Technology

GiveDirectly's approach to responsible AI/ML

July 8, 2024 by Vera Lummis, Stella Luk and Swathi Ramprasad

Contents

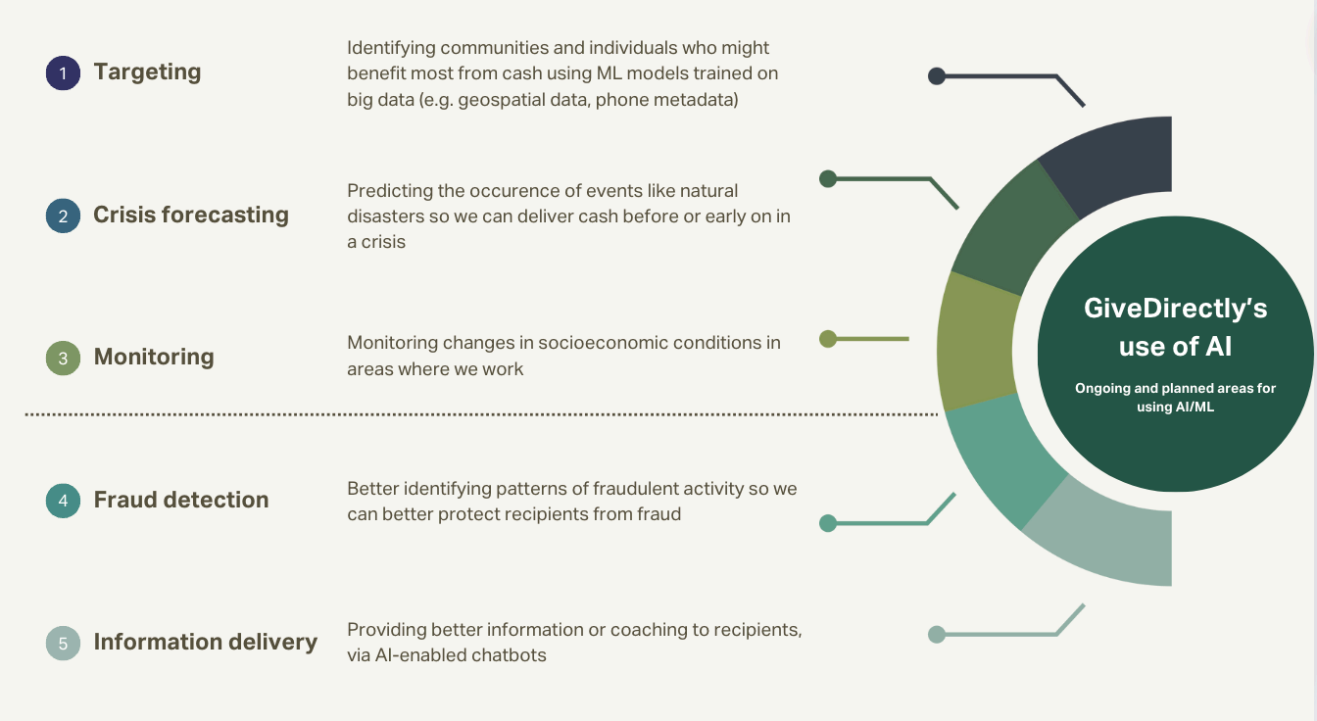
Summary

- Over the last five years, GiveDirectly has used AI/ML technology to improve our programs delivering cash aid to people living in poverty.
- AI/ML models come with risk, so we've developed a Responsible AI/ML Framework to illustrate how we are using AI/ML and share our current principles and guardrails.
- We hope this framework will encourage organisations to create guardrails for how they deploy AI/ML and collaborate with us to develop best practices.

GiveDirectly team conducting recipient consultations in Lilongwe Rural, Malawi

First - How does GiveDirectly use AI?

Over the past few years, we've worked with best-in-class researchers to test how big data and machine learning can help us overcome data gaps and better understand the geographies we work in. As of 2024, we are working on applying AI/ML models in three main ways: to identify communities and people who might benefit most from cash transfers, to respond early to natural disasters, and to monitor changes in socioeconomic conditions in places we work. We are also scoping opportunities to use AI/ML to better detect and mitigate fraud, and provide life-improving information and coaching to recipients via AI-enabled chatbots.



