who felt they had been targeted falsely or unfairly. Citizen groups began to question the choice to employ automated means of law enforcement when this automation implied reduced accountability.

As a democratic nation, the rulers of Panoptica felt obliged to respond to these concerns on the part of its citizens. They assured citizens that individuals would not be targeted for a chatbot intervention unless they were already under criminal investigation initiated by a law enforcement agent. Indeed, investigative units in Panoptica only have information on suspects of crimes, not law-abiding citizens, so they wouldn't be able to deploy JEREMY against the latter even if they were tempted to. Civilians were thus assured that only those with something to fear would be targeted and monitored. However, for obvious reasons, they could not alert individuals if they had been flagged for investigation.

Most importantly, Panoptican legislators noted that there would always be points for human intervention that would serve as a check on automated systems. Human investigators sort individuals into the threat categories from which JEREMY draws its targets. And human judges can use their discretion to evaluate individuals when they are brought before Panoptican courts. It is true, they noted, that judges have a reputation for being notoriously bad at dealing with complicated technologies, but judges do not have to know how JEREMY works to rely on its judgment. The judges can simply read the transcripts!

Discussion Question #2:

In this context, what is the difference between a human undercover agent with trained judgment and an AI tool programmed to act within precise boundaries when executing their tasks? To what extent—and to what end—does JEREMY replace human agents? What are some of the benefits of deploying JEREMY over a human agent? What value does a human agent have over JEREMY for a cybercrime investigation?

These guarantees did much to alleviate the concerns of law-abiding Panopticans; however, some citizens noted that these official statements still failed to address the morally tricky questions surrounding entrapment and did not offer solutions to hold JEREMY accountable for the decisions it did make.

Internationally, many lawmakers couldn't shake the larger political questions raised by the incident between Panoptica and Hedonia. Even if JEREMY did not run afoul of Panoptican law, the standards necessarily become trickier when national law enforcement is investigating citizens of other states – especially states in which the act in question is not a crime. This kind of conflict may be inevitable when law enforcement deploys a chatbot on an internationally accessible forum, but the threat to national sovereignty discomfited national leaders.

The university researchers and law enforcement officials responsible for JEREMY acknowledged the legal difficulties associated with its use and the potential cross-jurisdictional conflicts it raised. Until capable international law enforcement agencies that can competently handle cybercrimes like identity theft exist, however, they believed it to be their duty to pursue cybercriminals using the technological capacities at their disposal. As long as Panoptican servers were used in the commission of a cybercrime, Panoptican law enforcement was willing to claim jurisdiction. When the identities of Panoptican citizens are at stake, they insisted, that interest overrides the sovereignty claims of illiberal national regimes.

Discussion Question #3:

Were the law enforcement actors who deployed JEREMY justified in arresting someone from another jurisdiction based on their interactions with the chatbot? Should national courts of law have the authority to decide to deploy tools such as JEREMY? Does the inevitable cross-border nature of the online forums used by cybercriminals make the use of courts more pressing?

General Discussion Questions

Automation: At the current stage of technological development, AI systems require ongoing support from humans. On the frontend, humans can provide AI with relevant information so that it can improve its algorithms. On the backend, humans are needed to judge how a system's outputs accord with human values and standards. JEREMY's supporters noted that there would always be human oversight to judge its recommended actions and reign it in when necessary. But it is impossible to predict the relationship between humans and AI going forward. As the technologies improve, humans may begin to play smaller and smaller roles in what we think of as essentially human activities, like law enforcement.

- ○ If, in the future, removing humans from the equation would make AI systems more efficient at achieving their ends in law enforcement, do you think that should be allowed? If human judgment produces error and emotion that clouds efficiency, should we extricate ourselves entirely?
- ○ We have already seen the mass automation of industries like manufacturing and mining. Beyond the sociopolitical questions of what happens to workers (and the notion of work) when they are displaced by machines, some have argued that the outsourcing of human tasks to AI poses an inherent ethical dilemma. The blurring of lines between humans and machines serves to anthropomorphize the latter and dehumanize the former. This may undermine our very notion of what it means to be a human being with agency. How would you engage with this position? Consider this question from the perspective of both the AI developer and the law enforcement agent who might be replaced by JEREMY.

Research Ethics: Social scientists have long understood that it is difficult to work with human subjects. As soon as a researcher interacts with humans, they change. Just the fact of being observed influences a subject's choices. Active interventions, such as those performed by JEREMY, can have even more unpredictable impacts on behavior.

- ○ Strict regulations exist for research involving human subjects. The researchers behind JEREMY developed a technical tool to be used on humans by another actor. To what extent would JEREMY have been subject to research ethics regulations in its development phase? Would data collection during the chatbots operation (by law enforcement officials) also be subject to research ethics regulations if this data improved JEREMY's algorithms on an ongoing basis?
- ○ If talking to a chatbot makes an individual more likely to commit a crime, does that individual bear full responsibility for the crime? What is the research team's culpability, if any?

Sovereignty: A democratic state is meant to be responsive to its people. The cross-border nature of the internet and the rise of internet-mediated crime, however, has challenged traditional understandings of who constitutes "the people." These new systems have also changed expectations for governmental action. If governments ignore cybercrimes that affect their citizens, they are seen not to be doing enough to protect their citizens. On the other hand, if national governments take action against foreign citizens, they may be infringing on other nations' sovereignty.

- ○ Short of establishing international agencies to handle criminal activity across borders via digital networks, what can be done to balance the interests of individual citizens and nation-states?
- ○ Within the current international order, different nation-states have different resources for combatting cybercrimes. They also have different priorities. In the event that a nation like Panoptica has the resources and political capital to deploy a system like JEREMY, should it share these technologies with countries that may lack the necessary infrastructure to build such systems on their own? As a citizen of the world, how might you respond if Panoptica were to release the data collected by JEREMY to other law enforcement agencies in order to build an international database for online identity theft and trading?

AI Ethics Themes:

Automation
Research ethics
Sovereignty