DYNAMIC SOUND IDENTIFICATION

CASE STUDY: 2

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**

DIALOGUES ON
AI AND ETHICS

Also known as "query-by-example," dynamic sound recognition recently found commercial success as a means to identify music through short audio snippets, captured through a microphone. First-generation algorithms recognized unique signatures in a particular sound, which they could then match with a most likely source or an equivalent sound stored in a large database of previously identified auditory signatures. Early mobile apps employing these algorithms were amusing and effectively enabled music listeners to identify a song's title and the performing artist. One Los Angeles-based research and development company determined that the underlying technologies might have further, public-minded implications as well, and began exploring new uses for sound recognition algorithms. The most promising output of this research was a mobile app, dubbed Epimetheus.

Epimetheus was particularly proficient at recognizing music, advertisements and human voices. Unlike previous apps using dynamic sound identification, Epimetheus was also adept at picking up subtle auditory signals and sorting through environmental noise in order to accurately identify natural phenomena, such as the changing tides. This functionality was meant to benefit scientific researchers who could employ Epimetheus as a tool to track ecological change in remote locations. It also proved popular among students and casual hobbyists who enjoyed the app's educative and informative capabilities. In addition to identifying sounds with a high degree of accuracy, Epimetheus incorporated a machine learning algorithm that adapts to new inputs and provides users with useful information about the sounds being processed. For example, the app might identify personal information about those speaking, links to websites selling a product being advertised on television, encyclopedic entries about bird calls in the wild and other relevant resources.

It wasn't long before the titans of Silicon Valley recognized Epimetheus' commercial and scientific potential and started bidding to acquire the underlying software. At that point, the research team behind Epimetheus began preparing demos that leveraged the strengths of its sound classification engine. For example, engineers developed an entertaining demo that was able to identify with high accuracy the voice actors/actresses for cartoon characters. It even worked in cases where the cartoon characters were voiced by actors/actresses of the opposite sex (e.g. Bart Simpson is voiced by female voice actress Nancy Cartwright).

One company, Cronus Corp., was especially impressed by these demos, and was eager to acquire Epimetheus and incorporate its sensing technology, databases and information provisions into its own products. However, before negotiations could proceed, Cronus Corp.'s lawyers asked the research team behind Epimetheus to prove that they had minimized the risk of unexpected harmful results. Programming an algorithm that is sensitive to societal norms and cultural flux is notoriously difficult, and Cronus Corp. did not want to unwittingly produce a bad outcome or acquire a public relations scandal.

Discussion Question #1:

Regardless of how much testing is done in the development stage, it is impossible to predict all the potential harms that may occur when an AI system goes live. This is why Cronus Corp. only asked Epimetheus to show they have done as much as reasonably possible to "minimize" the risk of unanticipated harms. But while developers cannot predict everything, they should be able to anticipate common discriminatory harms. What are examples of such harms? What might companies do to minimize these risks?

A problem arose when one of the adversarial testers, Sybel, tested her voice on the system. Sybel, who had been born as a biological male, had recently begun sexual reassignment and now identified both psychologically and publicly as a woman. Based on her voice sample, however, Epimetheus identified Sybel as male and displayed further information about her known history, including a link to several online

videos that showed Sybel prior to her transition. The researchers only then realized the potential for this error to cause substantial dignitary and material harms for transgender individuals. When transgender users of Epimetheus are misidentified, they may feel like they are not being respected for who they are. For those who are "passing" as the gender with which they identify, being publicly identified by their biological sex might even make them targets for abuse. And while this one isolated error may have seemed minor now, researchers notes the potential for a larger, systematic problem. Transgender individuals comprise only a small percentage of the world population, but mass adoption of the Epimetheus app through an enormous technology company, like Cronus Corp., would mean that the algorithm might categorize individuals in ways that did not match their gender identity multiple times per day. As members of a historically marginalized group, this is would be no small thing.

Discussion Question #2:

Even when AI companies aim to be inclusive, limitations on data and human imagination often mean that minority populations lose out. As an engineer at Epimetheus, what might you do to make your products accessible to all? What can you do?

The research team revealed this issue to the acquisition team at Cronus Corp., apologized and promised Sybel a swift resolution of this rather embarrassing issue. However, time was running out for the Epimetheus team to devise a workable solution, lest the negotiators begin looking elsewhere for advanced dynamic sound recognition technologies. Unfortunately, all the usual solutions proved inadequate. Regardless of the amount or type of new data the researchers fed into Epimetheus's training sets, the engineers could only marginally reduce the error rate of categorizing the sex of transgender persons. Even efforts to create focused and auxiliary training data using a significantly diverse set of transgender persons did not yield the necessary results in subsequent tests. The team had to concede that this may not be a problem that can be solved with more data or improved calculations but would require a different strategy entirely.