1. Targeting - Identifying communities and people who might benefit most from cash

- **At the community level**: We are using geospatial AI models to more accurately and precisely identify which communities and households are most impacted by extreme poverty, climate change and natural disasters. We have done this in partnership with experts like the Center for Effective Global Action, Google Research, AtlasAI, Fathom, and Floodbase.
- **At the individual level**: In crisis contexts or regions without up-to-date social registries, it is challenging to figure out how to distribute limited resources to those most in need. In several programs, we're working with telecommunications companies and external researchers to test how evidence-backed, open-sourced ML methods can be used to identify phone users likely to be in poverty. This approach has enabled GiveDirectly and governments to remotely identify and send instant digital payments to individuals in extreme poverty in challenging contexts like the pandemic. We're piloting ways to use phone metadata and AI/ML alongside in-person needs assessments to more efficiently identify and pay people at scale, without leaving people behind.


Cures, a GiveDirectly recipient in Bomet, Kenya

2. Crisis forecasting - Enabling cash assistance before or early on in a crisis

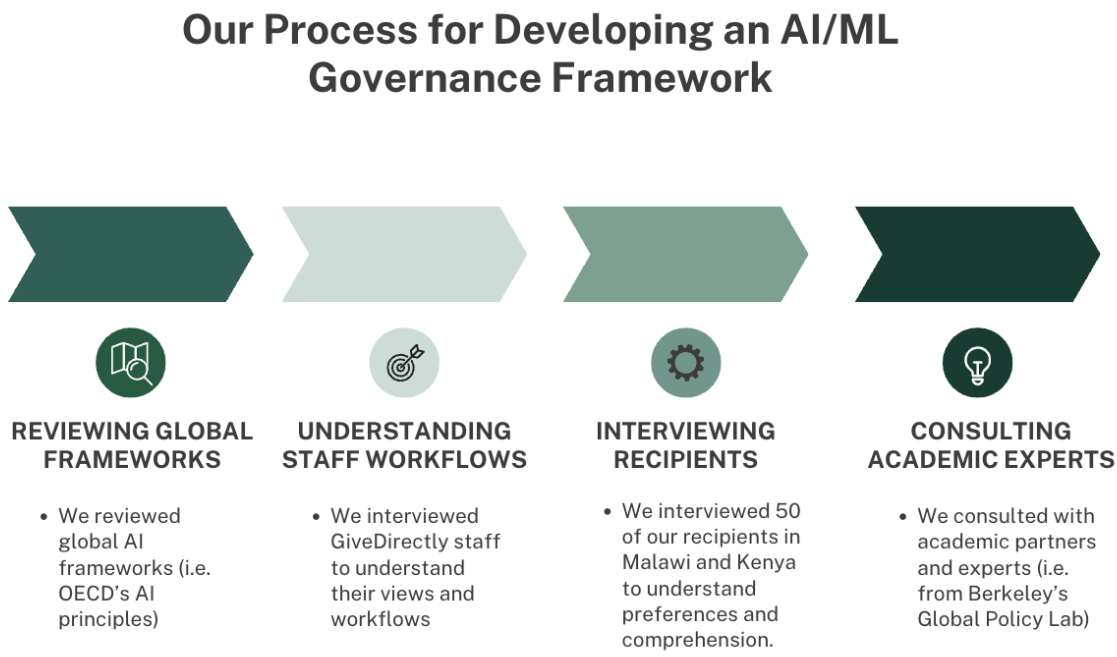
- Many of the regions we operate in are vulnerable to natural disasters such as floods. Our recipients stand to benefit if we can reach them with payments before floods strike, to help them better protect themselves and their assets. We use a combination of local hydrometeorological early warning systems and AI models - through partners like Google Research - to predict where floods might hit, so we can pay people in advance of those disasters. This can improve their resiliency and ability to protect themselves and their livelihoods.

3. Monitoring and learning - Adapting to changes in socioeconomic conditions

- Remote sensing data has transformed capabilities to monitor changes in socioeconomic conditions cheaply and at-scale. We are building a tool with AtlasAI that can help us monitor changes in geospatial indicators of wealth (e.g. infrastructure, farming productivity) in communities where we deliver large-scale lump-sum transfers. Efforts like these will improve our understanding of whether our programs are contributing to expected impacts.

Our process for developing AI/ML principles and protocols

Crafting a framework for responsible use of AI/ML involved a thorough review of our past learnings, industry standards, and consultations with our internal teams, recipients, and external stakeholders.



