AUTOMATED HEALTHCARE APP

CASE STUDY: 1

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

Type 2 diabetes is a chronic condition in which an individual's body does not produce or use insulin well, resulting in elevated blood glucose (or blood sugar). Over time, high blood glucose levels can lead to heart disease, stroke, blindness, amputations and other serious medical issues. Diabetes currently affects 30.3 million people in the United States (or 9.4% of the population), but it is unequally distributed across demographic groups.[1] Due to a number of sociopolitical factors, it occurs most frequently among low-income individuals, American Indians, Blacks and Hispanics. And while type 2 diabetes is manageable with proper care, treatment can be expensive and onerous. Beyond the mandate to eat well and exercise, individuals with diabetes must test their blood several times per day and self-administer insulin injections based on complex calculations of their current blood glucose levels, the amount (and type) of carbohydrates they expect to consume and projected physical activity. Many diabetics eventually become delinquent in their testing and insulin shots, especially those who also suffer from mental health issues or who lack adequate access to healthcare resources. And because the calculations are complicated and imprecise, even compliant patients often miscalculate their optimal insulin dosages.

Concerned about high rates of type 2 diabetes complications and their concentration among certain socio-economic and racial groups, medical researchers from St. Marcarius-Alexandria University Hospital pledged to make a change. Teaming up with a group of computer scientists, they developed a multi-platform application, named Charlie, which utilizes artificial intelligence technologies to make diabetic care easier, more holistic and more accessible. Taking advantage of smartwatches' biosensors to test blood glucose through the skin, the app's algorithms calculate the optimal level and type of insulin for each user. Individuals still administer their own insulin injections, but Charlie makes the process more efficient and effective. It provides them with a precise dosage calculation, which accounts for current blood glucose levels, as well as lifestyle data (e.g. predicted carbohydrate consumption, exercise levels) and other medical data (e.g. preexisting healthcare conditions).

Charlie distinguishes itself from similar medical devices that use biosensors to test blood glucose in two respects. First, it incorporates a data collection platform. This means that user data—along with anonymized datasets collected by the University Hospital in other experiments—is fed into Charlie's algorithms, enabling them to constantly improve and provide individuals with increasingly more accurate, individualized insulin dosage recommendations. Charlie also uses this data to provide subjects with personalized reminders to exercise, eat well and check their blood glucose levels.

Second, unlike other AI-enabled medical devices, Charlie contains a forum for information sharing and social networking. The developers hoped the forum would serve dual functions. First, by providing a space where users could post, view and discuss emerging and ongoing scientific developments in diabetes research, they wished to counteract what they saw as misinformation about type 2 diabetes—and healthcare in general—on the internet. The forum would be a neutral aggregator of information, and users would be expected to self-regulate. However, to ensure users were kept up-to-date on the latest research, Charlie's developers conceded the use of automated content moderation algorithms (ACMAs) to privilege the most recent information in results. Second, noting the high incidence of depression, anxiety and feelings of isolation among people with diabetes, Charlie's developers thought to design an informal space where users could communicate with one another, thus organically building a network of support. As a bonus, Charlie could use natural language processing techniques to analyze the emerging discourse in order to add contextual data points to individuals' profiles. These analyses could then be used to improve customized treatments.

---

[1] Centers for Disease Control and Prevention (CDC), *The National Diabetes Statistics Report, 2017*. https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf. Note that these figures do not differentiate between type 1 and type 2 diabetes; however, type 2 diabetes accounts for 90-90% of all diabetes cases in the US.

Charlie passed through the university IRB process with little trouble, as it promised substantial improvement in diabetes care with minimal risk to the individual research subjects. Best practices in diabetes care are already well-established in the medical community, and Charlie didn't aim to disrupt them. The technology merely uses existing frameworks to provide more efficient, accurate and personalized care. The review board did note that there was some precedent for assessing the social media platform separately from the medical device itself, but ultimately decided that the forum would not cause undue harm to individual users.

Upon IRB approval, Charlie was rolled out in a clinical trial at the University Hospital to generally positive results. In a survey administered at the end of the trial, users reported liking the regular, individualized, scientifically grounded analyses of their health. Data analysis revealed that those who used the system had higher rates of medical compliance, lower blood glucose levels and improvements in mental health over those in the control group. However, these results were not equally strong for all users. In particular, racial minorities did not experience the same positive results as white users.

Results for Charlie's social networking forum were also mixed. Discussion was frequently lively, but not all users found it productive. Conflicting reports and comments abounded, and would sometimes devolve into hostile arguments, in which users would gang up and call each other 'pseudoscientists' - or worse. Complaints that the negative tone and tenor of conversation was leading some users to disengage with the platform were accompanied by concerns over the emergence of homogenous mini-publics, or echo chambers, that had begun to form around shared scientific and/or philosophical beliefs. This kind of fractured discourse was in direct conflict with the developers' intentions to create a positive, productive space to encourage healthy living amongst those living with diabetes.