The researchers at Epimetheus organized several workshops and focus groups with experts from a variety of fields. Participants signed non-disclosure agreements before being invited to critique the approaches and help think through possible solutions. Experts were also asked to help identify any additional red flags or areas for concern. These review sessions produced several findings. Regarding the Sybel problem, some reviewers suggested that the team might want to rethink whether the benefits of using Epimetheus' algorithm on any particular sound would always be worthwhile. Epimetheus' low error rates—calculated at around 0.016% of identified issues—were well within the acceptable range for each interaction. However, given the scale of operations at Cronus Corp., even a tiny rate of error would likely be amplified beyond what the researchers and the interested companies may consider to be negligible levels. In instances where such an error might harm members of already marginalized groups, several reviewers argued that the only acceptable rate of error should be zero.

Discussion Question #3:

Just because a technological capacity exists, does that mean it should be pursued? What factors should companies take into account when determining whether or not to launch a product?

Discussion Question #4:

Epimetheus began as a small company without much market penetration. Being acquired by the much larger Cronus Corp. presents many opportunities, but also raised certain challenges. For one thing, low margins of error that may have seemed fine initially may no longer be acceptable when Epimetheus is scaled up and out. In that case, what would constitute an acceptable margin of error?

While they were at it, the reviewers also alerted the Epimetheus team to several ethical dilemmas they thought the company ought to consider prior to any sort of major expansion or buy-out. While not an exhaustive list, they identified three major areas for concern. These centered around questions of cultural insensitivities, concerns about the act of categorization, itself, and the lack of control over Epimetheus's uses once it had been made publicly available.

Ethical Objection #1: Cultural Insensitivity

Some experts pointed out that Epimetheus' identification of sounds and subsequent labeling had been well trained on American cultural norms and the subtleties of the English language. However, this training would not necessarily translate to non-Anglophone societies and different cultures. While privacy laws had been mapped around the world and were taken into account when processing the sounds, the researchers had not foreseen every culturally specific ethical issue at play. Even within the United States, experts warned that the sensitivity of certain labels and categories are disputed. For example, while some may consider the term "American Indian" to be a valid description of one whose ancestors lived in North America before European settlement, others may consider it offensive. It would be a struggle—perhaps impossible—to develop a categorization schema that did not offend anyone.

Ethical Objection #2: Categorization as Harm

Beyond causing offense when categorization goes wrong, the mere act of sorting people and ideas into groups struck some reviewers as wrong. To categorize a person, idea or thing is to assert a judgment about what they are. Each instance of labeling may be minor, but taken together, programming decisions about how to draw distinctions can influence society's values and cultural understandings of what is good, right and feasible. Where these categorization schemas are used to inform decisions or actions within a larger, more complex information system, the results can be real material harm.

Ethical Objection #3: Unforeseen Uses

Epimetheus' engineers had designed the app to do good. Even if it also produced some harmful side effects, the intention was to create something that would benefit society by increasing knowledge. But what about bad actors who might want to use the technology for more nefarious purposes? The expert panel argued that Epimetheus would need to think more about its moral and legal responsibilities in the event that the sound recognition capabilities were coopted to knowingly inflict harm.

The research team at Epimetheus was glad for the input and advice, but they struggled with how to implement it. They argued that it was technologically impossible to reduce error rates to zero in the case of misidentification of transgender voices—they'd tried!—and so they proposed a quick, though inelegant, compromise. They decided that the least harmful approach would be to delete certain labeling categories that had yielded insulting results for marginalized groups and would continue to do so as a service whenever Epimetheus' technology is included in new applications. This meant that, for example, Epimetheus would no longer differentiate between genders when identifying voices. Such an ad-hoc approach would essentially function as a band-aid solution, though one that might, in fact, do the trick.

Discussion Question #5:

Given that many outcomes cannot be predicted prior to a product's release, how should companies address individual or group harm after it has occurred? Are ad-hoc solutions ever acceptable? What are some alternatives?

Some observers were disappointed with this approach. They argued that deleting that one category didn't address the harm of labeling people in the first place. Furthermore, the act was one of erasure for a community that has fought hard to make themselves seen and heard. They would have preferred for Epimetheus to have committed more efforts into eliminating error rates for transgender voices.

The Epimetheus team defended their ad-hoc approach, while acknowledging that it is far from ideal. In the extreme unlikely situation where an error is also insulting, the engineers decided that it is best to remove the problem, rather than continuing funneling resources into efforts to marginally decrease the likelihood of it occurring further. In the current nascent stages of the development of machine learning approaches, they argued that it is not worth discarding the technology due to growing pains. Rather, ad-hoc solutions should be embraced to allow the technology to mature further.