Our approach began with an in-depth examination of global standards, such as the OECD's AI principles and the United States' AI Bill of Rights, as well as regulations like the General Data Protection Regulation (GDPR), from which we developed a core set of guiding principles. We then reviewed learnings from past programs and spoke with core teams (e.g. our Programs, Safeguarding, Data and Research teams) to understand their workflows, and together defined protocols to put these principles into practice. Interviewing 50 recipients in our programs in Malawi (Lilongwe Rural) and Kenya (Mathare, Nairobi, and Kilifi County) helped us improve our protocols related to transparency and explainability, and refine our approach to consulting communities on their perspectives and concerns on the use of AI/ML. Additionally, we sought expertise from academic scholars such as Zoe Kahn, a PhD Candidate at the UC Berkeley School of Information who studies how AI/ML systems may result in unanticipated dynamics, including harms to people and society. These conversations enriched our framework, which is described in detail below.

Our five core principles and how we're defining them

1. **Transparency & Explainability**
   We communicate about the design, goals, and data included in AI/ML models in contextualized, understandable ways and take into account community preferences prior to implementation. We also transparently document our models internally and support open-sourcing where feasible.

2. **Recipient Consent**
   Our recipients should be able to make informed choices about whether they consent to the use of AI/ML to assess their eligibility. Where possible, we offer alternative methods of assessment.

3. **Privacy**
   We protect and securely store data used to train or validate AI/ML models, in line with our policies and international best practice.

4. **Inclusion**
   We thoughtfully consider whether our AI/ML models represent and produce equitable outcomes for members of different communities before deployment.

5. **Rigor & Relevance**
   Models we use are rigorous (performing in line with the academic literature and robust technical standards), and relevant (context appropriate and used in the absence of better local, available data sources or in synergy with the same).

Principle #1: Transparency & Explainability

When using AI/ML for determining eligibility for cash assistance, we aim for the design, goals, and data to be transparent and clearly communicated to those impacted.[1] This can be challenging, since we often work with communities that have low rates of literacy and digital capacity. Therefore, we will consult with local communities to use AI/ML appropriately to ensure it is explained in a contextually appropriate way. We will also transparently communicate with relevant staff and external stakeholders about the key drivers and accuracy of our AI/ML models compared to other approaches (where data is available), enabling these audiences to effectively participate in decision-making related to their use.

Here are some of the key protocols we developed to ensure transparency and explainability:

- **Community consultation** - In programs where using phone metadata and AI/ML can help us prioritize which individuals should receive assistance more quickly and cost-effectively, we will consult with target communities before implementing an AI/ML-based approach to eligibility assessment. Through interviewing community members, we will gain a better understanding of how we can best deploy an AI/ML-based approach and develop messaging strategies that are understandable and locally appropriate.
- **Transparent documentation** - Transparent documentation is essential for helping our staff understand how AI/ML models work and are used. For all the models we build, we will describe the type of data that goes into building it, key ethical considerations, and how well the model performs (e.g. accuracy and fairness) and make this information available to relevant staff. Where feasible, we also support efforts to open-source models we use that have been validated by research, like we did in partnership with the University of Berkeley

[Case Study] Community consultation to promote transparency & explainability

In February 2024, we interviewed 50 adults from our programs in rural Malawi (Lilongwe Rural) and Kenya (Mathare, Nairobi and Kilifi County). We sought to understand their preferences, comprehension and concerns related to GiveDirectly assessing phone users' eligibility for cash assistance by using their phone metadata to estimate their spending power, in settings where this approach would enable us to pay more people faster. We explained to interviewees that researchers can study how people in a community use their phones and create ways to instantly predict a person's spending power and location based on their phone usage, especially because research has shown that wealthier people tend to use their phones differently.


Recipient consultations in Lilongwe Rural, Malawi

In gathering feedback, we learned that many recipients, especially in urban areas, prefer instant assessments of eligibility and enrollment on their phones if targeting is accurate and cash assistance is received faster. Almost all interviewees trusted GiveDirectly accessing phone metadata from telecommunications companies, given established trust, but many were concerned about the risk of this data being shared with community members who might cause them harm. Those less familiar with phones, including more women and elderly recipients in rural areas, needed more support to understand how phone usage patterns predict eligibility for programs.

[1] As a hypothetical example, we can describe the intuition behind how different patterns of phone usage might relate to a person's spending power, and give examples of the types of variables that are significant.