Discussion Question #1:

From the outset, the research team was explicitly concerned about inequality in type 2 diabetes care and outcomes. Their aim was not just to improve to improve diabetes treatment in general, but to do so for all. How can they best achieve this? Might different users sometimes require different treatment protocols? Does the research team have a responsibility to single out users who are not receiving the full benefit of Charlie's interventions for special attention and/or treatment?

Discussion Question #2:

Noting the failure of Charlie's social networking platform to achieve its intended goals of sharing information and building a community of support, what should the researchers have done? Arguably, social networking is not central to the purpose of a diabetes health care system. Should they have simply dropped this functionality, or should they have adopted a stronger role in moderating these forums? What are the tradeoffs of each decision?

The research team behind Charlie formed a subgroup of machine learning researchers to address these initial findings. The new "Benevolence Team," as it was named, was tasked to tackle the issues of inequality and speech on the social networking forum, as well as explore other ways of "doing good" with Charlie. After polling users and hosting several open discussion threads about ways to improve Charlie, the researchers decided on several initiatives.

First, in order to ensure that all users benefitted equally from Charlie, they had to determine why some users did not. One possible explanation for the racial gap in outcomes was that the data sets they had been using, which underrepresented racial minorities, were producing skewed results. The team resolved to start using more representative data sets immediately. But they also wanted to know why some users were more compliant with the app than others, and what could be done about it. Using its natural language processing

and machine learning capabilities, Charlie was able to predict which users were least likely to comply with the app's healthcare recommendations by correlating their behavior with that of previous users. The Benevolence Team wished to reach out to these users especially.

One potential solution to the problem of varying compliance was found in Charlie's social networking forum. Upon deeper analysis of the trial's results, a correlation emerged between users who had been presented with certain types of scientific information and noncompliance with the app's healthcare recommendations. For example, users who viewed articles claiming that obesity is not a risk factor for type 2 diabetes were less likely to test their blood glucose when prompted by Charlie than those who were presented with the opposite evidence. A similar correlation was present for those who read articles about disagreements in the scientific community, the theory being that this made them less trusting of medical advice in general. Charlie's research team concluded that they could improve compliance by inundating high risk users with scientific information that tied type 2 diabetes to lifestyle choice and minimizing their exposure to disagreements within the scientific community.

Second, in order to combat hostile discussion and the development of echo chambers, the Benevolence Team would revise the ACMAs on its social media forum. Rather than merely privileging the most recent content, ACMAs would be trained to prioritize information that is most generally accepted or approved by users. At the same time, in order to attract more users to Charlie's social media forums, they would introduce individualized content filtering—based on users' specific circumstances and contexts, inferred from the profiles created through the app's extensive data collection and analysis—to create a more pleasurable, personally relevant experience.

Third, the Benevolence Team realized that they could improve healthcare outcomes for *all* users if they knew which of the common scientifically supported approaches to type 2 diabetes management were most effective for each individual. Without deviating from the canon of type 2 diabetes care, they decided to perform controlled mini-experiments on a subset of Charlie's users, using a "multi-armed bandits" (MAB) approach. Also known as the "explore v. exploit" model, this technique enables researchers to test different treatment options and collect data about which interventions work best under which circumstances. Rather than provide each user with the treatment protocols that appear most appropriate for them given the limited data available at the outset, researchers purposely provide some users with sub-optimal solutions in order to "explore" outcomes and gather additional data to enrich Charlie's algorithms. While the Benevolence Team conceded that this approach may not have been ideal for individual research subjects in each instance, they justified the experiments by explaining that they would increase understanding about healthcare protocols and produces better outcomes in the long-run.

Discussion Question #3:

The Benevolence Team's three major initiatives—improving the representativeness of datasets and targeting at-risk users; moderating discussion; and experimenting on users to find best practices—were aimed at improving the experience for Charlie's users. But with each change, Charlie also came to play a more active role in nudging users towards particular behaviors. In what ways, if any, does Charlie's shift from aggregator/provider of information to curator/nudger impact the issue of "informed consent"?

Discussion Question #4:

AI has created opportunities for highly personalized medical care. But in order to provide such care, apps like Charlie need to collect lots of data about their users. Many users are happy to share their data for the purposes of improving healthcare but worry about how sensitive information will be used in the future. What responsibilities, if any, does Charlie have to be transparent about their data collection, use and sharing policies?