Indeed, the team went one step further, arguing that the development of new technologies is always going to require a learning curve. Technology companies need room to experiment, and it is impossible to predict with perfect accuracy the challenges a new product will present once it is unleashed in the wild. This is especially true in the case of products that are likely to be used millions or billions of times a day, in which the question may not be whether one can avoid inaccuracies as such inaccuracies are inevitable, but how best to deal with them when problems arise.

In an attempt to address emerging concerns and dilemmas, the Epimetheus team committed to organizing further stakeholder meetings, conducting interdisciplinary research and ensuring diversity of races, genders, ages and socioeconomic backgrounds in development teams. They hoped that these measures would contribute to the ongoing development of products and services that would not only push development forward but would do so in a way to best serve social welfare.

Reflection & Discussion Questions

Rights: The complexity of machine learning technologies makes it difficult to mitigate error rates entirely. Furthermore, it is nearly impossible to judge the true impact of error rates during the testing phase. What appears to be a negligible error rate in advance of a product launch may turn out to have significant consequences once the technology is up-and-running and used millions or billions of times per day by people across the globe. In many cases, the overall utility of these AI systems may outweigh the harms associated with the large-scale effects of low error rates, which can range from the benign to the highly detrimental. However, especially in cases where the consequences of a materialized errors constitute significant dignitary and/or material harms, the rights of those on the losing end of a technology may need to be balanced against the overall utility of the system.

- If technology is imperfect and some error is inevitable, how should companies, product managers, engineers, etc. balance the overall utility of a product with its potential harms to individual and group rights? What are the factors that should be considered in making that determination? Are these decisions that should be made in boardrooms and labs, or should they involve societal input?

- Due to technological limitations, the Epimetheus research team chose to accept low error rates in certain instances and not in others. In particularly sensitive cases, researchers decided it would be best to remove a search category altogether rather than accept some inevitable degree of error, which may cause harm. Which other solutions might have been available during the testing and negotiation phases? Does a "better" solution exist?

Representational harms: Technologies that assess and sort the messy physical and social worlds into predefined or emerging categories contribute towards human understanding and enable us to explore connections that might otherwise have remained invisible given the limits of human intelligence. However, the act of categorization may also inflict harm on the social standing, peer perception and/or self-understanding of some people. To name a person and place her in a particular box is to detract from her individuality and undermine her complexity. When that label contains negative social connotations or harmful political associations, or when it is one with which the individual does not identify, the experienced harm can be especially grievous.

- What responsibilities, both to its users and society in general, does Epimetheus have to protect against labeling that is not only improper but harmful? What role, if any, might diversity in development have to play in reducing such instances?

- Epimetheus is based in the United States, embedded with some American set of values. Virtually any technical decision it makes could be considered a form of moral imperialism, imposing their understanding of the problem onto a population through their technology. How, if at all, can Epimetheus engage meaningfully with this accusation? If you were a manager at Epimetheus, how would you balance conflicting definitional claims and standards between communities?

Neutrality: The process of drawing distinctions is never entirely neutral. By choosing certain categories rather than others, and by defining those categories in particular ways, Epimetheus is implicitly making value judgments about what is good, right and possible. For example, to categorize one sound as "music" and another as "noise" indicates something about what Epimetheus believes both those ideal types represent. These kinds of value judgments may then go on to influence the values of those humans who use Epimetheus, creating a self-reinforcing pattern.

- How should companies like Epimetheus decide which values to promote through its use (or non-use) of particular categorizations?
- Given that a technology can never be perfectly value-neutral, what if Epimetheus were to decide to take a proactive approach towards promoting social values by, for example, designing its categorization function to over-represent the prevalence of female scientists? Is it right or even desirable for private companies to engage in social engineering? What might a "free speech" absolutist have to say about such practices?

Downstream Responsibility: Technologies can be used in a variety of ways, or they may influence others to create similar technologies for other ends. Once a system has left the hands of the original engineers, they may not have much say in how their technologies are used. Sometimes, this means systems that were designed to produce positive social ends get coopted to negative purposes, such as facial recognition software being used by authoritarian regimes to identify and persecute political dissidents.

- Could the researchers at Epimetheus be held responsible when lives are endangered on the basis of erroneous identifications or labeling? Should they be responsible? Think of this in terms of not only legal liability, a well-defined jurisdictional term, but moral culpability as well.
- Can we expect engineers to foresee how the data they create through machine learning inferences may be used in further systems that make decisions about people? Up to which hypothetical data reuse moment should an engineer think ahead? What if the inferred data would be directly useful for authoritarian governments who could justify crackdown on minorities or special interest groups based on erroneously inferred or collected data?

AI Ethics Themes:

Rights
Representational harms
Neutrality
Downstream responsibility