Principle #2: Recipient Consent

Our recipients should be able to make informed choices about whether they consent to the use of their data and an AI/ML approach to assess their eligibility. We will mitigate the risk of 'murky consent' by clearly communicating with potential participants (via community meetings, our call center, IVR, etc.) about how and why their data would be used, and their data rights (in line with our data privacy policies) when sharing opportunities to enroll. We will also provide alternatives to AI/ML-based approaches to eligibility assessment where feasible. For example, in our post-pandemic programs in Malawi and Bangladesh in 2022-2023, where we tested using phone metadata and AI/ML to identify and deliver cash transfers to adults below the poverty line, we offered all potential participants the opportunity to be surveyed and enrolled in-person rather than digitally. We also make recipients aware that they can speak with our field staff or our call center to make requests related to how their data is used, stored, or deleted.

Principle #3: Privacy

We protect and securely store data used to train or validate AI/ML models, in line with our data privacy policies. We also ensure that any data collected for training or validating AI/ML models are kept private, securely stored and used only for intended purposes.

Principle #4: Inclusion

We are mindful of the ways AI/ML algorithms can disproportionately impact particular communities based on categories such as ethnicity, gender, or other sensitive sub-groups, and evaluate the demographic parity of their impacts before deployment. In addition to assessing the accuracy of our models, we also ensure we understand the context and composition of the communities we serve and check that model results are equitable.

Here are some of the key protocols we developed to ensure inclusion:

- **Definition of sensitive subgroups -** Before we operate in a new area, we will take steps to understand community dynamics (e.g. through field scoping visits and a 'community intelligence' gathering process, led by our Safeguarding and Program teams) and work with community members to define vulnerable sub-groups.
- **Bias and fairness audits** - We will run checks to identify if an AI/ML model we have developed systematically or incorrectly excludes (or includes) people from any particular sub-group (i.e. ethnicity or tribal group) outcomes against any particular sub-group. In cases where we observe biased results, we will seek to rectify the model. If rectification still does not produce systematically fair outcomes, we will not use the model altogether.

[Case Study] Bias checks to ensure inclusion

In 2022, we implemented a pilot in rural Malawi to test the accuracy, cost and speed of various methods for identifying and enrolling adults living under the extreme poverty line into our programs. We developed a proxy-means-test model from household consumption surveys (considered the 'gold standard' for estimating household consumption) and an ML model trained using phone metadata and consumption surveys. Our checks found that the ML model ('cdr' charts in below figures) did not cause women or individuals of certain age groups to be more likely to be incorrectly excluded from the program relative to the proxy-means test model. In fact, the proxy-means-test model indicated more biased results. However, neither targeting approach achieved perfect demographic parity.

Principle #5: Rigor & Relevance

We will only use AI/ML models that are high-quality (performing in line with the academic literature and robust technical standards), and relevant (context appropriate and used in the absence of better local data sources). In particular, we aim to only use AI/ML models when they provide added value in comparison to other available data sources (e.g. more accurate, cost-effective, scalable or instantaneous data).

Here are some of the key protocols we developed to ensure rigor and relevance:

- **Benchmarking and assessment** - Before we use an AI/ML model for program decision-making, we will benchmark our models' results or assess them against available alternatives, and only use AI/ML models when they perform better than alternative methods on metrics like accuracy, speed, and cost. This helps us only use models when there is true value add.
- **Evaluation of third-party models** - Prior to procuring third-party AI/ML models, our team will vet external vendors to ensure that their models are of sufficient quality in contexts in which we operate, including by checking model performance against ground truth data where available. This review will examine the academic reputation of the vendor, compliance with data privacy and security best practices, model performance (across factors like accuracy or bias), and the quality and recency of training data utilized.

- **Validation with local data and human review** - Sometimes, AI/ML does not output the nuance that a human or local knowledge will. Before we use an AI/ML model to determine where to operate, our staff will scope areas to understand dynamics on the ground, and consult local government and key informants to sense-check the information provided by models, before making final decisions.

> **[Case Study] Benchmarking and assessment**
>
> In 2023, we partnered with the Bangladesh government (a2i) and the University of California, Berkeley to test various targeting approaches as part of a program to use best-possible targeting methods to identify and deliver cash transfers to 22,600 households in Cox's Bazar living in extreme poverty. The Berkeley team evaluated several different targeting approaches including AI/ML models trained on census data, phone metadata, census + phone metadata, as well as non-AI/ML methods like geographic targeting, proxy means tests, and random targeting. After this benchmarking exercise, we chose to use the AI/ML model trained on census data only to target the most needy households, given it demonstrated the most accurate targeting performance.

Looking Ahead

Looking ahead, we will regularly iterate on our framework as our understanding of risks, preferences, and AI/ML technology develops, and are eager to hear from and share learnings with organizations developing their thinking around responsible AI as we create best practices in the sector.