The new version of Charlie was released to the previous research subjects as part of a follow-up study. Subjects were not explicitly informed about the particular changes instituted by the Benevolence Team, and simply went about using the app as they had before. Initial results were encouraging. Healthcare metrics improved over the previous iteration of Charlie, and these results were concentrated among minority and at-risk users. Research subjects, themselves, noted that the new version of Charlie had reduced the "cacophony of voices" on the social networking forum. While the free-form discussion sometimes did still turn hostile, the issue was at least somewhat alleviated by the revisions to Charlie's ACMAs, which simultaneously favored generally accepted advice and provided individualized content. This led to a boom of activity on the forum.

This positive mood changed, however, when the research team began to publish their results and methodologies. Some users and critical observers began to complain, citing concerns about censorship, the singling out of at-risk users for special treatment and the use of the MAB approach. Debates sprang up around these several ethical issues.

Ethical Dilemma #1: Paternalism

Civil libertarians and patients' rights advocates claimed that the Benevolence Team was unjustified in taking paternalistic action against research subjects without their knowledge or explicit consent. They cited both Charlie's use of data to determine at-risk users as well as its policy of curating content to achieve particular outcomes as examples of overreaching. This kind of active interference in the name of promoting an individual's welfare, they argued, was inappropriate for an experimental research platform – especially one tied to essential healthcare services. They claimed that each user should be able to engage with all relevant information and decide for themselves what to believe and act upon. In other words, rather than nudging users towards particular attitudes or behaviors, Charlie should enable users to "pursue their own good in their own way" – even if this means they might make suboptimal choices.

Ethical Dilemma #2: Individual Liberty v. Social Welfare

Some research subjects reported feeling manipulated by the new version of Charlie—especially those who suspected they may have been targeted for more active interventions—and they argued that these tactics had changed the purpose of the forum away from the original stated intent. Those users who believed they may have received sub-optimal healthcare recommendations as part of the MAB experiments were particularly aggrieved. While such an approach may increase social welfare over time, individuals were unhappy about having been treated as means, no matter how laudable the end may be.

Ethical Dilemma #3: Consent and Transparency

The idea that treatments had occurred without the explicit consent of Charlie's users eroded trust in the system. In order to restore trust and provide opportunities for true consent, users demanded transparency. They wanted insight into how Charlie's algorithms worked to construct detailed individual profiles, how it was determined which advice was presented to individual users, and how the algorithms decided to offer sub-optimal solutions to persons. The algorithms are an integral part to the Benevolence Team's scientific method, and yet no information about them had been published in the methodology sections of resulting papers.

Charlie's researchers convened a small conference to respond to the ethical concerns raised by their research subjects, which was live-streamed and later made accessible in Charlie's forums. The researchers began by refuting the claim that they were wrong to offer their research subjects healthcare recommendations by automated means. They argued that the positive initial response in their forums and surveys clearly showed that most users were happy to receive personalized and scientifically supported suggestions to improve their lives, based on their individual health profiles. Furthermore, this strategy was highly effective at improving healthcare results. As long as the technological capacity exists, Charlie's researchers believed they had a moral responsibility to act in response to such user demand, especially given their function as a healthcare app and their position of knowledge about their users.

The researchers also disputed the value of being transparent with machine learning algorithms. First, they claimed that they, themselves, could not decipher or reverse-engineer how specific algorithms operate after having been trained by vast datasets and deployed. The value of these algorithms rests in their ability to adapt themselves to situations and contexts in ways that are superior to human understanding. Algorithmic transparency would, therefore, neither contribute to user understanding, nor enable research subjects to suggest meaningful changes to how the platform or research study operates.

Finally, the researchers addressed the concern about their MAB experiments. They argued that it was obviously more beneficial to optimize outcomes for all platform users, even if that meant that some users were provided with suboptimal advice in the short-term. Since the experiments were designed so that they only tested approaches that were well-established in Diabetes healthcare protocols, the research subjects were never really at risk, whereas the potential benefit was enormous. Furthermore, while research subjects had not explicitly consented to participating in this experiment, if they were to read to fine print on their initial release forms, they would see that they had actually given Charlie's researchers wide latitude to include them in experiments on the platform.

Discussion Question #5:

Is the team behind Charlie right to introduce features based on what their users want? How can they be sure what that is? As a healthcare company, are there other values they should consider in their design choices besides appealing to consumers?

Discussion Question #6:

One way to think about algorithmic transparency in this case is to compare Charlie with a human doctor. Patients don't need to know how a human doctor's brain work or how its decisions were made in order to feel comfortable following them. Is an AI system any different? What roles might trust or fear play?

Discussion Question #7:

MAB is a common technique to improve the overall accuracy of AI systems. Which factors should be included in the decision about the final implementation of the MAB algorithm on research subjects, or whether such an approach should be implemented at all?

General Discussion Questions

Foundations of Legitimacy: The research team behind Charlie claims that the app responds to user/consumer demand for innovative features. In this case, many users had explicitly expressed approval of the app's individualized approach to healthcare through polling and discussion threads with the research team. Furthermore, they argued that users' complaints about the forum's atmosphere amounted to an implicit demand for Charlie's researchers to take proactive measures to regulate discussion. If users hadn't seemed to want it, they wouldn't have done it.

- To what extent do you think consumer sovereignty (i.e., the belief that consumer demand justifies particular actions) is a relevant and useful driving force for the implementation (or prevention) of innovations in machine learning technologies? Is the satisfaction of individual user demand a sufficient source of legitimacy to justify particular design choices or the dissatisfaction suffice to delegitimize?
- Beyond consumer sovereignty, which additional values ought to be considered when deciding which technical features to implement? Democratic self-governance? Utilitarian efficiency? Humanitarian or rights concerns? Others?
- This app is not yet available for public consumption, but a general rollout may be on the horizon. In that event, the academic researchers currently making decisions about Charlie would be replaced by a privately held, for-profit company. Would that change the role that user demand should play in influencing more meta-level decision-making about which ends to pursue and how? What would be a legitimate mechanism to assess consumer demands, as well as inform them of the benefits and harms?

Paternalism: Charlie's research team recognizes that it is in a unique position to provide support of various kinds to users. As an issue of public health, those behind Charlie sincerely wish to do well by the people who experience the world through their platform. Sometimes this means simply monitoring their behaviors, and sometimes it means proactive engagement. While users may ultimately benefit from paternalistic practices, many worry that Charlie is projecting values and life choices that they, themselves, may not necessarily endorse. In doing so, they argue, Charlie's research team is violating their individual liberty.

- Do intentions matter? Or are the results all that matter? Would Charlie's experimental methods and data collection practices be judged differently if the researchers' stated aim was something other than "doing good" (e.g. profitmaking)?
- How should Charlie strike a balance between respecting individual freedom and difference, and pursuing socially advantageous ends?

Transparency: Calls for transparency often take the form of open access codes or Terms of Service agreements, but there is also a richer notion of transparency that involves sharing the ends, means and thought processes behind an engineering decision with those who will be affected by it. Transparency, in this sense, may not always mean that users get to have a final say in company policy, but it does invest them in the process.

- Can and should Charlie's ends and means be made transparent to individual users? In your response, consider both the narrower and the richer definitions of transparency.
- When platforms engage in particularly personal spheres of action, such as healthcare, do they take on special responsibilities of openness and transparency?

Censorship: Charlie's automated content moderation algorithms were originally designed to put the newest information at the top of feeds. When that proved suboptimal, developers decided to privilege generally accepted information instead. But as this standard might still contain a great deal of variety, they also introduced revisions to their ACMAs that would provide more personalized results.

- What challenges are associated with ACMAs that privilege content according to these various values: newness, general acceptance, individualization? What other values might be considered when designing ACMAs to sort through vast quantities of information for users?
- When a social media forum hosts information, everything it features is given an implicit (if, in many cases, weak) endorsement. Because Charlie is a Diabetes healthcare device, users are likely to assume that information that makes its way onto the forums is more accurate than information that does not. Where healthcare questions remain unsettled, what responsibility does the team behind Charlie have to present fair, accurate and/or representative content?

Inequality: Type 2 diabetes is a disease whose impact is felt disproportionately by racial minorities and those on the lower end of the socio-economic spectrum. The developers of Charlie truly wanted to address these underserved populations, and one of the Benevolence Team's first actions was to start using more representative datasets to assure that all users receiving equally good advice and treatment. However, there are also more subtle ways in which Charlie may have been reproducing inequality. For example, the language the app and its developers use is reflective of the racial and class stratification in healthcare. Terms like "compliance" and "at-risk" may make some users feel as if they're being targeted as a threat. Rather than being treated as patients who are in need of care, they are inadvertently framed as problems that need to be solved. This represents not only a dignitary harm, but in practical terms, it may put some users off from the app.

- The language around healthcare is contested and inflected with the industry's history regarding discrimination and dehumanization of certain subject-patients. How might new apps like Charlie be a source of change? Is there better language they could be using?
- How can apps like Charlie avoid causing these kinds of inadvertent but meaningful harms? Is increasing diversity in the development phases—not just of research subjects, but the researchers themselves—sufficient? What are the downsides?

AI Ethics Themes:

Foundations of legitimacy
Paternalism
Transparency
Censorship
Inequality